Learning Dyadic Associations of Diagnoses of the Alzheimer's Disease and Longitudinal Imaging-Genetic Biomarkers

Lvjian Lu, *Member, IEEE,* Zhennan Shi, *Member, IEEE,* Hua Wang and Feiping Nie, *Member, IEEE*

***Abstract*—Alzheimer's Disease (AD) is a clinical syndrome, characterized by progressive impairment of cognitive and memory functions. How to combine information from various data sources has become an important research topic in longitudinal AD study. In this paper, we propose a framework for simultaneous prediction of cognitive status of patients from phenotypic and genotypic data and learning dyadic association between genetic and neuroimaging data. To uncover the temporal and modality structures in the genetic and neuroimaging data, the sparsity-induced regularization is introduced in our model. To optimize the non-smooth objective, an efficient algorithm is proposed, whose convergence could be theoretically proved. The promising experimental results in extensive empirical studies performed on the ADNI cohort have validated the effectiveness of the proposed method.**

***Index Terms*—Longitudinal Study, Multi-level Logistic Classification, and Alzheimer's Disease.**

I. Introduction

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

Due to these facts, the research of AD gains more and more attention in brain science. Benefiting from the advances of imaging technology, different types of brain imaging data have been collected [2]. The structural magnetic resonance imaging (sMRI) scans provide the morphometry of the brain such as the gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF), and the positron-emission tomography (PET) scans measure the metabolic processes of the brain. These neuroimaging data obtained by different imaging technologies, measuring the same brain from different perspectives, might carry complementary information [3]. In addition, genetic variations such as single nucleotide polymorphism (SNP) are also studied to explore the genetic basis of brain structures, brain functions, and brain disorders, such as AD [4]. The genetic variations such as SNPs and neuroimaging quantitative traits are usually analyzed together. Thus, a large number of studies are proposed to uncover the relationship between genetic and neuroimaging data [5], [3]. These approaches consider that phenotype can be explained by a sum of effects from genetic variants [6]. Different sparsity induced norms are leveraged to identify the key dyadic associations between the genetic and neuroimaging data.

Since AD is a progressive brain disorder, another interesting problem is about combining genetic and neuroimaging data for automatic monitoring progression the disease status of patients. To integrate heterogeneous data to predict the disease status of patients longitudinally, different longitudinal multi-modal approaches are proposed [7]. However, these approaches only implicitly consider the potential effects between genetic and imaging data for predictions. they put on the same level genetic and imaging data, although these data do not provide the same type of information.

Thereby, in this paper, we propose a longitudinal hierarchical framework for ~~simultaneously~~ prediction of cognitive status of patients from phenotypic and genotypic data. This framework also works for explicitly learning dyadic association between genetic and neuroimaging data. As illustrated in Figure. 1, we started with the idea that learning AD diagnosis from imaging data already provides good results. Then, we considered that the decision function parameters learned from imaging data could be modulated, depending on each subject's genetic data. In other words, genes would express themselves through these parameters. Considering a linear regression that links these parameters and the genetic data, ~~it~~ leads to a multilevel model between imaging and genetics. Furthermore, to explore the temporal and multi-modal structure among the genetic and neuroimaging data, sparsity induce norms are leveraged. To learn all the decision function parameters, an efficient algorithm has been developed. The promising experimental results in extensive empirical studies performed on the ADNI cohort have validated the effectiveness of the proposed method.

II. Problem Formulation

In the prediction of AD diagnosis, we are given N training samples $\{(\mathbf{x}_{\mathcal{G}}^k, \mathbf{X}_{\mathcal{I}}^k; \mathbf{y}^k)\}_{k=1}^N$ where $\mathbf{x}_{\mathcal{G}}^k \in \Re^{d_{\mathcal{G}}}$ denotes the ge-

Lvjian Lu, Zhennan Shi and Hua Wang are with the Department of Computer Science, Colorado School of Mines Golden, CO 80401, U.S.A.(e-mail: lyujianlu@mymail.mines.edu, zhennanshi@mymail.mines.edu and huawangcs@gmail.com).

Feiping Nie is with the School of Computer Science and Center for OPTical IMagery Analysis and Learning Northwestern Polytechnical University(e-mail:feipingnie@gmail.com).

Manuscript received April 19, 2005; revised August 26, 2015.

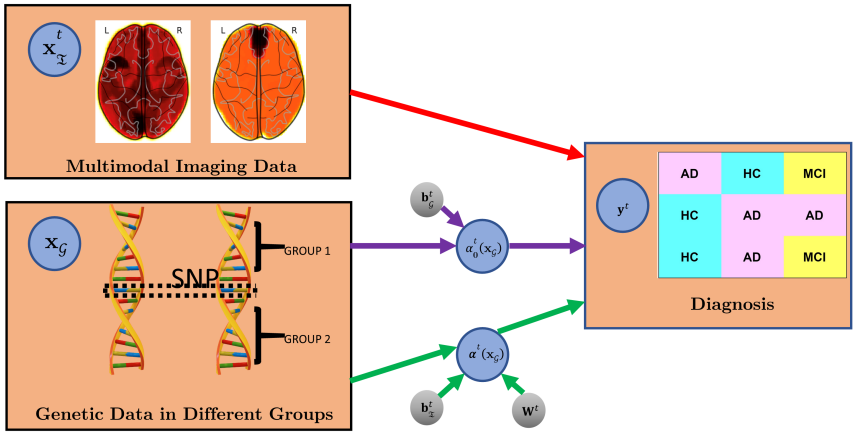

Fig. 1: Illustration of hierarchical classification for Alzheimer's Disease Prediction at time point t. In the first level, the disease status is directly predicted from the neuroimaging data at time points t. In the second level, a parametric model with parameters $\alpha_0(\mathbf{X}_\mathcal{G})$ and $\boldsymbol{\alpha}^t(\mathbf{X}_\mathcal{G})$ is computed ~~form~~ genetic data to predict the longitudinal disease status.

netic features of k-th participant, $\mathbf{X}_\mathcal{I}^k = [\mathbf{x}_\mathcal{I}^{k1}, \mathbf{x}_\mathcal{I}^{k2}, \cdots, \mathbf{x}_\mathcal{I}^{kT}] \in \Re^{d_\mathcal{I} \times T}$ collects neuroimaging features of k-th participant from time 1 to T and $\mathbf{y}^k = [y^{k1}; y^{k2}; \cdots; y^{kT}] \in \{0,1\}^T$ is the AD diagnosis of k-th participant from time 1 to T.

For each time point $t (1 \leq t \leq T)$, we combine the genetic and neuroimaging features to predict the diagnosis of AD at time t, following the multilevel logistic regression model:

$$p\left(y^t = 1 \mid \mathbf{x}_\mathcal{G}, \mathbf{x}_\mathcal{I}^t\right) = \sigma\left(\boldsymbol{\alpha}\left(\mathbf{x}_\mathcal{G}\right)^\top \mathbf{x}_\mathcal{I}^t + \alpha_0\left(\mathbf{x}_\mathcal{G}\right)\right), \quad \sigma : x \mapsto \frac{1}{1+\mathrm{e}^{-x}}, \tag{1}$$

where $\alpha_0^t\left(\mathbf{x}_\mathcal{G}\right)$ is the intercept and $\boldsymbol{\alpha}^t\left(\mathbf{x}_\mathcal{G}\right) \in \Re^{d_\mathcal{I}}$ is the parameter vector at time point t. Different from an additive model, we propose a multilevel model, for which the parameter vector $\boldsymbol{\alpha}^t\left(\mathbf{x}_\mathcal{G}\right)$ and the intercept $\alpha_0^t\left(\mathbf{x}_\mathcal{G}\right)$ at time point t depend on genetic data $\mathbf{x}_\mathcal{G}$. Here we assume that $\boldsymbol{\alpha}^t$ and α_0^t are affine functions of genetic data $\mathbf{x}_\mathcal{G}$: $\boldsymbol{\alpha}^t\left(\mathbf{x}_\mathcal{G}\right) = \mathbf{W}^t\mathbf{x}_\mathcal{G} + \mathbf{b}_\mathcal{I}^t$ and $\alpha_0^t\left(\mathbf{x}_\mathcal{G}\right) = \left(\mathbf{b}_\mathcal{G}^t\right)^\top \mathbf{x}_\mathcal{G}$, where $\mathbf{W}^t \in \Re^{d_\mathcal{I} \times d_\mathcal{G}}$, $\mathbf{b}_\mathcal{I}^t \in \Re^{d_\mathcal{I}}$ and $\mathbf{b}_\mathcal{G}^t \in \Re^{d_\mathcal{G}}$.

In order to predict the diagnosis of AD along the time point from 1 to T, longitudinal multilevel logistic regression model is used in our study, which minimizes the following objective:

$$\begin{aligned}\mathcal{J}_1 = \frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\{-y^{kt}((\mathbf{x}_\mathcal{G}^k)^\top\mathbf{W}^t\mathbf{x}_\mathcal{I}^{kt} + (\mathbf{b}_\mathcal{G}^t)^\top\mathbf{x}_\mathcal{G}^k \\ + (\mathbf{b}_\mathcal{I}^t)^\top\mathbf{x}_\mathcal{I}^{kt}) + \log\left(1 + e^{(\mathbf{x}_\mathcal{G}^k)^\top\mathbf{W}^t\mathbf{x}_\mathcal{I}^{kt} + (\mathbf{b}_\mathcal{G}^t)^\top\mathbf{x}_\mathcal{G}^k + (\mathbf{b}_\mathcal{I}^t)^\top\mathbf{x}_\mathcal{I}^{kt}}\right)\} \\ + \gamma_1\sum_{t=1}^{T}\left\|\mathbf{W}^t\right\|_2^2 + \gamma_2\sum_{t=1}^{T}\left\|\mathbf{b}_\mathcal{G}^t\right\|_2^2 + \gamma_3\sum_{t=1}^{T}\left\|\mathbf{b}_\mathcal{I}^t\right\|_2^2,\end{aligned} \tag{2}$$

where $\mathbf{b}_\mathcal{G}^t$ and $\mathbf{b}_\mathcal{I}^t$ are coefficient vectors for genetic data and neuroimaging data at time point t respectively and $\mathcal{W} = \{\mathbf{W}^1, \mathbf{W}^2, \cdots, \mathbf{W}^T\} \in \Re^{d_\mathcal{I} \times d_\mathcal{G} \times T}$ denotes a tensor of logistic regression weights.

The objective $\mathcal{J}_1$ in Eq. (2) does not take into account the temporal correlations over time, because the optimization problem can be decoupled for each individual time point separately. During the AD progression, a small set of features is predictive of the progression and the logistic regression models from different points stay stable [8]. To this end, following the work [8], [9], we introduce the sparse regularization as follows:

$$\begin{aligned}\mathcal{J}_2 = \frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\{-y^{kt}((\mathbf{x}_\mathcal{G}^k)^\top\mathbf{W}^t\mathbf{x}_\mathcal{I}^{kt} + (\mathbf{b}_\mathcal{G}^t)^\top\mathbf{x}_\mathcal{G}^k \\ + (\mathbf{b}_\mathcal{I}^t)^\top\mathbf{x}_\mathcal{I}^{kt}) + \log\left(1 + e^{(\mathbf{x}_\mathcal{G}^k)^\top\mathbf{W}^t\mathbf{x}_\mathcal{I}^{kt} + (\mathbf{b}_\mathcal{G}^t)^\top\mathbf{x}_\mathcal{G}^k + (\mathbf{b}_\mathcal{I}^t)^\top\mathbf{x}_\mathcal{I}^{kt}}\right)\} \\ + \gamma_1\sum_{t=1}^{T}\left\|\mathbf{W}^t\right\|_2^2 + \gamma_2\left\|\mathbf{B}_\mathcal{G}\right\|_{2,1} + \gamma_3\left\|\mathbf{B}_\mathcal{I}\right\|_{2,1},\end{aligned} \tag{3}$$

where $\mathbf{B}_\mathcal{G} = [\mathbf{b}_\mathcal{G}^1, \mathbf{b}_\mathcal{G}^2, \cdots, \mathbf{b}_\mathcal{G}^T] \in \Re^{d_\mathcal{G} \times T}$ and $\mathbf{B}_\mathcal{I} = [\mathbf{b}_\mathcal{I}^1, \mathbf{b}_\mathcal{I}^2, \cdots, \mathbf{b}_\mathcal{I}^T] \in \Re^{d_\mathcal{I} \times T}$ collect all the coefficient vectors for genetic data and neuroimaging data along all the time points respectively.

During the progression of AD, genetic data and neuroimaging data are interrelated to each other. Although correlation between genetic data and neuroimaging data at different time point are different, its associations share similar patterns at two consecutive time points [8], [10]. To capture the correlations between genetic data and neuroimaging at different time points, we impose low-rank regularization to discover the common subspace shared by genetic and neuroimaging data as follows:

$$\begin{aligned}\mathcal{J}_3 = \frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\{-y^{kt}((\mathbf{x}_\mathcal{G}^k)^\top\mathbf{W}^t\mathbf{x}_\mathcal{I}^{kt} + (\mathbf{b}_\mathcal{G}^t)^\top\mathbf{x}_\mathcal{G}^k \\ + (\mathbf{b}_\mathcal{I}^t)^\top\mathbf{x}_\mathcal{I}^{kt} + \log\left(1 + e^{(\mathbf{x}_\mathcal{G}^k)^\top\mathbf{W}^t\mathbf{x}_\mathcal{I}^{kt} + (\mathbf{b}_\mathcal{G}^t)^\top\mathbf{x}_\mathcal{G}^k + (\mathbf{b}_\mathcal{I}^t)^\top\mathbf{x}_\mathcal{I}^{kt}}\right)\} \\ + \gamma_1(\left\|\mathbf{W}_{(1)}\right\|_* + \left\|\mathbf{W}_{(2)}\right\|_*) + \gamma_2\left\|\mathbf{B}_\mathcal{G}\right\|_{2,1} + \gamma_3\left\|\mathbf{B}_\mathcal{I}\right\|_{2,1},\end{aligned} \tag{4}$$

where $\mathbf{W}_{(1)} = [\mathbf{W}^1, \mathbf{W}^2, \cdots, \mathbf{W}^T] \in \Re^{d_\mathcal{I} \times d_\mathcal{G} T}$ and $\mathbf{W}_{(2)} = [(\mathbf{W}^1)^\top, (\mathbf{W}^2)^\top, \cdots, (\mathbf{W}^T)^\top] \in \Re^{d_\mathcal{G} \times d_\mathcal{I} T}$ are unfolded forms of tensor $\mathcal{W}$.

In addition, the neuroimaging data in our study is a concatenation of voxel-based morphometry (VBM) and FreeSurfer. Moreover, Genetic data are a sequence of Single-nucleotide polymorphism (SNP) counted by minor allele. Following the AlzGene database [11], 1224 SNPs in our study are grouped in 37 Alzheimer's genes. From a multi-modal perspective, the features of a specific biomarker modalities can be more or less discriminative for different clusters. Motivated by previous work [12], a group ℓ_1-norm (G1-norm) is introduced to untangle the multi-modality structure, which is defined as $\|\mathbf{M}\|_{G_1} = \sum_{i=1}^{m_\mathcal{I}} \sum_{j=1}^{m_\mathcal{G}} \left\|\mathbf{M}_i^j\right\|_2$. Here we write $\mathbf{M} = [\mathbf{M}_1^1, \cdots, \mathbf{M}_c^1; \cdots; \mathbf{M}_1^k, \cdots, \mathbf{M}_c^k]$ and $\mathbf{M}_i^j \in \Re^{d_i \times d_j}$ indicates the weights of dyadic association between i-th view of imaging data and j-th view genetic data, where d_i denotes the dimensionality of i-th modality. Moreover, in certain cases, even if most features in one modality are not discriminative for a group of objects, a small number of features in the same modality can still be highly discriminative. To identify these small group of features, we add an additional G1-norm regularization for AD diagnosis task as follows:

$$\begin{aligned}\mathcal{J}_4 = &\frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\{-y^{kt}((\mathbf{x}_\mathcal{G}^k)^\top \mathbf{W}^t \mathbf{x}_\mathcal{I}^{kt} + (\mathbf{b}_\mathcal{G}^t)^\top \mathbf{x}_\mathcal{G}^k \\ &+ (\mathbf{b}_\mathcal{I}^t)^\top \mathbf{x}_\mathcal{I}^{kt}) + \log\left(1 + e^{(\mathbf{x}_\mathcal{G}^k)^\top \mathbf{W}^t \mathbf{x}_\mathcal{I}^{kt} + (\mathbf{b}_\mathcal{G}^t)^\top \mathbf{x}_\mathcal{G}^k + (\mathbf{b}_\mathcal{I}^t)^\top \mathbf{x}_\mathcal{I}^{kt}}\right)\} \\ &+ \gamma_1(\left\|\mathbf{W}_{(1)}\right\|_* + \left\|\mathbf{W}_{(2)}\right\|_*) + \sum_{t=1}^{T}\gamma_2 \left\|\mathbf{W}^t\right\|_{G_1} \\ &+ \gamma_3 \left\|\mathbf{B}_\mathcal{G}\right\|_{2,1} + \gamma_4 \left\|\mathbf{B}_\mathcal{I}\right\|_{2,1}, \end{aligned} \tag{5}$$

Moreover, SNPs on the same chromosome with close distance tend to be inherited together and correlated with each other. Thus, they should be considered together when we predict the cognitive status of the AD participants. Motivated by previous work [13], group $\ell_{2,1}$ norm is introduced to enforce the group sparsity among SNPs, which is defined as $\|\mathbf{M}\|_{G_{2,1}} = \sum_{i=1}^{m_\mathcal{G}} \sqrt{\sum_{j=1}^{m_\mathcal{I}} \left\|\mathbf{M}_i^j\right\|_2^2}$. Thus we can rewrite the objective as follows:

$$\begin{aligned}\mathcal{J}_4 = &\frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\{-y^{kt}((\mathbf{x}_\mathcal{G}^k)^\top \mathbf{W}^t \mathbf{x}_\mathcal{I}^{kt} + (\mathbf{b}_\mathcal{G}^t)^\top \mathbf{x}_\mathcal{G}^k \\ &+ (\mathbf{b}_\mathcal{I}^t)^\top \mathbf{x}_\mathcal{I}^{kt}) + \log\left(1 + e^{(\mathbf{x}_\mathcal{G}^k)^\top \mathbf{W}^t \mathbf{x}_\mathcal{I}^{kt} + (\mathbf{b}_\mathcal{G}^t)^\top \mathbf{x}_\mathcal{G}^k + (\mathbf{b}_\mathcal{I}^t)^\top \mathbf{x}_\mathcal{I}^{kt}}\right)\} \\ &+ \gamma_1(\left\|\mathbf{W}_{(1)}\right\|_* + \left\|\mathbf{W}_{(2)}\right\|_*) + \sum_{t=1}^{T}\gamma_2 \left\|\mathbf{W}^t\right\|_{G_1} \\ &+ \sum_{t=1}^{T}\gamma_3 \left\|\mathbf{W}^t\right\|_{G_{2,1}} + \gamma_3 \left\|\mathbf{B}_\mathcal{G}\right\|_{2,1} + \gamma_5 \left\|\mathbf{B}_\mathcal{I}\right\|_{2,1}, \end{aligned} \tag{6}$$

III. The Optimization Algorithm

Although the motivations of the formulation of our new method in Eq. (6) is clear and justifiable, it is a non-smooth objective, which is difficult to efficiently solve in general. Thus we derive the solution of our objective in this section. Motivated by earlier works that use the iterative reweighted method [14] to solve non-smooth objectives, we propose a novel smoothed iterative reweighted method to solve the general optimization problem as our proposed objective in Eq. (6).

A. Smoothed Iterative Reweighted Method

Algorithm 1 Solve the optimization problem in Eq. (10).

Initialization: $x \in \mathcal{C}$;

1. For each i, calculate $D_i = \frac{p}{2}\left(\|g(x)_i\|_2^2 + \delta I\right)^{\frac{p-2}{2}}$;

2. Update x by solving the problem $\min_{x\in\mathcal{C}} f(x) + \sum_i \mathrm{tr}\left(g_i^T(x) g_i(x) D_i\right)$;

Output: x.

In this subsection, we focus on solving a general problem as follows:

$$\min_{x\in\mathcal{C}} f(x) + \sum_i \mathrm{tr}\left(\left(g_i^T(x) g_i(x)\right)^{\frac{p}{2}}\right), \tag{7}$$

note that $g_i(x)$ is scalar, vector or matrix output function, then $\left(g_i^T(x) g_i(x)\right)^{\frac{p}{2}}$ becomes the following terms respectively:

$$\mathrm{tr}\left(\left(g_i^T(x) g_i(x)\right)^{\frac{p}{2}}\right) = \begin{cases} |g_i(x)|^p & g_i(x) \text{ is scalar} \\ \|g_i(x)\|_2^p & g_i(x) \text{ is vector} \\ \|g_i(x)\|_{S_p}^p & g_i(x) \text{ is matrix} \end{cases} \tag{8}$$

For the case that $p = 1$, $\mathrm{tr}\left(\left(g_i^T(x) g_i(x)\right)^{\frac{p}{2}}\right)$ denotes ℓ_1-norm, ℓ_2-norm and trace norm respectively:

$$\mathrm{tr}\left(\left(g_i^T(x) g_i(x)\right)^{\frac{1}{2}}\right) = \begin{cases} |g_i(x)| & g_i(x) \text{ is scalar} \\ \|g_i(x)\|_2 & g_i(x) \text{ is vector} \\ \|g_i(x)\|_* & g_i(x) \text{ is matrix} \end{cases} \tag{9}$$

For the Eq. (7) is non-smooth, we can turn to solve an approximation problem of it, a smooth problem formulated as follows:

$$\min_{x\in C} f(x) + \mu \sum_i \mathrm{tr}\left(\left(g_i^T(x) g_i(x) + \delta I\right)^{\frac{p}{2}}\right), \tag{10}$$

when $\delta \to 0$, Eq. (10) is reduced to Eq. (7) since the following equations hold:

$$\begin{aligned}&\lim_{\delta\to 0} \mathrm{tr}\left(\left(g_i^T(x) g_i(x) + \delta I\right)^{\frac{p}{2}}\right) \\ &= \begin{cases} |g_i(x)| & g_i(x) \text{ is scalar} \\ \|g_i(x)\|_2 & g_i(x) \text{ is vector} \\ \|g_i(x)\|_* & g_i(x) \text{ is matrix} \end{cases} \end{aligned} \tag{11}$$

Before deriving the algorithm for optimizing problem (10), we need some significant lemmas as follows:

Lemma 1 (chain rule)**.** *Suppose $g(x)$ is a matrix output function, $h(x)$ is a scalar output function, x is a scalar, vector or matrix variable, then we have:*

$$\begin{aligned}\frac{\partial h(g(x))}{\partial x} &= \frac{\sum_{i,j} \frac{\partial h(g(x))}{\partial g_{ij}(x)} \partial g_{ij}(x)}{\partial x} \\ &= \frac{\mathrm{tr}\left(\left(\frac{\partial h(g(x))}{\partial g(x)}\right)^T \partial g(x)\right)}{\partial x}. \end{aligned} \tag{12}$$

According to the chain rule in Lemma 1, we have the following two lemmas:

Lemma 2. *Suppose $g(x)$ is a scalar, vector or matrix output function, x is a scalar, vector or matrix variable, then we have:*

$$\begin{aligned}&\frac{\partial \operatorname{tr}\left(\left(g^T(x)g(x)+\delta I\right)^{\frac{p}{2}}\right)}{\partial x}\\&=\frac{\operatorname{tr}\left(2\frac{p}{2}\left(g^T(x)g(x)+\delta I\right)^{\frac{p-2}{2}}g^T(x)\partial g(x)\right)}{\partial x}.\end{aligned}\tag{13}$$

Proof. Let $h(x)=\operatorname{tr}(x^Tx+\delta I)^{\frac{p}{2}}$, we have:

$$\frac{\partial h(x)}{\partial x}=2\frac{p}{2}x\left(x^Tx+\delta I\right)^{\frac{p-2}{2}},\tag{14}$$

further, we can obtain:

$$\frac{\partial h(g(x))}{\partial g(x)}=2\frac{p}{2}g(x)\left(g^T(x)g(x)+\delta I\right)^{\frac{p-2}{2}}\tag{15}$$

According to Lemma 1, we get the Eq. (13). ■

Lemma 3. *Suppose $g(x)$ is a scalar, vector or matrix output function, x is a scalar, vector or matrix variable, D is a constant and D is symmetrical if D is a matrix, then we have*

$$\frac{\partial \operatorname{tr}\left(g^T(x)g(x)D\right)}{\partial x}=\frac{\operatorname{tr}\left(2Dg^T(x)\partial g(x)\right)\right)}{\partial x}.\tag{16}$$

Proof. Let $h(x)=\operatorname{tr}\left(x^TxD\right)$, we have $\frac{\partial h(x)}{\partial x}$, then we have $\frac{\partial h(g(x))}{\partial x}=2g(x)D$. So according to the chain rule in Lemma 1, we get the Eq. (16). ■

Now we derive the algorithm for optimizing the problem (10). The Lagrangian function of the problem (10) is:

$$\begin{aligned}\mathcal{L}(x,\lambda)=&f(x)+\mu\sum_i\operatorname{tr}\left(\left(g_i^T(x)g_i(x)+\delta I\right)^{\frac{p}{2}}\right)\\&-r(x,\lambda),\end{aligned}\tag{17}$$

where $r(x,\lambda)$ is a Lagrangian term for the constraint $x\in\mathcal{C}$. By setting the derivative of Eq. (17) w.r.t. x to zero, we have:

$$\begin{aligned}\frac{\partial L(x,\lambda)}{\partial x}=&f'(x)+\mu\sum_i\frac{\partial \operatorname{tr}\left(\left(g_i^T(x)g_i(x)+\delta I\right)^{\frac{p}{2}}\right)}{\partial x}\\&-\frac{\partial r(x,\lambda)}{\partial x}\\&=0.\end{aligned}\tag{18}$$

According to Lemma 2, Eq. (18) can be rewritten as:

$$\begin{aligned}&f'(x)+\mu\sum_i\frac{\operatorname{tr}\left(2\frac{p}{2}\left(g_i^T(x)g_i(x)+\delta I\right)^{\frac{p-2}{2}}g_i^T(x)\partial g_i(x)\right)}{\partial x}\\&-\frac{\partial r(x,\lambda)}{\partial x}=0\end{aligned}\tag{19}$$

If we can find a solution x that satisfies the Eq. (17), then we usually find a stationary point or global optimal solution to the problem (10) according to the Karush-Kuhn-Tucker conditions. However, directly finding a solution x that satisfies Eq. (17) is generally not an easy task. In this paper, we propose an iterative algorithm to find it. A basic observation is that, if $D=\frac{p}{2}\left(g_i(x)^Tg_i(x)+\delta I\right)^{\frac{p-2}{2}}$ is a given constant, then Eq. (17) is reduced to

$$f'(x)+\mu\sum_i\frac{\operatorname{tr}\left(2D_ig_i^T(x)\partial g_i(x)\right)}{\partial x}-\frac{\partial r(x,\lambda)}{\partial x}=0.\tag{20}$$

which is equivalent to solving the following problem:

$$\min_{x\in\mathcal{C}}f(x)+\mu\sum_i\operatorname{tr}\left(g_i^T(x)g_i(x)D_i\right)\tag{21}$$

Based on the observation, we first guess a solution x, then we calculate D_i based on the current solution x and then update the current solution x by the optimal solution of the problem (21) based on the calculated D_i. We iteratively perform this procedure until it converges.

B. Convergence Analysis of Smoothed Iterative Reweighted Method

Before proving the convergence of the Algorithm 1, we first introduce several significant lemmas:

Lemma 4. *For any $\sigma>0$, the following inequality holds when $0<p\leq 2$:*

$$\frac{p}{2}\sigma-\sigma^{\frac{p}{2}}+\frac{2-p}{2}\geq 0.\tag{22}$$

Proof. Denoting $f(\sigma)=p\sigma-2\sigma^{\frac{p}{2}}+2-p$, we have the following derivatives:

$$f'(\sigma)=p(1-\sigma^{\frac{p-2}{2}}),\quad\text{and}\quad f''(\sigma)=\frac{p(2-p)}{2}\sigma_i^{\frac{p-4}{2}}.$$

When $\sigma>0$ and $0<p\leq 2$, $f''(\sigma)\geq 0$ and $\sigma=1$ is the only point that $f'(\sigma)=0$. Note that $f(1)=0$, thus when $\sigma>0$ and $0<p\leq 2$, then $f(\sigma)\geq 0$, which indicates inequality in Eq. (22) holds true. ■

Lemma 5. *[15] For any positive definite matrices $\tilde{M},M$ with the same size, suppose the eigen-decomposition $\tilde{M}=U\Sigma U^T$, $M=V\Lambda V^T$, where the eigenvalues in Σ are in increasing order and the eigenvalues in Λ are in decreasing order. Then the following inequality holds:*

$$\operatorname{tr}(\tilde{M}M)\geq\operatorname{tr}(\Sigma\Lambda).\tag{23}$$

Lemma 6. *For any positive definite matrices $\tilde{M},M$ with the same size, the following inequality holds when $0<p\leq 2$:*

$$\operatorname{tr}(\tilde{M}^{\frac{p}{2}})-\frac{p}{2}\operatorname{tr}(\tilde{M}M^{\frac{p-2}{2}})\leq\operatorname{tr}(M^{\frac{p}{2}})-\frac{p}{2}\operatorname{tr}(MM^{\frac{p-2}{2}}).\tag{24}$$

Proof. For any $\sigma>0$, $\lambda>0$ and $0<p\leq 2$, according to Lemma 4, we have $\frac{p}{2}(\frac{\sigma}{\lambda})-(\frac{\sigma}{\lambda})^{\frac{p}{2}}+\frac{2-p}{2}\geq 0$, which indicates

$$\frac{p}{2}\sigma\lambda^{\frac{p-2}{2}}-\sigma^{\frac{p}{2}}+\frac{2-p}{2}\lambda^{\frac{p}{2}}\geq 0.\tag{25}$$

Suppose the eigen-decomposition $\tilde{M} = U\Sigma U^T$, $M = V\Lambda V^T$, where the eigenvalues in Σ is in increasing order and the eigenvalues in Λ is in decreasing order. Then according to Inequality (25), we have

$$\frac{p}{2}\operatorname{tr}(\Sigma\Lambda^{\frac{p-2}{2}}) - \operatorname{tr}(\Sigma^{\frac{p}{2}}) + \frac{2-p}{2}\operatorname{tr}(\Lambda^{\frac{p}{2}}) \geq 0. \tag{26}$$

and according to Lemma 5, we have:

$$\frac{p}{2}\operatorname{tr}(\tilde{M}M^{\frac{p-2}{2}}) - \frac{p}{2}\operatorname{tr}(\Sigma\Lambda^{\frac{p-2}{2}}) \geq 0. \tag{27}$$

$$\frac{p}{2}\operatorname{tr}(\tilde{M}M^{\frac{p-2}{2}}) - \operatorname{tr}(\Sigma^{\frac{p}{2}}) + \frac{2-p}{2}\operatorname{tr}(\Lambda^{\frac{p}{2}}) \geq 0. \tag{28}$$

Note that $\operatorname{tr}(\tilde{M}^{\frac{p}{2}}) = \operatorname{tr}(\Sigma^{\frac{p}{2}})$ and $\operatorname{tr}(M^{\frac{p}{2}}) = \operatorname{tr}(\Lambda^{\frac{p}{2}})$, so we have

$$\begin{aligned}
&\tfrac{p}{2}\operatorname{tr}(\tilde{M}M^{\frac{p-2}{2}}) - \operatorname{tr}(\tilde{M}^{\frac{p}{2}}) + \tfrac{2-p}{2}\operatorname{tr}(M^{\frac{p}{2}}) \geq 0 \\
&\Rightarrow \operatorname{tr}(\tilde{M}^{\frac{p}{2}}) - \tfrac{p}{2}\operatorname{tr}(\tilde{M}M^{\frac{p-2}{2}}) \leq \tfrac{2-p}{2}\operatorname{tr}(M^{\frac{p}{2}}) \\
&\Rightarrow \operatorname{tr}(\tilde{M}^{\frac{p}{2}}) - \tfrac{p}{2}\operatorname{tr}(\tilde{M}M^{\frac{p-2}{2}}) \leq \operatorname{tr}(M^{\frac{p}{2}}) - \tfrac{p}{2}\operatorname{tr}(MM^{\frac{p-2}{2}}).
\end{aligned}$$

which completes the proof. ■

Lemma 7. *For any matrices $\tilde{A}, A$ with the same size and $\delta > 0$, the following inequality holds when $0 < p \leq 2$:*

$$\begin{aligned}
&\operatorname{tr}((\tilde{A}^T\tilde{A} + \delta I)^{\frac{p}{2}}) - \tfrac{p}{2}\operatorname{tr}(\tilde{A}^T\tilde{A}(A^TA + \delta I)^{\frac{p-2}{2}}) \\
&\leq \operatorname{tr}((A^TA + \delta I)^{\frac{p}{2}}) - \tfrac{p}{2}\operatorname{tr}(A^TA(A^TA + \delta I)^{\frac{p-2}{2}}).
\end{aligned} \tag{29}$$

Proof. Note that $\tilde{A}^T\tilde{A}+\delta I$ and $A^TA+\delta I$ are positive definite matrices since $\delta > 0$. Then according to Lemma 6, we have

$$\begin{aligned}
&\operatorname{tr}((\tilde{A}^T\tilde{A} + \delta I)^{\frac{p}{2}}) - \tfrac{p}{2}\operatorname{tr}((\tilde{A}^T\tilde{A} + \delta I)(A^TA + \delta I)^{\frac{p-2}{2}}) \\
&\leq \operatorname{tr}((A^TA + \delta I)^{\frac{p}{2}}) - \tfrac{p}{2}\operatorname{tr}((A^TA + \delta I)(A^TA + \delta I)^{\frac{p-2}{2}}).
\end{aligned} \tag{30}$$

which indicates that inequality in Eq. (29) holds true. ■

Theorem 1. *The Algorithm 1 will monotonically decrease the objective of the problem (10) in each iteration until the algorithm converges.*

Proof. In step 2 of Algorithm 1 we denote the updated x as $\tilde{x}$. According to step 2, we know:

$$f(\tilde{x}) + \sum_i \operatorname{tr}(g_i^T(\tilde{x})g_i(\tilde{x})D_i) \leq f(x) + \sum_i \operatorname{tr}(g_i^T(x)g_i(x)D_i). \tag{31}$$

where the inequality holds when and only when the algorithm converges.

For each i, according to Lemma 7, we have

$$\begin{aligned}
&\operatorname{tr}((g_i^T(\tilde{x})g_i(\tilde{x}) + \delta I)^{\frac{p}{2}}) - \tfrac{p}{2}\operatorname{tr}(g_i^T(\tilde{x})g_i(\tilde{x})(g_i^T(x)g_i(x) + \\
&\delta I)^{\frac{p-2}{2}}) \leq \operatorname{tr}((g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}}) - \tfrac{p}{2}\operatorname{tr}(g_i^T(x)g_i(x) \\
&(g_i^T(x)g_i(x) + \delta I)^{\frac{p-2}{2}}).
\end{aligned} \tag{32}$$

Note that $D_i = \frac{p}{2}(g_i^T(x)g_i(x)+\delta I)^{\frac{p-2}{2}}$, so for each i we have

$$\begin{aligned}
&\operatorname{tr}((g_i^T(\tilde{x})g_i(\tilde{x}) + \delta I)^{\frac{p}{2}}) - \operatorname{tr}(g_i^T(\tilde{x})g_i(\tilde{x})D_i) \\
&\leq \operatorname{tr}((g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}}) - \operatorname{tr}(g_i^T(x)g_i(x)D_i).
\end{aligned} \tag{33}$$

Then we have

$$\begin{aligned}
&\sum_i \operatorname{tr}((g_i^T(\tilde{x})g_i(\tilde{x}) + \delta I)^{\frac{p}{2}}) - \sum_i \operatorname{tr}(g_i^T(\tilde{x})g_i(\tilde{x})D_i) \\
&\leq \sum_i \operatorname{tr}((g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}}) - \sum_i \operatorname{tr}(g_i^T(x)g_i(x)D_i).
\end{aligned} \tag{34}$$

Summing inequality in Eq. (31) and Inequality (34) on both sides, we arrive at :

$$\begin{aligned}
&f(\tilde{x}) + \sum_i \operatorname{tr}((g_i^T(\tilde{x})g_i(\tilde{x}) + \delta I)^{\frac{p}{2}}) \\
&\leq f(x) + \sum_i \operatorname{tr}((g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}}).
\end{aligned} \tag{35}$$

Note that Inequality (35) holds only when the algorithm converges. Thus the Algorithm 1 will monotonically decrease the objective of the problem in Eq. (6) in each iteration until the algorithm converges. ■

In the convergence, equality in Eq. (19) will hold, thus the KKT condition in Eq. (10) is satisfied. Therefore, Algorithm 1 will converge to a local optimum solution to Eq. (10). If problem in Eq. (10) is convex, Algorithm 1 will converge to a global optimum solution.

C. Optimization Algorithm

Equipment with our new optimization framework, we rewrite the objective Eq. (6) to solve the following problem:

$$\begin{aligned}
\mathcal{J}_{\mathrm{R}} = &\frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\{-y^{kt}((\mathbf{x}_{\mathcal{G}}^k)^\top\mathbf{W}^t\mathbf{x}_{\mathcal{I}}^{kt} + (\mathbf{b}_{\mathcal{G}}^t)^\top\mathbf{x}_{\mathcal{G}}^k \\
&+ (\mathbf{b}_{\mathcal{I}}^t)^\top\mathbf{x}_{\mathcal{I}}^{kt}) + \log\left(1 + e^{(\mathbf{x}_{\mathcal{G}}^k)^\top\mathbf{W}^t\mathbf{x}_{\mathcal{I}}^{kt} + (\mathbf{b}_{\mathcal{G}}^t)^\top\mathbf{x}_{\mathcal{G}}^k + (\mathbf{b}_{\mathcal{I}}^t)^\top\mathbf{x}_{\mathcal{I}}^{kt}}\right)\} \\
&+ \gamma_1\left[\operatorname{tr}\left(\mathbf{W}_{(1)}^\top\hat{\mathbf{D}}_1\mathbf{W}_{(1)}\right) + \operatorname{tr}\left(\mathbf{W}_{(2)}\tilde{\mathbf{D}}_1\mathbf{W}_{(2)}^\top\right)\right] \\
&+ \gamma_2\sum_{t=1}^{T}\sum_{i=1}^{m_{\mathcal{I}}}\sum_{j=1}^{m_{\mathcal{G}}}\operatorname{tr}((\mathbf{W}_i^j)^t)^\top(\mathbf{D}_i^j)^t(\mathbf{W}_i^j)^t) \\
&+ \gamma_3\sum_{t=1}^{T}\operatorname{tr}((\mathbf{W}^t)^\top\mathbf{D}_4\mathbf{W}^t) + \gamma_4\operatorname{tr}(\mathbf{B}_{\mathcal{G}}^\top\mathbf{D}_4\mathbf{B}_{\mathcal{G}}) \\
&+ \gamma_5\operatorname{tr}(\mathbf{B}_{\mathcal{I}}^\top\mathbf{D}_5\mathbf{B}_{\mathcal{I}})
\end{aligned} \tag{36}$$

where $\hat{\mathbf{D}}_1 = \frac{1}{2}(\mathbf{W}_{(1)}\mathbf{W}_{(1)}^\top)^{-1/2}$, and $\tilde{\mathbf{D}}_1 = \frac{1}{2}(\mathbf{W}_{(2)}\mathbf{W}_{(2)}^\top)^{-1/2}$. $(\mathbf{D}_i^j)^t = \frac{1}{2\sqrt{\|(\mathbf{W}_i^j)^t\|_2^2+\sigma}}(\mathbf{I}_i^j)^t$ and $\sigma \to 0$. $\mathbf{D}_3$ is a block diagonal matrix whose j-th diagonal element is $\frac{1}{2||\sqrt{\mathbf{W}_i^j||_2^2+\sigma}}\mathbf{I}_j$. $\mathbf{D}_4 = \frac{1}{2}(\mathbf{B}_{\mathcal{G}}\mathbf{B}_{\mathcal{G}}^\top)^{-1/2}$ and $\mathbf{D}_5 = \frac{1}{2}(\mathbf{B}_{\mathcal{I}}\mathbf{B}_{\mathcal{I}}^\top)^{-1/2}$.

In this section, we will give a detailed introduction of Alternating Direction Method of Multipliers (ADMM), which was proposed in [16], [17] to solve convex optimization problems by breaking them into smaller pieces that are easier to handle. Specifically, given the following objective with the equality constraint:

$$\min_{x,z} f(x) + g(z), \qquad s.t. \quad h(x,z) = 0, \tag{37}$$

Algorithm 2 solves the problem by decoupling it into subproblems and optimizing each variable while fixing others [16], [17], where y is the Lagrangian multiplier to the constraint h.

Algorithm 2 The ADMM algorithm.

Set $1 < \rho < 2$ and initialize $\mu > 0$ and y;

1. Update x by solving $x^{k+1} = \arg\min_x (f(x) + \frac{\mu}{2}\|h(x, z^k) + \frac{y^k}{\mu}\|^2)$;
2. Update z by solving $z^{k+1} = \arg\min_z (g(z) + \frac{\mu}{2}\|h(x^{k+1}, z) + \frac{y^k}{\mu}\|^2)$;
3. Update y by $y^{k+1} = y^k + \mu h(x^{k+1}, z^{k+1})$;
4. Update μ by $\mu = \rho\mu$.

Algorithm 3 Solve the optimization problem in Eq. (38).

Initialization: $\mathbf{W}^t$, $\mathbf{P}^t$, $\mathbf{\Lambda}^t$, $1 < \rho < 2$, $\mu, \gamma_1, \gamma_2, \gamma_3, \gamma_4 > 0$;
1. Update $\mathbf{W}^t$ by solving by Sylvester equation:
$\left(2\gamma_1\hat{\mathbf{D}}_1 + \mu\mathbf{I} + 2\gamma_3\mathbf{D}_3\right)\mathbf{W}^t + 2\gamma_1\mathbf{W}^t\tilde{\mathbf{D}}_1 = \frac{1}{N}\sum_{k=1}^{N}\{y^{kt}\mathbf{x}_{\mathcal{I}}^{kt}(\mathbf{x}_{\mathcal{G}}^k)^\top - \frac{m^{kt}}{1+m^{kt}}\mathbf{x}_{\mathcal{I}}^{kt}(\mathbf{x}_{\mathcal{G}}^k)^\top\} + \mu\mathbf{P}^t - \mathbf{\Lambda}^t$, where $m^{kt} = e^{(\mathbf{x}_{\mathcal{G}}^k)^\top\mathbf{W}^t\mathbf{x}_{\mathcal{I}}^{kt} + (\mathbf{b}_{\mathcal{G}}^t)^\top\mathbf{x}_{\mathcal{G}}^k + (\mathbf{b}_{\mathcal{I}}^t)^\top\mathbf{x}_{\mathcal{I}}^{kt}}$;
2. Update $(\mathbf{P}_i^j)^t$ by $(\mathbf{P}_i^j)^t = (2\gamma_2(\mathbf{D}_i^j)^t + \mu\mathbf{I})^{-1}((\mathbf{W}_i^j)^t + \frac{1}{\mu}(\mathbf{\Lambda}_i^j)^t)$;
3. Update $\mathbf{b}_{\mathcal{G}}^t$ by $\mathbf{b}_{\mathcal{G}}^t = -(2\gamma_4\mathbf{D}_4)^{-1}\frac{1}{N}\sum_{k=1}^{N}(-y^{kt}\mathbf{x}_{\mathcal{G}}^k + \frac{m^{kt}}{1+m^{kt}}\mathbf{x}_{\mathcal{G}}^k)$;
4. Updating $\mathbf{b}_{\mathcal{I}}^t$ by $\mathbf{b}_{\mathcal{I}}^t = -(2\gamma_5\mathbf{D}_5)^{-1}\frac{1}{N}\sum_{k=1}^{N}(-y^{kt}\mathbf{x}_{\mathcal{I}}^{kt} + \frac{m^{kt}}{1+m^{kt}}\mathbf{x}_{\mathcal{I}}^{kt})$;
5. Updating $\mathbf{\Lambda}^t$ by $\mathbf{\Lambda}^t = \mathbf{\Lambda}^t + \mu(\mathbf{W}^t - \mathbf{P}^t)$;
6. Updating μ by $\mu = \mu\rho$;.
Output: $\mathbf{W}^t$, $\mathbf{B}_{\mathcal{G}}^t$, $\mathbf{B}_{\mathcal{I}}^t$.

It is worth noting that Algorithm 2 was proved to converge Q-linearly to the optimal solution [16].

Following the framework Alternating Direction Method of Multipliers (ADMM), which was proposed in [17], we further rewrite the objective in Eq. (36) as follows:

$$\begin{aligned}\mathcal{J}_{\text{ADMM}} = &\frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\{-y^{kt}((\mathbf{x}_{\mathcal{G}}^k)^\top\mathbf{W}^t\mathbf{x}_{\mathcal{I}}^{kt} + (\mathbf{b}_{\mathcal{G}}^t)^\top\mathbf{x}_{\mathcal{G}}^k \\ &+ (\mathbf{b}_{\mathcal{I}}^t)^\top\mathbf{x}_{\mathcal{I}}^{kt}) + \log\left(1 + e^{(\mathbf{x}_{\mathcal{G}}^k)^\top\mathbf{W}^t\mathbf{x}_{\mathcal{I}}^{kt} + (\mathbf{b}_{\mathcal{G}}^t)^\top\mathbf{x}_{\mathcal{G}}^k + (\mathbf{b}_{\mathcal{I}}^t)^\top\mathbf{x}_{\mathcal{I}}^{kt}}\right)\} \\ &+ \gamma_1\left[\operatorname{tr}\left(\mathbf{W}_{(1)}^\top\hat{\mathbf{D}}_1\mathbf{W}_{(1)}\right) + \operatorname{tr}\left(\mathbf{W}_{(2)}\tilde{\mathbf{D}}_1\mathbf{W}_{(2)}^\top\right)\right] \\ &+ \gamma_2\sum_{t=1}^{T}\sum_{i=1}^{m_{\mathcal{I}}}\sum_{j=1}^{m_{\mathcal{G}}}\operatorname{tr}((\mathbf{P}_i^j)^t)^\top(\mathbf{D}_i^j)^t(\mathbf{P}_i^j)^t) \\ &+ \gamma_3\sum_{t=1}^{T}\operatorname{tr}((\mathbf{W}^t)^\top\mathbf{D}_4\mathbf{W}^t) + \gamma_4\operatorname{tr}(\mathbf{B}_{\mathcal{G}}^\top\mathbf{D}_4\mathbf{B}_{\mathcal{G}}) \\ &+ \gamma_5\operatorname{tr}(\mathbf{B}_{\mathcal{I}}^\top\mathbf{D}_5\mathbf{B}_{\mathcal{I}}) + \sum_{t=1}^{T}\frac{\mu}{2}\left\|\mathbf{W}^t - \mathbf{P}^t + \frac{1}{\mu}\mathbf{\Lambda}^t\right\|_2^2, \end{aligned} \tag{38}$$

where $\mathbf{\Lambda}^t\,(1 \le i \le T)$ is the Lagrangian multiplier for the constraint of $\mathbf{W}^t = \mathbf{P}^t$. The algorithm to solve our objective in Eq. (38) is summarized in Algorithm 3.

IV. Experiment

A. Description of the Experimental Dataset

Data used in the preparation of all our experiments are obtained from the ADNI (adni.loni.usc.edu). We download 1.5 T MRI scans and demographic information for 821 ADNI-1 participants and processed using Freesurfer (version 4, http://surfer.nmr.mgh.harvard.edu/, Boston, MA) and VBM as implemented in SPM5 (www.fil.ion.ucl.ac.uk/spm/, London, UK) as described in work [18]. 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 $192 \times 192 \times 166$ (x, y, z dimensions), to yield a standard voxel size of $1.25 \times 1.25 \times 1.2$ mm. Images were then reconstructed to give a $256 \times 256 \times 166$ matrix and voxel size of approximately $1 \times 1 \times 1.2$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.

TABLE I: Classification results for different modalities of our proposed methods.

Modality	F_1 (AD)	F_1 (MCI)	F_1 (HC)
Only VBM	0.431 ± 0.037	0.393 ± 0.041	0.463 ± 0.071
Only FreeSurfer	0.394 ± 0.063	0.361 ± 0.027	0.427 ± 0.042
VBM+FreeSurfer	0.473 ± 0.043	0.431 ± 0.079	0.493 ± 0.051
VBM+SNP	0.523 ± 0.051	0.466 ± 0.051	0.548 ± 0.031
FreeSurfer+SNP	0.485 ± 0.029	0.383 ± 0.023	0.491 ± 0.032
VBM+FreeSurfer+SNP	**0.677 ± 0.043**	**0.474 ± 0.053**	**0.571 ± 0.044**

The analysis of VBM ~~are~~ performed using previously described methods [19], as implemented in SPM5 (www.fil.ion.ucl.ac.uk/spm/, London, UK). The scans are 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 [18]. GM maps are then normalized to MNI atlas space as $1 \times 1 \times 1$ 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 attempt to use the second MP-RAGE. This was successful for only 1 of 8 cases.

A hippocampal region of interest (ROI) template is 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 [20]. These ROIs are 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, are extracted using FreeSurfer (version 4, http://surfer.nmr.mgh.harvard.edu/, Boston, MA). FreeSurfer is 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.

B. Numerical Performance Comparison

In this section, we explore the performance of the proposed longitudinal hierarchy multilevel logistical classification method. We investigate the overall performance of classification task averaged over all the time points used in the ADNI

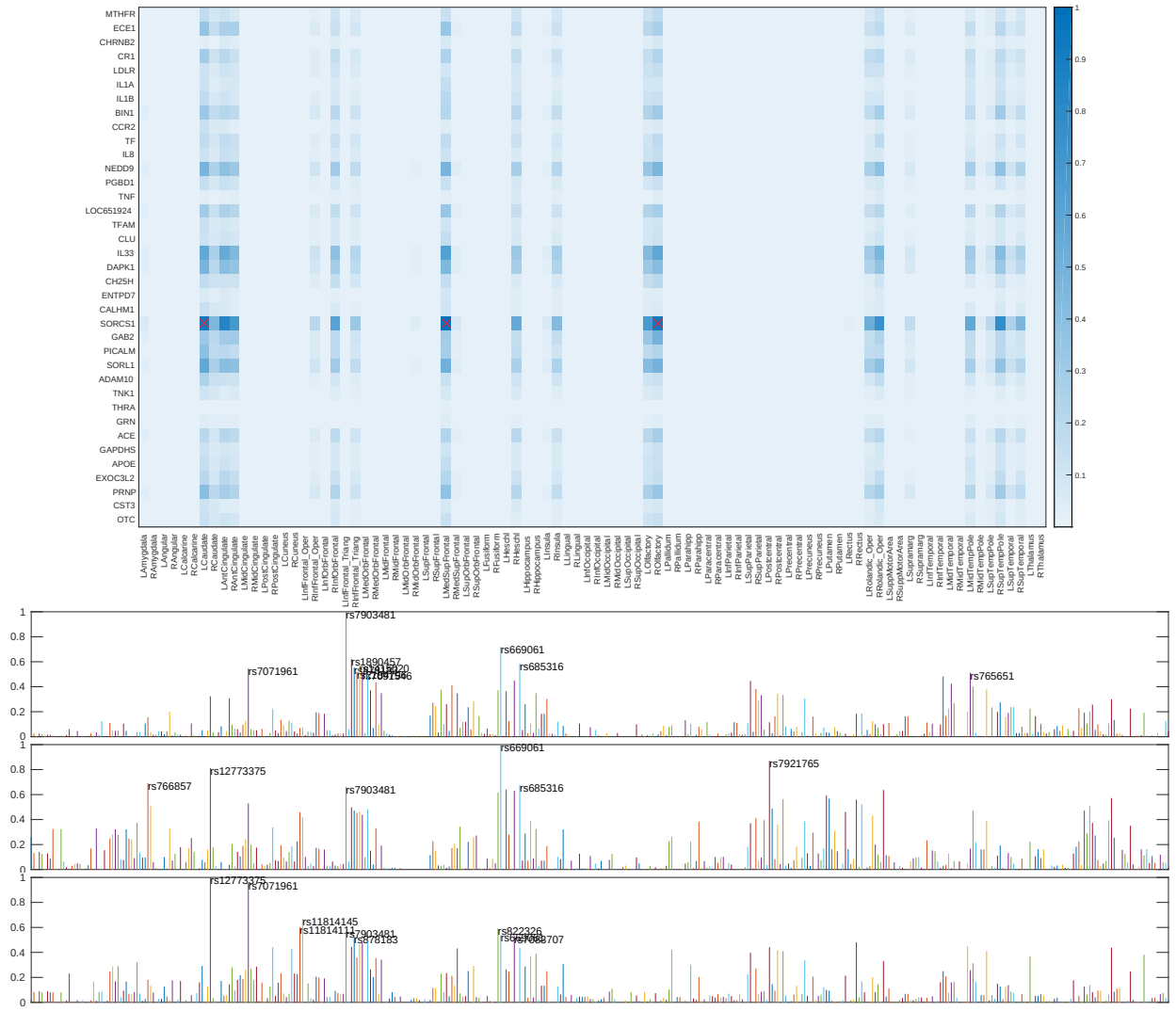

Fig. 2: Visualization of dyadic association between Alzheimer's Gene and imaging markers of VBM. Visualization of top 3 dyadic association between the SNPs in the Alzheimer's gene "SORCS1" and VBM biomarker LMedSupFrontal, LCaudat and ROlfactory respectively.

TABLE II: Classification results using all modalities for different methods

Methods	F_1 (AD)	F_1 (MCI)	F_1 (HC)
Logistic	0.282 ± 0.043	0.383 ± 0.026	0.351 ± 0.045
Random Forest	0.347 ± 0.047	0.403 ± 0.042	0.396 ± 0.040
SVM	0.272 ± 0.033	0.362 ± 0.030	0.384 ± 0.038
KNN	0.334 ± 0.051	0.362 ± 0.030	0.401 ± 0.035
MLP	0.310 ± 0.048	0.422 ± 0.054	0.414 ± 0.040
ElasticNet	0.301 ± 0.052	0.385 ± 0.048	0.369 ± 0.090
TGF [21]	0.487 ± 0.032	0.446 ± 0.042	0.512 ± 0.031
JMMLRC [22]	0.476 ± 0.028	0.403 ± 0.021	0.504 ± 0.032
Our Method - (w/o trace-norm)	0.432 ± 0.017	0.438 ± 0.031	0.501 ± 0.039
Our Method - (w/o group ℓ_1-norm)	0.631 ± 0.021	0.462 ± 0.046	0.539 ± 0.028
Our Method - (w/o group $\ell_{2,1}$-norm)	0.642 ± 0.028	0.471 ± 0.048	0.563 ± 0.037
Our Method	**0.677 ± 0.043**	**0.474 ± 0.053**	**0.571 ± 0.044**

study (BL, M6, M12, and M24). The performance of the classification task is determined by calculating class-specific F_1 scores. Here we note that best performance is reported based on the grid research of $\{10^{-5}, \ldots, 10^{-1}, 1, 10, \cdots, 10^{5}\}$ for each hyperparameter of γ_1, γ_2 and γ_3.

To validate the effectiveness of our proposed method, we compare ours with its degenerative models. We first compare our methods with its non-hierarchical counterparts named "only VBM", "Only FreeSurfer" and "VBM+FreeSurfer". In these experiments, we directly ~~using~~ proposed neuroimaging modality to predict the disease status of the patients and determine the performance of classification through calculating class-specific F_1 scores. From Table I, we can see that the hierarchical framework using genetic data achieves better performance. We then compare our proposed ~~the~~ methods with its hierarchical degenerative models, named "VBM+SNP", "FS+SNP". In these experiments, hierarchical logistic classification is deployed. From Table I, we can see that the hierarchical framework combining all neuroimaging modalities achieve the best performance. This observation can be attributed to the following reasons. First, compared to its non-hierarchical counterparts, our model could benefit from the multilevel framework by integrating the information of genetic data to improve the prediction of methods. Second, compared to hierarchical degenerative models, our model makes use of all neuroimaging ~~model achieve~~ the best performance because the neuroimaging data are obtained by different image technologies, measuring the same brain from different perspectives,

might carry complementary information.

We also compare the proposed method against six baseline classification methods – logistic classification, random forest, support vector machine (SVM), K-nearest neighbour (KNN), Multilayer perceptron (MLP) and logistic regression with elastic net regularization (ElasticNet). Because Logistic, random forest, KNN, MLP and ElasticNet are designed naively to deal with data at static time point, these methods are performed for each cognitive measures at each time point separately. Thus they can not make use of the temporal correlations across tasks. Apart from the comparison with traditional machine learning methods, we compare our method against its degenerative model – without trace norm term, without group ℓ_1-norm term and without group $\ell_{2,1}$-norm term. From Table II, we can see that our method with all regularization terms outperforms its degenerative models, which demonstrates the effectiveness of each regularization term leveraged in our proposed method. Besides the comparison among its degenerative models, we also compare with two longitudinal multi-modal methods – Temporal Group Feature (TGF) method [21] and Joint Multi-Modal Longitudinal Regression and Classification for Alzheimer's Disease Prediction (JMMLRC) [22]. We report the comparison results in Table II and it shows that our new method achieves the best performance, which again demonstrates the effectiveness of our new method.

C. Association Studies between Genetic and Neuroimaging Data

Besides the prediction of the disease status of a patient, another primary goal of our classification analysis is to explore the relationship between genetic and neuroimaging data. From the formulation in Eq. (5), we learn a tensor projection $\mathcal{W}$ which summarizes the dyadic associations between genetic and neuroimaging data from time point 1 to T. Therefore, we plot association weights between each region of VBM and each group SNP from the average of normalized temporal projection $\mathbf{W}$, shown in Figure 2. From Figure 2, we can see that top 3 coefficient weight between of the SNPs in the Alzheimer's Gene and VBM biomarkers are SORCS1–LMedSupFrontal, SORCS1–LCaudat, SORCS1–ROlfactory and SORCS1–LAntCingulate respectively. The visualization of dyadic association between SNP in Gene ORCS1 and VBM biomarkers (LMedSupFrontal, LCaudat, ROlfactory and LAntCingulate) are also shown in Figure 2.

SorCS1 belongs to the sortilin family of vacuolar protein sorting-10 (Vps10) domain-containing proteins, which is genetically associated with AD. The medical research finds that SORL1 expression is decreased in the brains of patients suffering from AD and SORL1 contorls generation of Alzheimer's amyloid-β peptide [23]. In summary, the identified gene group SorCS1 is strongly in accordance with existing medical research findings with regards to AD, which warrant the effectiveness of proposed study.

D. Identification of Disease Relevant Imaging Biomarkers

Apart from the cognitive outcomes prediction task, another primary goal of our regression analysis is to identify a subset of biomarkers which are highly correlated to AD progressions. Thus, we examine the biomarkers identified by the proposed methods encoded by the cognitive scores.

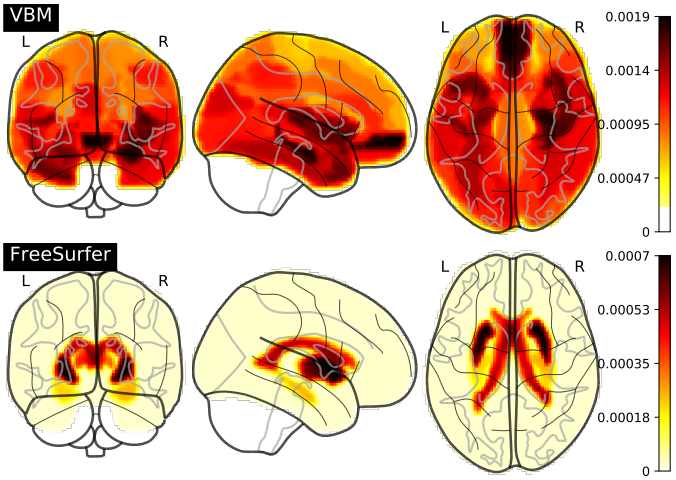

Fig. 3: Weights of imaging markers of VBM.

From the formulation in Eq. (5), we learn a longitudinal projection $\mathbf{B}_{\mathcal{I}}$ which summarizes all the most important biomarkers across time point 1 to T. Therefore, we plot the weights of each region of VBM from average of normalized longitudinal projection $\mathbf{B}_{\mathcal{I}}$, shown in Figure 3. From Figure 3, we can see that the bilateral hippocampus, amygdala regions in VBM and the bilateral cerebral white matter regions in FreeSurfer are found to be in the top selected biomarkers by our model. The hippocampus is a small organ located within the brain's medial temporal lobe and forms an important part of the limbic system, the region that regulates emotions. The hippocampus is associated mainly with memory, in particular long-term memory. The amygdala performs a primary role in the processing of memory, decision-making and emotional response.

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

V. Conclusion

Longitudinal prediction of diseases status of AD patients and Association between genetic and imaging data are two major problems in AD studies. In this paper, we propose a longitudinal multilevel logistic model for simultaneously prediction of cognitive status of patients from phonetic and genotypic data and explicitly learning dyadic association between genetic and neuroimaging data. In the first level, the disease status is directly predicted from the neuroimaging data at time points. In the second level, a parametric model is computed form genetic data to predict the longitudinal disease status. When applied to genetic and imaging data from ADNI dataset, the model is able to highlight brain regions and genes

group which have been verified by medical research, further demonstrating the correctness of our approach for AD study.

VI. Conclusion

Longitudinal prediction of diseases status of AD patients and Association between genetic and imaging data are two major problems in AD studies. In this paper, we propose a longitudinal multilevel logistic model for simultaneously prediction of cognitive status of patients from phonetic and genotypic data and explicitly learning dyadic association between genetic and neuroimaging data. In the first level, the disease status is directly predicted from the neuroimaging data at time points. In the second level, a parametric model is computed form genetic data to predict the longitudinal disease status. When applied to genetic and imaging data from ADNI dataset, the model is able to highlight brain regions and genes group which have been verified by medical research, further demonstrating the correctness of our approach for AD study.

Appendix A
Derivation Detail

1. Updating $\mathbf{W}$: Take the derivation of $\mathcal{J}_1$ with respect to $\mathbf{W}$:

$$\begin{aligned}\frac{\partial \mathcal{J}_1}{\mathbf{W}} &= \frac{1}{N}\sum_{k=1}^{N}\{-y^k\mathbf{x}_{\mathcal{I}}^k(\mathbf{x}_{\mathcal{G}}^k)^\top\mathbf{U}^\top\mathbf{x}_{\mathcal{C}}^k + \frac{m^k}{1+m^k}\mathbf{x}_{\mathcal{I}}^k(\mathbf{x}_{\mathcal{C}}^k)^\top\mathbf{U}^\top\mathbf{x}_{\mathcal{C}}^k\} \\ &\quad + 2\gamma_1\mathbf{D}_1\mathbf{W} + 2\gamma_2\mathbf{D}_2\mathbf{W} + 2\gamma_3\mathbf{D}_3 \\ &= \mathbf{0}, \end{aligned} \tag{39}$$

where $m^k = e^{(\mathbf{x}_{\mathcal{I}}^k)^\top\mathbf{W}\mathbf{x}_{\mathcal{G}}^k\mathbf{U}^\top\mathbf{x}_{\mathcal{C}}^k+(\mathbf{x}_{\mathcal{I}}^k)^\top\beta_{\mathcal{I}}\mathbf{U}^\top\mathbf{x}_{\mathcal{C}}^k+\beta_{\mathcal{G}}^\top\mathbf{x}_{\mathcal{G}}^k+\beta_{\mathcal{C}}^\top\mathbf{x}_{\mathcal{C}}^k}$,

2. Updating $\mathbf{U}$: Take the derivation of $\mathcal{J}_1$ with respect to $\mathbf{U}$:

$$\begin{aligned}\frac{\partial \mathcal{J}_1}{\mathbf{U}} &= \frac{1}{N}\sum_{k=1}^{N}\{-y^k\left((\mathbf{x}_{\mathcal{I}}^k)^\top\mathbf{W}^\top\mathbf{x}_{\mathcal{G}}^k\mathbf{x}_{\mathcal{C}}^k + (\mathbf{x}_{\mathcal{I}}^k)^\top\beta_{\mathcal{I}}\mathbf{x}_{\mathcal{C}}^k\right) \\ &\quad + \frac{m^k}{1+m^k}\left((\mathbf{x}_{\mathcal{I}}^k)^\top\mathbf{W}^\top\mathbf{x}_{\mathcal{G}}^k\mathbf{x}_{\mathcal{C}}^k + (\mathbf{x}_{\mathcal{I}}^k)^\top\beta_{\mathcal{I}}\mathbf{x}_{\mathcal{C}}^k\right)\} \\ &\quad + 2\gamma_3\mathbf{D}_3\mathbf{U} \\ &= \mathbf{0}, \end{aligned} \tag{40}$$

where $m^k = e^{(\mathbf{x}_{\mathcal{I}}^k)^\top\mathbf{W}\mathbf{x}_{\mathcal{G}}^k\mathbf{U}^\top\mathbf{x}_{\mathcal{C}}^k+(\mathbf{x}_{\mathcal{I}}^k)^\top\beta_{\mathcal{I}}\mathbf{U}^\top\mathbf{x}_{\mathcal{C}}^k+\beta_{\mathcal{G}}^\top\mathbf{x}_{\mathcal{G}}^k+\beta_{\mathcal{C}}^\top\mathbf{x}_{\mathcal{C}}^k}$

3. Updating $\beta_{\mathcal{I}}$: Take the derivation of $\mathcal{J}_1$ with respect to $\beta_{\mathcal{I}}$:

$$\begin{aligned}\frac{\partial \mathcal{J}_1}{\beta_{\mathcal{I}}} &= \frac{1}{N}\sum_{k=1}^{N}\{-y^k\mathbf{U}^\top\mathbf{x}_{\mathcal{C}}^k\mathbf{x}_{\mathcal{I}}^k + \frac{m^k}{1+m^k}\mathbf{U}^\top\mathbf{x}_{\mathcal{C}}^k\mathbf{x}_{\mathcal{I}}^k\} + 2\gamma_4\mathbf{D}_4\beta_{\mathcal{I}} \\ &= \mathbf{0}, \end{aligned} \tag{41}$$

4. Updating $\beta_{\mathcal{G}}$: Take the derivation of $\mathcal{J}_1$ with respect to $\beta_{\mathcal{G}}$:

$$\begin{aligned}\frac{\partial \mathcal{J}_1}{\beta_{\mathcal{G}}} &= \frac{1}{N}\sum_{k=1}^{N}\{-y^k\mathbf{x}_{\mathcal{G}}^k + \frac{m^k}{1+m^k}\mathbf{x}_{\mathcal{G}}^k\} + 2\gamma_5\mathbf{D}_5\beta_{\mathcal{G}} \\ &= \mathbf{0}, \end{aligned} \tag{42}$$

5. Updating $\beta_{\mathcal{C}}$: Take the derivation of $\mathcal{J}_1$ with respect to $\beta_{\mathcal{C}}$:

$$\begin{aligned}\frac{\partial \mathcal{J}_1}{\beta_{\mathcal{C}}} &= \frac{1}{N}\sum_{k=1}^{N}\{-y^k\mathbf{x}_{\mathcal{C}}^k + \frac{m^k}{1+m^k}\mathbf{x}_{\mathcal{C}}^k\} + 2\gamma_6\mathbf{D}_6\beta_{\mathcal{C}} \\ &= \mathbf{0}, \end{aligned} \tag{43}$$

Appendix B
Association Studies between Genetic and Neuroimaging Data

In addition to the relationship between SNP group and VBM biomarker, we also plot the association weights between each region of FreeSurfer and each group SNP from the average of normalized temporal projection $\mathbf{W}$, shown in Fig. 4. From Fig. 4, we can see that top 3 coefficient weight between of the SNPs in the Alzheimer's Gene and VBM biomarkers are SORCS1–LHippVol, SORCS1–RHippVol, SORCS1–RAmygVol and SORCS1–RInfFrontal respectively. The visualization of dyadic association between SNP in Gene "ORCS1" and VBM biomarkers (LHippVol, RHippVol, RAmygVol and RInfFrontal) are also shown in Fig. 4.

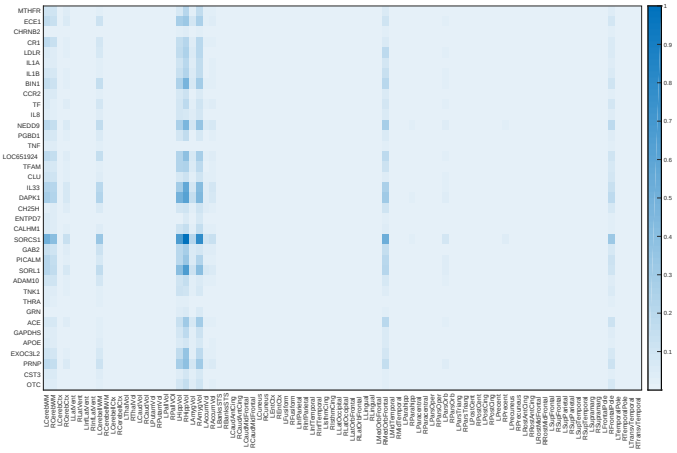

Fig. 4: Top: Visualization of dyadic association between Alzheimer's Gene and imaging markers of FreeSurfer.

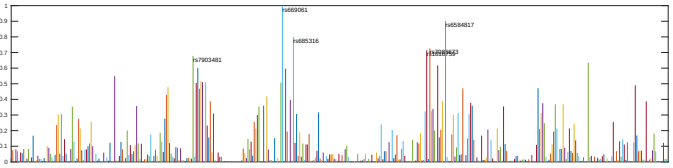

Fig. 5: Visualization of dyadic association between the SNPs in the Alzheimer's gene "SORCS1" and FreeSurfer biomarker LHippVol.

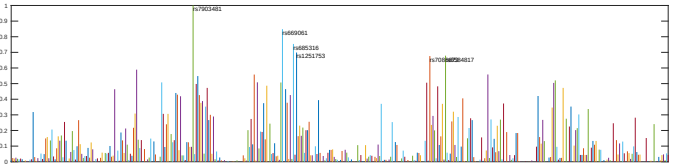

Fig. 6: Visualization of dyadic association between the SNPs in the Alzheimer's gene "SORCS1" and FreeSurfer biomarker RHippVol.

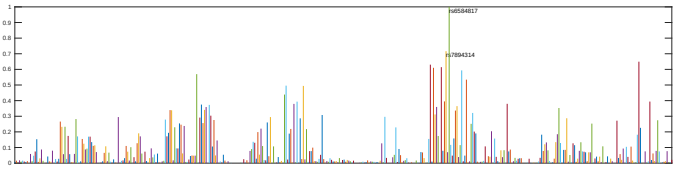

Fig. 7: Visualization of dyadic association between the SNPs in the Alzheimer's Gene "SORCS1" and FreeSurfer biomarker RAmygVol.

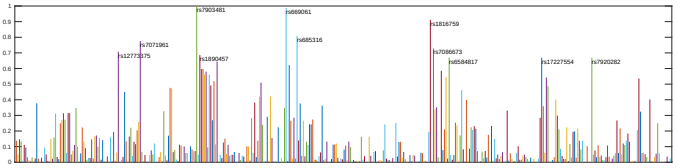

Fig. 8: Visualization of dyadic association between the SNPs in the Alzheimer's Gene "SORCS1" and FreeSurfer biomarker RInfFrontal.

ACKNOWLEDGMENT

The authors first would like to thank the support from 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) for their assistance in collecting and sharing data. In addition, the Canadian Institutes of Health Research funds to support ADNI clinical sites in Canada and the support from dissemination of the ADNI data by the Laboratory for Neuro Imaging at the University of Southern California were highly appreciated as well. 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. This project was supported in part by the National Science Foundation (NSF) under the grants of 1 and 2.

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] L. Shen, P. M. Thompson, S. G. Potkin, L. Bertram, L. A. Farrer, T. M. Foroud, R. C. Green, X. Hu, M. J. Huentelman, S. Kim *et al.*, "Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers," *Brain imaging and behavior*, vol. 8, no. 2, pp. 183–207, 2014.

[3] L. Du, F. Liu, K. Liu, X. Yao, S. L. Risacher, J. Han, A. J. Saykin, and L. Shen, "Associating multi-modal brain imaging phenotypes and genetic risk factors via a dirty multi-task learning method," *IEEE Transactions on Medical Imaging*, 2020.

[4] A. J. Saykin, L. Shen, X. Yao, S. Kim, K. Nho, S. L. Risacher, V. K. Ramanan, T. M. Foroud, K. M. Faber, N. Sarwar *et al.*, "Genetic studies of quantitative mci and ad phenotypes in adni: Progress, opportunities, and plans," *Alzheimer's & Dementia*, vol. 11, no. 7, pp. 792–814, 2015.

[5] 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.

[6] P. Lu, O. Colliot, A. D. N. Initiative *et al.*, "Multilevel modeling with structured penalties for classification from imaging genetics data," in *GRAIL*. Springer, 2017, pp. 230–240.

[7] X. Wang, D. Shen, and H. Huang, "Prediction of memory impairment with mri data: a longitudinal study of alzheimer's disease," in *MICCAI*. Springer, 2016, pp. 273–281.

[8] 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 KDD*, 2011, pp. 814–822.

[9] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 1–45, 2017.

[10] 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 *NeurIPS*, 2012, pp. 1277–1285.

[11] L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi, "Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database," *Nature genetics*, vol. 39, no. 1, pp. 17–23, 2007.

[12] H. Wang, F. Nie, and H. Huang, "Multi-view clustering and feature learning via structured sparsity," in *International conference on machine learning*, 2013, pp. 352–360.

[13] H. Wang, F. Nie, H. Huang, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, and A. D. N. Initiative, "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort," *Bioinformatics*, vol. 28, no. 2, pp. 229–237, 2012.

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

[15] A. Ruhe, "Perturbation bounds for means of eigenvalues and invariant subspaces," *BIT Numerical Mathematics*, vol. 10, no. 3, pp. 343–354, 1970.

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

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] R. F. Lane, J. W. Steele, D. Cai, M. E. Ehrlich, A. D. Attie, and S. Gandy, "Protein sorting motifs in the cytoplasmic tail of sorcs1 control generation of alzheimer's amyloid-β peptide," *Journal of Neuroscience*, vol. 33, no. 16, pp. 7099–7107, 2013.