Modeling Disease Progression via Multisource Multitask Learners: A Case Study With Alzheimer's Disease

Liqiang Nie, Luming Zhang, Lei Meng, Xuemeng Song, Xiaojun Chang, and Xuelong Li, *Fellow, IEEE*

***Abstract*— Understanding the progression of chronic diseases can empower the sufferers in taking proactive care. To predict the disease status in the future time points, various machine learning approaches have been proposed. However, a few of them jointly consider the dual heterogeneities of chronic disease progression. In particular, the predicting task at each time point has features from multiple sources, and multiple tasks are related to each other in chronological order. To tackle this problem, we propose a novel and unified scheme to coregularize the prior knowledge of source consistency and temporal smoothness. We theoretically prove that our proposed model is a linear model. Before training our model, we adopt the matrix factorization approach to address the data missing problem. Extensive evaluations on real-world Alzheimer's disease data set have demonstrated the effectiveness and efficiency of our model. It is worth mentioning that our model is generally applicable to a rich range of chronic diseases.**

***Index Terms*— Disease progression modeling, future health prediction, multisource analysis, source consistency, temporal regularization.**

I. Introduction

CHRONIC disease can be controlled but not cured.[1] It typically lasts over a long duration with slow progression. Some examples of lifelong progressive chronic diseases include stroke, asthma, diabetes, and hypertension. Chronic disease affects the population and wellness system worldwide. As reported by the Centers for Disease Control,[2] chronic diseases are the leading cause of death and disability in the U.S., which account for 70% of all deaths. As a nation, the U.S. spends 86% of healthcare dollars on the treatment of chronic diseases.[3] Data from the World Health Organization[4] show that chronic diseases are also the major cause of premature death around the world even in the places, where infectious diseases are rampant. Although chronic diseases are among the most common and costly health problems, they progress over a long-period time to become fully established, which offers us great opportunities for prevention.

Many clinical measures have been designed to evaluate the disease status and used as essential criteria for clinical diagnosis of probable chronic diseases. For example, mini mental state examination (MMSE) and Alzheimer's disease (AD) assessment scale cognitive subscale (ADAS-Cog) are prevailing to estimate the severity and progression of cognitive impairment of AD [1]–[3]. A comprehensive understanding of the chronic disease progression based on these clinical measures is the key to preventive care and personalized medicine. However, progression modeling for chronic diseases is nontrivial due to the following reasons. First, recent advances in medical technologies have made it popular to collect various complementary types of data of the same patient, which describe his/her disease statuses from different view points. Take the study of AD as an example. Different types of clinical measurements, such as subject characteristics, medical history, genetic information, and imaging data, are usually collected, because their combination can potentially provide a more accurate and rigorous assessment of disease status and likelihood of progression. How to effectively integrate information from multiple heterogeneous sources to comprehensively characterize the given patient is a challenge. Second, a multisource analysis may suffer from the problem of missing data for some specific sources. This is especially the case for expensive measures such as positron emission tomography (PET) scans, where patients have a high chance of dropout or partial attendance in a longitudinal study. Besides, missing data are frequently occurred due to privacy concerns. Last but not least, the progression prediction at each time point is highly correlated. How to identify and model their intrinsic relatedness is of vital importance.

To address the above challenges, there already exist several machine learning efforts dedicated to chronic disease

Manuscript received July 8, 2015; revised November 3, 2015; accepted December 31, 2015. Date of publication February 24, 2016; date of current version June 15, 2017. This work was supported in part by the Qilu Scholar Grant of Shandong University and in part by the National Natural Science Foundation of China under Grant 61572169.

L. Nie is with the School of Computer Science and Technology, Shandong University, Jinan 250100, China (e-mail: nieliqiang@gmail.com).

L. Zhang is with the Department of Electric Engineering and Information System, Hefei University of Technology, Hefei 230009, China (e-mail: zglumg@gmail.com).

L. Meng is with the Joint NTU-University of British Columbia Research Center of Excellence in Active Living for the Elderly, Nanyang Technological University, Singapore 639798 (e-mail: lmeng@ntu.edu.sg).

X. Song is with the School of Computing, National University of Singapore, Singapore 119077 (e-mail: sxmustc@gmail.com).

X. Chang is with the Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: cxj273@gmail.com).

X. Li is with the State Key Laboratory of Transient Optics and Photonics, Center for Optical Imagery Analysis and Learning, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xuelongli@opt.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2016.2520964

[1] http://cmcd.sph.umich.edu/what-is-chronic-disease.html

[2] http://www.cdc.gov/chronicdisease/resources/publications/aag/chronic.htm

[3] http://www.cdc.gov/chronicdisease/

[4] http://www.who.int/chp/chronic_disease_report/contents/part2.pdf

progression modeling. These efforts generally fall into three categories. One is the single source single-task learning. In this context, the disease statuses at different time points are estimated separately by using data from a single source [4]–[6]. Neither the correlation among tasks nor the complementary information across sources is explored. Another line of efforts is the multiple task learning [7], [8]. They formulated the prediction of clinical scores at a sequence of time points as a multitask regression problem, where each task aims to predict a clinical score at one time point. Existing approaches focus on improving the generalization performance by learning multiple related tasks jointly. The relatedness is modeled by assuming that they share either a common representation space or some parameters. The third category of approaches is the multisource learning [9]–[11]. It analyzes how multiple clinical data sources describing the same subjects can be combined to extract more comprehensive information for disease status prediction at one time point. It is noteworthy that the weaknesses of the latter two approaches are the existing multitask learning explores the relatedness among tasks, but disregards the consistency among different sources of a single task; whereas the existing multisource learning ignores the label information from other related tasks.

The problem of progression modeling for chronic diseases exhibits dual heterogeneities: every task in the problem has features from multiple sources, and multiple tasks are related to each other in a chronological sequence. Therefore, by jointly regularizing the relatedness of tasks and sources, multisource multitask (MSMT) learning would be a better choice for chronic disease progression. We propose a novel MSMT regression model to predict the chronic disease progression, because most of the clinical variables are continuous. Our model takes two kinds of prior knowledge into consideration. One is the temporal smoothness. In particular, the sudden changes of disease statuses between neighboring time points should be penalized. The other one is the source consistency. The disagreement among multisources is also penalized, since they are supposed to reflect the same disease status. Our model differs from the traditional multiview learning approaches [12], [13], which mainly focus on semisupervised learning and employ unlabeled data to maximize the agreement between different views. We focus on multisource learning in a supervised setting, which is free from a sufficient amount of unlabeled data. In addition, we utilize a fast matrix factorization (MF) approach to efficiently complete the missing data.

The contributions of this paper are in threefold.

1) We proposed a novel MSMT learning approach to model the chronic disease progression, which regularizes source consistency and temporal smoothness simultaneously.
2) We theoretically proved that our proposed model is a linear model and empirically demonstrated its efficiency.
3) We verified our model on the real-world and public data set of AD.

The remainder of this paper is organized as follows. Sections II and III, respectively, review the related work and detail our proposed disease progression model. Section IV introduces the data preprocessing. Experimental results and analyses are presented in Section V. Finally, the conclusions are drawn in Section VI.

II. Related Work

Medical record search has attracted increasing research attentions from information retrieval communities [14]–[21]. They all aim to return informative knowledge for health seekers to take reactive care. In contrast, predicting disease progression enables patients or clinicians to take proactive management of their health problems. The future health status can be measured by the clinically defined categories [22], [23] or the continuous clinical scores [1], [3], such as MMSE and ADAS-Cog. Broadly speaking, the existing efforts on disease progression modeling can be grouped into three categories: 1) single source single-task learning; 2) multitask learning; and 3) multisource learning.

Single source and/or single-task learning approaches in the past decades dominated the literatures of disease progression modeling. They focused on estimating the disease status separately and usually utilized data from only a single source. Various popular regression models were proposed to predict the target at a single time point with one specific source, such as exploring the magnetic resonance image (MRI) scans to infer the targets at the time point of baseline [3] or in one year [1]. Besides regression models, survival models were introduced in [4] to predict the future disease status of liver transplant patients by considering historical clinical variables individually. In addition, some other approaches [5] considered a small number of input features, and each feature was individually fed into the model to examine its effectiveness. However, when there are a large number of features highly correlated, these approaches are suboptimal. Meanwhile, they neither consider the intrinsic correlation shared among different tasks, nor utilize the complementary information hidden in multiple sources. Their performance is thus far from satisfactory to be clinically useful.

Multitask learning has attracted great attention in the past decades [24]–[28] and were recently proposed to model the disease progression. They discover the commonality among different tasks and simultaneously learn a problem together with other related ones. This often leads to a better model than that of learning the individual tasks separately. The key issue in multitask learning is how to identify and characterize the relatedness among multiple tasks. Two kinds of relatedness have been studied in disease progression modeling. One is that the multiple tasks are assumed to share parameters or prior distributions of the hyperparameters. For example, Zhou *et al.* [7] formulated the prediction of clinical scores at a sequence of time points as a multitask regression problem and captured the intrinsic relatedness among the different tasks by a temporal group lasso (TGL) regularizer. The other way of exploring the intertask relatedness is to assume that they share a common underlying representation. For instance, the work in [8] formulated a novel convex fused sparse group lasso to select the common features for multiple tasks and specific features of individual task in parallel.

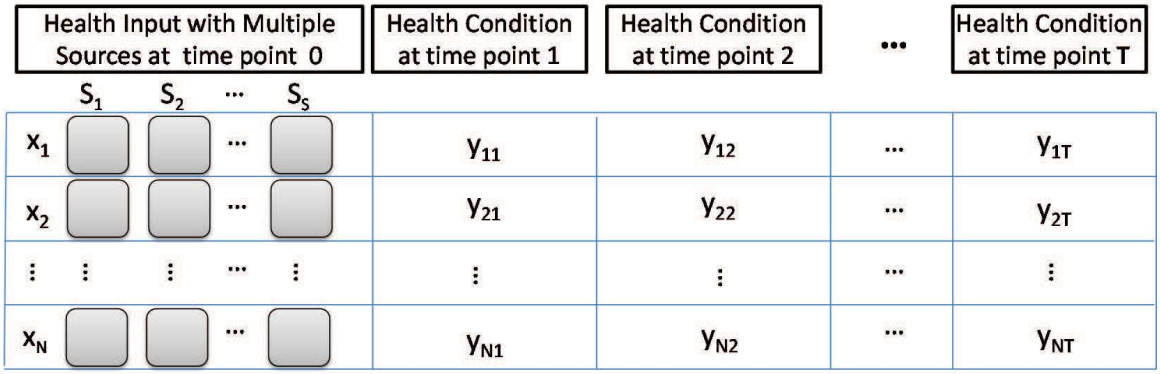


Fig. 1. Illustration of the context of our proposed MSMT learning model. At the baseline time point, we are given N subjects, and each subject is associated with S sources. y_{nt} is the label for the nth patient at time point t.

Multisource learning was initially proposed to integrate data from multiple channels [29]–[32]. It was applied to seamlessly sew clinical data, such as genetic, imaging, and medical history, which is able to improve the accurate and rigorous assessment of the disease status and likelihood of progression. Ye *et al.* [9] proposed a multiple kernel learning method for integrating imaging and nonimaging data for AD study and extended the kernel framework for selecting features from heterogeneous data sources. Experiments showed that the integration of multiple data sources leads to a considerable improvement in the prediction accuracy. One disadvantage of multisource learning is the prevalence of missing data. To address this problem, two novel multisource learning methods [10] were proposed to jointly analyze the incomplete multimodality neuroimaging data, where subjects with missing measures were also kept for training. One year later, Xiang *et al.* [11] presented a bilevel learning model to handle multisource blockwise missing data at both feature level and source level.

However, disease progression modeling exhibits dual heterogeneities. In particular, a single learning task might have features from multiple sources, and multiple learning tasks might be highly correlated by sharing some commonalities. Existing multitask learning or multiview learning algorithms only capture one type of heterogeneities. Zhang and Shen [33] noticed such limitation and proposed a multimodal multitask learning approach. It treats the estimation of different regression and classification variables as different tasks and adopts one existing multitask learning model to learn a common feature subset. Following that, it uses a separate multimodal support vector machine method to fuse these features. Instead of exploring the task relatedness and source relatedness separately, in retrospect, there are a few literatures on unified multisource multitask learning framework [34], but none of them has been applied to disease progression modeling. He and Lawrence [35] proposed a graph-based iterative framework for multiview multitask learning (IteM^2) with its applications to text classification. IteM^2 projects task pairs to a new reproducing kernel Hilbert space based on the common views shared by them. However, it is specifically designed to handle nonnegative feature values. Even worse, as a transductive model, it fails to generate predictive models on independent and unknown samples. To address the intrinsic limitations of transductive models, an inductive multiview multitask learning model regMVMT was introduced in [12]. regMVMT uses coregularization to obtain functions consistent with each other on the unlabeled samples from different views. Across different tasks, additional regularization functions are utilized to ensure that the learned functions are similar. However, simply assuming that all tasks are similar without prior knowledge might be inappropriate. As a generalized model of regMVMT, an inductive convex shared structure learning algorithm for multiview multitask problem (CSL-MTMV) was developed in [13]. As an improvement to regMVMT, CSL-MTMV considers the shared predictive structure among multiple tasks. Noticeably, IteM^2, regMVMT, and CSL-MTMV are all binary classification models, which require nontrivial extensions in order to handle multiclass problems, especially when the number of classes is large. Furthermore, Jin *et al.* [36] pointed out that all previous multiview multitask learning approaches were based on the implicit assumption that all tasks shared a common class label set. Many multiview applications, however, are built upon tasks with different class label sets.

III. Disease Progression Modeling

Let us first define some symbols and notations. In the training set, we assume that we are given N chronic patients $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}]^T$ at the baseline time, and their corresponding disease statuses at the following T time points $\mathbf{Y} = [\mathbf{y^1}, \mathbf{y^2}, \ldots, \mathbf{y^T}] \in \mathbb{R}^{N\times T}$. Each patient is characterized by S complementary sources. For example, the nth patient can be represented by $\mathbf{x_n} = [\mathbf{x_{n1}^T}, \mathbf{x_{n2}^T}, \ldots, \mathbf{x_{nS}^T}]^T$, where $\mathbf{x_{ns}} \in \mathbb{R}^{D_s}$, and D_s denotes the dimensionality of feature space for the sth source. All training samples described by the sth source and by all the sources are, respectively, represented as $\mathbf{X_s} = [\mathbf{x_{1s}}, \mathbf{x_{2s}}, \ldots, \mathbf{x_{Ns}}]^T \in \mathbb{R}^{N\times D_s}$ and $\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_S}] \in \mathbb{R}^{N\times\sum_{s=1}^{S} D_s}$. Fig. 1 shows the context of our model. Our objective is to generalize the disease progression models from training patients to predict the future disease statuses of new patients, given their health information at the baseline time point.

A. Multisource Multitask Learning

We denote $f_s^t(\mathbf{x_{ns}})$ as the predictive function for patient n at time t with the knowledge from source s. We define a linear

predictive function for all patients in a vectorwise form as

$$\mathbf{f_s^t}(\mathbf{X_s}) = \mathbf{X_s}\mathbf{w_s^t} \tag{1}$$

where $\mathbf{w_s^t} \in \mathbb{R}^{D_s}$ is the parameter vector we aim to learn. The disease statuses for all patients at time point t is modeled by averaging the prediction results from all sources

$$\mathbf{f^t}(\mathbf{X}) = \frac{1}{S}\sum_{s=1}^{S}\mathbf{f_s^t}(\mathbf{X_s}) = \frac{1}{S}\sum_{s=1}^{S}\mathbf{X_s}\mathbf{w_s^t}. \tag{2}$$

To model disease progression, we should simultaneously consider two kinds of prior knowledge.

1) *Source Consistency:* We assume that heterogeneous sources of the same patient describe a disease from multiple views, but they should consistently reflect the same disease status. In particular, for a given patient at time point t, the disease status estimated by different sources should be the same or very close.
2) *Temporal Smoothness:* Chronic diseases are the long-term medical conditions that generally progress smoothly.[5] Hence, the sudden change of disease statuses between neighboring time points should be penalized.

Mathematically, we can formulate the above two properties into the following objective function, $O(\mathbf{w_s^t})$:

$$\min_{\mathbf{w_s^t}} \sum_{t=1}^{T}\left\{\frac{1}{2}\left\|\mathbf{y^t} - \frac{1}{S}\sum_{s=1}^{S}\mathbf{X_s}\mathbf{w_s^t}\right\|^2 + \frac{\lambda}{2}\sum_{s=1}^{S}\sum_{s'\neq s}^{S}\left\|\mathbf{X_s}\mathbf{w_s^t} - \mathbf{X_{s'}}\mathbf{w_{s'}^t}\right\|^2 + \frac{\eta}{2}\sum_{s=1}^{S}\left\|\mathbf{w_s^t} - \mathbf{w_s^{t+1}}\right\|^2 + \frac{\mu}{2}\sum_{s=1}^{S}\left\|\mathbf{w_s^t}\right\|^2\right\}. \tag{3}$$

The first term is the widely adopted least square loss function that measures the empirical error on the training data. The second and the third terms control the source consistency and temporal smoothness, respectively; while the last term penalizes the generalization errors. λ and η are parameters that, respectively, regularize the disagreement of heterogeneous sources for the same task and difference between chronologically adjacent tasks on the same sources. μ is a parameter that regulates the strength of the l_2-norm regularization on MSMT learning function.

It is notable that we define $\mathbf{w_s^{T+1}} = \mathbf{0}$ in the temporal smoothness term in (3). As the original formulation of temporal smoothness is infeasible for optimization, we redefine it as

$$\sum_{t=1}^{T}\sum_{s=1}^{S}\left\|\mathbf{w_s^t} - \mathbf{w_s^{t+1}}\right\|^2 = \sum_{s=1}^{S}\left\|\mathbf{W_s}\mathbf{H}\right\|^2 = \sum_{s=1}^{S}\left\|\sum_{t=1}^{T}\mathbf{w_s^t}\mathbf{h_t}^T\right\|^2 \tag{4}$$

where matrix $\mathbf{W_s} = [\mathbf{w_s^1}, \mathbf{w_s^2}, \ldots, \mathbf{w_s^T}] \in \mathbb{R}^{D_s \times T}$ and matrix $\mathbf{H} = [\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_T}]^T \in \mathbb{R}^{T\times(T-1)}$. Matrix $\mathbf{H}$ is precalculated by the following definition:

$$H_{ij} = \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } i = j+1 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

[5]http://www.hpb.gov.sg/HOPPortal/health-article/3396

B. Optimization

By substituting (4) into (3) and taking the derivative of (3) with respect to $\mathbf{w_s^t}$, we have

$$\frac{\partial O}{\partial \mathbf{w_s^t}} = \frac{1}{S}\mathbf{X_s}^T\left(\frac{1}{S}\sum_{s=1}^{S}\mathbf{X_s}\mathbf{w_s^t} - \mathbf{y^t}\right) + \lambda\mathbf{X_s}^T\sum_{s'\neq s}^{S}\left(\mathbf{X_s}\mathbf{w_s^t} - \mathbf{X_{s'}}\mathbf{w_{s'}^t}\right) + \eta\sum_{j=1}^{T}\mathbf{w_s^j}\mathbf{h_j}^T\mathbf{h_t} + \mu\mathbf{w_s^t}. \tag{6}$$

We set (6) as zero and rearrange its elements. We afterward obtain the following equation:

$$\frac{1}{S}\mathbf{X_s}^T\mathbf{y^t} = \left\{\frac{1}{S^2}\mathbf{X_s}^T\mathbf{X_s} + \lambda(S-1)\mathbf{X_s}^T\mathbf{X_s} + \mu\mathbf{I} + \eta\mathbf{h_t}^T\mathbf{h_t}\mathbf{I}\right\}\mathbf{w_s^t} + \left(\frac{1}{S^2} - \lambda\right)\mathbf{X_s^T}\sum_{s'\neq s}^{S}\mathbf{X_{s'}}\mathbf{w_{s'}^t} + \eta\sum_{t'\neq t}^{T}\mathbf{h_t}^T\mathbf{h_{t'}}\mathbf{w_s^{t'}} \tag{7}$$

where $\mathbf{I}$ is the identity matrix. To facilitate the optimization analysis, we define some notations and rewrite (7) in the following form:

$$\mathbf{A_s^t} = \mathbf{B_s^t}\mathbf{w_s^t} + \sum_{s'\neq s}^{S}\mathbf{C_{ss'}}\mathbf{w_{s'}^t} + \sum_{t'\neq t}^{T}\mathbf{D^{tt'}}\mathbf{w_s^{t'}}. \tag{8}$$

By aligning (7) with (8), we derive the following set of equations:

$$\begin{cases} \mathbf{A_s^t} = \dfrac{1}{S}\mathbf{X_s}^T\mathbf{y^t} \\ \mathbf{B_s^t} = \dfrac{1}{S^2}\mathbf{X_s}^T\mathbf{X_s} + \lambda(S-1)\mathbf{X_s}^T\mathbf{X_s} + \mu\mathbf{I} + \eta\mathbf{h_t}^T\mathbf{h_t}\mathbf{I} \\ \mathbf{C_{ss'}} = \left(\dfrac{1}{S^2} - \lambda\right)\mathbf{X_s^T}\mathbf{X_{s'}} \\ \mathbf{D^{tt'}} = \eta\mathbf{h_t}^T\mathbf{h_{t'}}\mathbf{I}. \end{cases} \tag{9}$$

Equations (8) and (9) explicitly imply that we must jointly learn $\mathbf{w_s^t}$ and $\mathbf{w_{s'}^{t'}}$ from a large set of equations, where $s' \neq s$ and $t' \neq t$. After combining the equations for all tasks on all sources, we obtain a linear system (10), as shown at the top the next page. Equivalently, we can represent this linear system as

$$\mathbf{E}\mathbf{w} = \mathbf{a} \tag{11}$$

where each entry in $\mathbf{E} \in \mathbb{R}^{(S\times T)\times(S\times T)}$ is a block matrix. Each block corresponds to a specific task on a specific source, and its size is the dimensionality of the feature extracted from the corresponding source. Similarly, $\mathbf{w}$ and $\mathbf{a}$ are block vectors with $S \times T$ blocks. So far, we have successfully transferred our proposed MSMT learning model to a linear model. Intuitively, if $\mathbf{E}$ is invertible, we can easily derive an analytical solution of $\mathbf{w}$.

In this paper, $\mathbf{E}$ is invertible. Before proving this property, we first introduce three preliminaries.

Preliminary 1: S denotes the number of sources. We are considering multiple sources, and it is thus reasonable to assume that $S \geq 2$. Therefore, $\lambda(S-1) \geq \lambda$, when $\lambda > 0$.

$$\begin{pmatrix} \mathbf{B_1^1} & \mathbf{C_{12}} & \mathbf{C_{13}} & .. & \mathbf{C_{1S}} & \mathbf{D^{12}} & \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & \mathbf{D^{13}} & \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & .... & \mathbf{D^{1T}} & \mathbf{0} & \mathbf{0} & .. & \mathbf{0} \\ \mathbf{C_{21}} & \mathbf{B_2^1} & \mathbf{C_{23}} & .. & \mathbf{C_{2S}} & \mathbf{0} & \mathbf{D^{12}} & \mathbf{0} & .. & \mathbf{0} & \mathbf{0} & \mathbf{D^{13}} & \mathbf{0} & .. & \mathbf{0} & .... & \mathbf{0} & \mathbf{D^{1T}} & \mathbf{0} & .. & \mathbf{0} \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ \mathbf{C_{S1}} & \mathbf{C_{S2}} & \mathbf{C_{S3}} & .. & \mathbf{B_S^1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & .. & \mathbf{D^{12}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & .. & \mathbf{D^{13}} & ... & \mathbf{0} & \mathbf{0} & \mathbf{0} & .. & \mathbf{D^{1T}} \\ \mathbf{D^{21}} & \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & \mathbf{B_1^2} & \mathbf{C_{12}} & \mathbf{C_{13}} & .. & \mathbf{C_{1S}} & \mathbf{D^{23}} & \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & .... & \mathbf{D^{2T}} & \mathbf{0} & \mathbf{0} & .. & \mathbf{0} \\ \mathbf{0} & \mathbf{D^{21}} & \mathbf{0} & .. & \mathbf{0} & \mathbf{C_{21}} & \mathbf{B_2^2} & \mathbf{C_{23}} & .. & \mathbf{C_{2S}} & \mathbf{0} & \mathbf{D^{23}} & \mathbf{0} & .. & \mathbf{0} & ... & \mathbf{0} & \mathbf{D^{2T}} & \mathbf{0} & .. & \mathbf{0} \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & \mathbf{D^{21}} & \mathbf{C_{S1}} & \mathbf{C_{S2}} & \mathbf{C_{S3}} & .. & \mathbf{B_S^2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & .. & \mathbf{D^{32}} & ... & \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & \mathbf{D^{2T}} \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ \mathbf{D^{T1}} & \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & \mathbf{D^{T2}} & \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & ... & \mathbf{D^{T(T-1)}} & \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & \mathbf{B_1^T} & \mathbf{C_{12}} & \mathbf{C_{13}} & .. & \mathbf{C_{1S}} \\ \mathbf{0} & \mathbf{D^{T1}} & \mathbf{0} & .. & \mathbf{0} & \mathbf{0} & \mathbf{D^{T2}} & \mathbf{0} & .. & \mathbf{0} & ... & \mathbf{0} & \mathbf{D^{T(T-1)}} & \mathbf{0} & .. & \mathbf{0} & \mathbf{C_{21}} & \mathbf{B_2^T} & \mathbf{C_{23}} & .. & \mathbf{C_{2S}} \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ \mathbf{0} & \mathbf{0} & .. & \mathbf{0} & \mathbf{D^{T1}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & .. & \mathbf{D^{T2}} & ... & \mathbf{0} & \mathbf{0} & \mathbf{0} & .. & \mathbf{D^{T(T-1)}} & \mathbf{C_{S1}} & \mathbf{C_{S2}} & \mathbf{C_{S3}} & .. & \mathbf{B_S^T} \end{pmatrix} \begin{pmatrix} \mathbf{W_1^1} \\ \mathbf{W_2^1} \\ . \\ . \\ \mathbf{W_S^1} \\ \mathbf{W_1^2} \\ \mathbf{W_2^2} \\ . \\ . \\ \mathbf{W_S^2} \\ . \\ . \\ . \\ \mathbf{W_1^T} \\ \mathbf{W_2^T} \\ . \\ . \\ \mathbf{W_S^T} \end{pmatrix} = \begin{pmatrix} \mathbf{A_1^1} \\ \mathbf{A_2^1} \\ . \\ . \\ \mathbf{A_S^1} \\ \mathbf{A_1^2} \\ \mathbf{A_2^2} \\ . \\ . \\ \mathbf{A_S^2} \\ . \\ . \\ . \\ \mathbf{A_1^T} \\ \mathbf{A_2^T} \\ . \\ . \\ \mathbf{A_S^T} \end{pmatrix} \tag{10}$$

Preliminary 2: We assume that $\mathbf{v_i}$ is an arbitrary block vector with the following property:

$$\sum_{i=1}^{K} \mathbf{v_i^T v_i} + \sum_{i=1}^{K} \sum_{j \neq i}^{K} \mathbf{v_i^T v_j} = \frac{1}{2} \sum_{i=1}^{K} \| \mathbf{v_i} \|^2 + \frac{1}{2} \left\| \sum_{i=1}^{K} \mathbf{v_i} \right\|^2 \geq 0.$$

Preliminary 3: Without loss of generality, we denote $\mathbf{z_i}$ as an arbitrary block vector with the following property:

$$\sum_{i=1}^{K} \mathbf{z_i^T z_i} - \sum_{i=1}^{K} \sum_{j \neq i}^{K} \mathbf{z_i^T z_j} = \frac{1}{2} \| \mathbf{z_1} - \mathbf{z_K} \|^2 + \frac{1}{2} \sum_{i=2}^{K} \| \mathbf{z_i} - \mathbf{z_{i-1}} \|^2 \geq 0. \tag{12}$$

To prove that $\mathbf{E}$ is invertible, we need to prove that $\mathbf{E}$ is a positive definite matrix first. Without loss of generality, we define $\mathbf{g}^T$ as a nonzero $S \times T$ block vector with $\mathbf{g}^T = [\mathbf{g_{11}^T}, \mathbf{g_{21}^T}, \ldots, \mathbf{g_{S1}^T}, \mathbf{g_{12}^T}, \ldots, \mathbf{g_{st}^T}, \ldots, \mathbf{g_{1T}^T}, \ldots, \mathbf{g_{ST}^T}]$. We thus have

$$\begin{aligned} \mathbf{g}^T \mathbf{E} \mathbf{g} &= \sum_{s=1}^{S} \sum_{t=1}^{T} \mathbf{g_{st}^T B_s^t g_{st}} + \sum_{s=1}^{S} \sum_{t=1}^{T} \mathbf{g_{st}^T} \\ &\quad \times \sum_{s' \neq s}^{S} \mathbf{C_{ss'} g_{s't}} + \sum_{s=1}^{S} \sum_{t=1}^{T} \mathbf{g_{st}^T} \sum_{t' \neq t}^{T} \mathbf{D^{tt'} g_{st'}} \\ &= \sum_{s=1}^{S} \sum_{t=1}^{T} \mathbf{g_{st}^T} \left\{ \frac{1}{S^2} \mathbf{X_s^T X_s} + \lambda(S-1) \mathbf{X_s^T X_s} \right. \\ &\qquad \left. + \mu \mathbf{I} + \eta \mathbf{h_t^T h_t I} \right\} \mathbf{g_{st}} \\ &\quad + \sum_{s=1}^{S} \sum_{t=1}^{T} \sum_{s' \neq s}^{S} \mathbf{g_{st}^T} \left\{ \left(\frac{1}{S^2} - \lambda \right) \mathbf{X_s^T X_{s'}} \right\} \mathbf{g_{s't}} \\ &\quad + \eta \sum_{s=1}^{S} \sum_{t=1}^{T} \sum_{t' \neq t}^{T} \mathbf{g_{st}^T h_t^T h_{t'} g_{st'}}. \end{aligned} \tag{13}$$

According to the first preliminary, $S >= 2$, hence, $\lambda(S-1) \geq \lambda$. We can further derive that $\mathbf{g}^T \mathbf{E}\, \mathbf{g}$ is greater than or equal to the following:

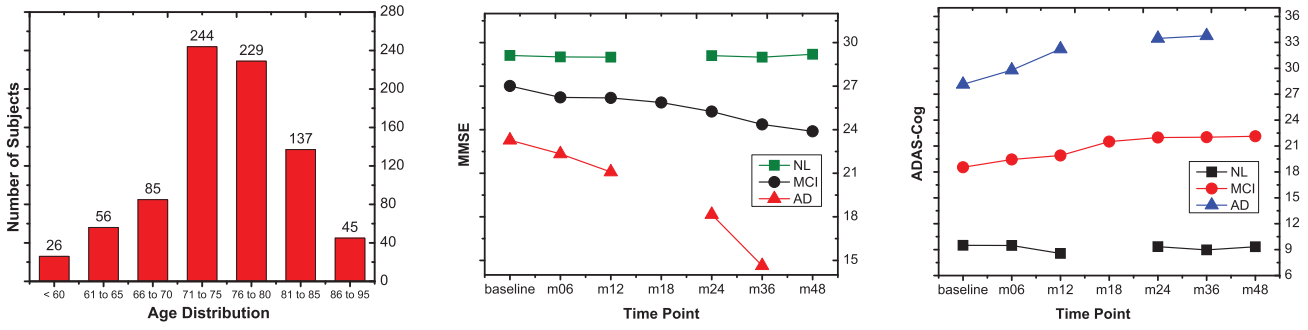
$$\begin{aligned} &\geq \sum_{t=1}^{T} \left\{ \sum_{s=1}^{S} \mathbf{g_{st}^T} \left(\frac{1}{S^2} \mathbf{X_s^T X_s} + \lambda \mathbf{X_s^T X_s} \right) \mathbf{g_{st}} \right. \\ &\qquad \left. + \sum_{s=1}^{S} \sum_{s' \neq s}^{S} \mathbf{g_{st}^T} \left(\frac{1}{S^2} \mathbf{X_s^T X_{s'}} - \lambda \mathbf{X_s^T X_{s'}} \right) \mathbf{g_{s't}} \right\} \\ &\quad + \eta \sum_{s=1}^{S} \left\{ \sum_{t=1}^{T} \mathbf{g_{st}^T h_t^T h_t g_{st}} + \sum_{t=1}^{T} \sum_{t' \neq t}^{T} \mathbf{g_{st}^T h_t^T h_{t'} g_{st'}} \right\} \\ &\quad + \mu \sum_{s=1}^{S} \sum_{t=1}^{T} \mathbf{g_{st}^T g_{st}}. \end{aligned} \tag{14}$$

Let us, respectively, denote block vector $\mathbf{v_s} = (1/S) \mathbf{X_s g_{st}}$, block vector $\mathbf{u_t} = \mathbf{h_t g_{st}}$, and block vector $\mathbf{z_s} = \sqrt{\lambda} \mathbf{X_s g_{st}}$. We can restate the above formulas as follows:

$$\begin{aligned} &= \sum_{t=1}^{T} \left\{ \left\{ \sum_{s=1}^{S} \mathbf{v_s^T v_s} + \sum_{s=1}^{S} \sum_{s' \neq s}^{S} \mathbf{v_s^T v_{s'}} \right\} \right. \\ &\qquad \left. + \left\{ \sum_{s=1}^{S} \mathbf{z_s^T z_s} - \sum_{s=1}^{S} \sum_{s' \neq s}^{S} \mathbf{z_s^T z_{s'}} \right\} \right\} \\ &\quad + \eta \sum_{s=1}^{S} \left\{ \sum_{t=1}^{T} \mathbf{u_t^T u_t} + \sum_{t=1}^{T} \sum_{t' \neq t}^{T} \mathbf{u_t^T u_{t'}} \right\} + \mu \sum_{s=1}^{S} \sum_{t=1}^{T} \mathbf{g_{st}^T g_{st}}. \end{aligned} \tag{15}$$

Based upon the second and third preliminaries, (15) is larger than or equal to the following:

$$\geq \mu \sum_{s=1}^{S} \sum_{t=1}^{T} \mathbf{g_{st}^T g_{st}} = \mu \sum_{s=1}^{S} \sum_{t=1}^{T} \| \mathbf{g_{st}} \|^2 > 0. \tag{16}$$



(a) (b) (c)

Fig. 2. Statistics of our selected data set. (a) Age distribution of the subjects at the baseline time. (b) and (c) Cognitive score progression along time in terms of MMSE and ADAS-Cog. It is worth emphasizing that some time points do not have cognitive scores.

According to the definition of positive definite matrix, we drive that $\mathbf{E}$ is a positive definite matrix. Consequently, $\mathbf{E}$ is invertible.

IV. Data Set

A. Data Collection

To verify the effectiveness and efficiency of our proposed progression model, we conducted experiments on the real-world data sets available from the AD Neuroimaging Initiative[6] (ADNI). In this paper, the date when the patient performs the screening in the hospital for the first time is called baseline, and the time point for the follow-up visits is denoted by the duration starting from the baseline. Take the notation M12 as an example. It denotes the time point 12 months or one year after the first visit.

We verified our proposed disease progression model on AD due to the following reasons.

1) *Severity:* AD is an irreversible, progressive brain disease that slowly destroys memory and thinking skills, and eventually even the ability to carry out the simplest tasks. In 2010, dementia resulted in about 486 000 deaths.
2) *Prevalence:* Worldwide, nearly 36 million people have AD or a related dementia,[7] with a significant increase predicted in the near future if there are no disease altering therapeutics developed [37].
3) *Mystery:* The cause of AD is poorly understood to date, especially the discriminant source of AD.
4) *Accessibility:* The representative AD data are available in ADNI, which is a longitudinal multisite observational study of normal elders (NLs), mild cognitive impairment (MCI), and AD. It has collected various sources of each subject, such as clinical assessment, at multiple time points. All the data are cross-linked and made available to the general scientific community conditioned on official request.

We requested all the data of ADNI-1, which started from 2004. Its 822 participants were recruited from 59 sites across the U.S. and Canada. These include 405 subjects diagnosed with MCI, 188 subjects with AD, and 229 normal healthy control subjects, the so-called NL. Of these subjects, 58.15% are male, and 61.07% of them had been well-educated for at least 16 years. Fig. 2(a) shows their age distribution at the baseline time. It can be seen that the ages of majorities range from 70 to 80 years.

In ADNI-1, we selected 818 subjects (229 NL, 401 MCI, and 188 AD), who all received 1.5T MRI scans at baseline. Therein, 419 subjects have another imaging modality, FDG-PET (PET). In addition to these imaging sources, some subjects also have nonimaging information. In particular, 818, 415, and 566 subjects, respectively, have META information, cerebrospinal fluid (CSF), and proteomics (PROT) measurements. An overview information of the META data is summarized in Table I. Each selected subject has at least two of the five data sources available: MRI and META. Before training and testing each model, we first employed our fast data completion method to complete the missing sources for some specific subjects.

The MRI and PET features were extracted with the image analysis suite FreeSurfer[8] and SPM8 tool,[9] respectively. CSF features were acquired by the ADNI Biomarker Core laboratory at the University of Pennsylvania Medical Center [38], and PROT features were produced by the Biomarkers Consortium Project titled Use of Targeted Multiplex Proteomic Strategies to Identify Plasma-Based Biomarkers in AD. Ultimately, we extracted 624-D features for each subject with all five sources available. Table I shows the subjects, sources, feature types, and feature dimensions of our studied data set.

B. Data Preprocessing

Data missing is highly prevalent in chronic disease data sets. In this paper, we consider two kinds of data missing issues, as shown in Fig. 3. One is missing source. Due to privacy, security, and other concerns, the participants may not provide all their complete health information at the baseline time, such as personal and family medical histories, demographics, and specific body imaging scans. The other one is missing label.

[6]http://adni.loni.usc.edu/
[7]http://www.alzheimers.net/resources/alzheimers-statistics/
[8]http://surfer.nmr.mgh.harvard.edu/
[9]http://www.fil.ion.ucl.ac.uk/spm/

TABLE I

STATISTICS OF SOURCES, SUBJECTS, FEATURE TYPES, AND FEATURE DIMENSIONS IN THIS PAPER. WE SELECTED 818 SUBJECTS IN TOTAL AND EACH SUBJECT HAS AT LEAST TWO SOURCES AT THE SAME TIME: MRI AND META. WE EXTRACTED 624-D FEATURES FOR EACH SUBJECT WITH ALL FIVE SOURCES AVAILABLE. DIM MEANS FEATURE DIMENSION

Sources	Subject #	Feature Types	Dim
MRI	818	average cortical thickness, standard deviation in cortical thickness, volumes of cortical parcellations, volumes of white matter parcellations, total surface area of the cortex.	305
PET	419	116 anatomical volumes of interest (AVOI) and derived average image values from AVOI.	116
CSF	415	levels of beta amyloid 1-42 ($A\beta_{1-42}$), tau protein (Tau), phosphorylated -tau protein 181 ($pTau_{181p}$), two CSF ratios (Tau/$A\beta_{1-42}$ and $pTau_{181p}$/$A\beta_{1-42}$).	5
PROT	566	produced by the Biomarkers Consortium Project "Use of Targeted Multiplex Proteomic Strategies to Identify Plasma-Based Biomarkers in AD".	147
META	818	3-D demographic(age, years of education, gender), 7-D genetic (ApoE -$\varepsilon 4$), 23-D baseline cognitive scores (CDR, FAQ, GDS, etc.), 18-D lab tests (RCT1, RCT11, RCT12, etc.).	51

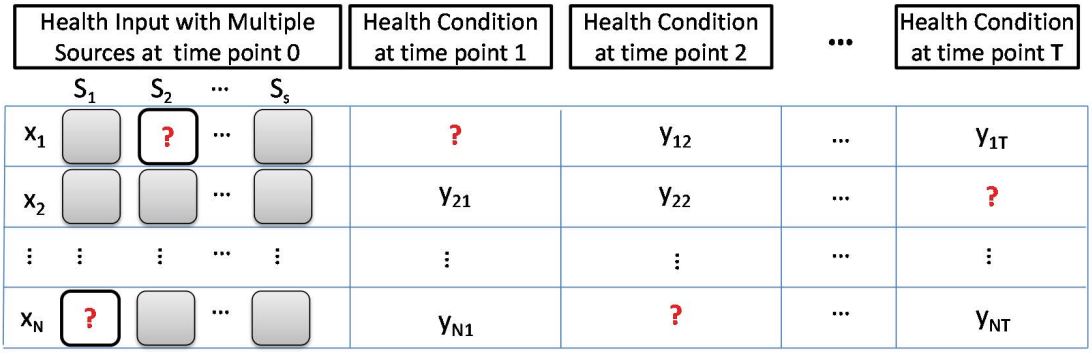


Fig. 3. Illustration of missing data scenarios in our work. One scenario is the missing source, and the other is the missing label.

The health conditions of given patients are periodically evaluated by some standard measures. It is not unusual to take several years to track health status and collect data for chronic diseases. During the long-term and periodic measurement, some patients may die or are partially absent from some time points of health status evaluation, which causes the problem of label missing.

One naive approach is to simply remove the patients with missing items. This results in information loss as patients with partial items will be abandoned. Meanwhile, this way dramatically reduces the training size, which is prone to a suboptimal model. Therefore, it is necessary to complete the missing information beforehand. MF is competent for this assignment, which is able to discover the latent features underlying the interactions between two different kinds of entities, such as users and products they are rating on, and is popular in collaborative recommendation systems [39]–[42].

We denote $\mathbf{M} = [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_S}, \mathbf{Y}] \in \mathbb{R}^{N\times D}$, where $D = T + \sum_s^S D_s$. $\mathbf{X_s}$ and $\mathbf{Y}$, respectively, refer to all patients on the sth source and their health statuses at T time points. Our target is to seek for two matrices $\mathbf{P} \in \mathbb{R}^{N\times L}$ and $\mathbf{Q} \in \mathbb{R}^{L\times D}$, such that their products approximate $\mathbf{M}$

$$\mathbf{M} \approx \hat{\mathbf{M}} = \mathbf{P} \times \mathbf{Q} \tag{17}$$

where matrix $\mathbf{Q}$ is the basis matrix in the latent space, and matrix $\mathbf{P}$ is the latent representation of patients in the latent space. Equation (17) can be intuitively interpreted as follows: the observed instances can be generated by additive combination of underlying set of hidden basis.

To obtain the estimated value $\hat{M}_{ij}$, we multiply the ith row of $\mathbf{P}$ and the jth column of $\mathbf{Q}$. We have

$$\hat{M}_{ij} = \sum_{r=1}^{L} P_{ir} Q_{rj}. \tag{18}$$

The bias between all the estimated and true values over all nonmissing items is formulated as

$$\begin{aligned}\epsilon &= \sum_{ij} e_{ij}^2 + \frac{\gamma}{2}(\| \mathbf{P} \|^2 + \| \mathbf{Q} \|^2) \\ &= \sum_{i}^{n}\sum_{j}^{d}(M_{ij} - \hat{M}_{ij})^2 + \frac{\gamma}{2}(\| \mathbf{P} \|^2 + \| \mathbf{Q} \|^2)\end{aligned} \tag{19}$$

where the regularization is incorporated to avoid overfitting and $\gamma > 0$ is a regularization parameter.

A general algorithm for minimizing the objective function ϵ is a gradient descent. For our problem, the gradient descent leads to the following additive update rules:

$$\begin{cases} P_{ir}^{(t+1)} = P_{ir}^{(t)} - \alpha \dfrac{\partial \epsilon}{\partial P_{ir}} \\ Q_{rj}^{(t+1)} = Q_{rj}^{(t)} - \alpha \dfrac{\partial \epsilon}{\partial Q_{rj}} \end{cases} \tag{20}$$

TABLE II
NUMBER OF SUBJECTS AVAILABLE FOR MMSE AT DIFFERENT TIME POINTS. THE ADAS-COG IS THE SAME. M18 AND M48 ARE NOT CONSIDERED IN THIS PAPER DUE TO ITS SEVERE MISSING LABEL PROBLEM

Type	baseline	M06	M12	*M18*	M24	M36	*M48*
AD	188	176	158	0	139	11	2
MCI	401	384	362	328	304	252	33
NL	229	221	212	0	203	187	56
Total	818	781	732	328	646	450	111

where the derivative results are

$$\begin{cases} \dfrac{\partial \epsilon}{\partial P_{ir}} = -2(M_{ij} - \hat{M}_{ij})Q_{rj} + \gamma P_{ir} \\ \dfrac{\partial \epsilon}{\partial Q_{rj}} = -2(M_{ij} - \hat{M}_{ij})P_{ir} + \gamma Q_{rj} \end{cases} \tag{21}$$

where α is the learning rate. One intuitive solution for the choice of learning rate is to have a constant rate. As long as α is sufficiently small, the above updates should reduce ϵ unless P and Q are at a stationary point. However, it will take a long time to converge. Another simple rule of thumb is to decrease the learning rate over time, $(\alpha_0/1+\tau)$, where α_0 and τ are, respectively, the initial learning rate and the number of epoches. However, they all suffer from the sensitivity of initializations. In this paper, we implement an adaptive learning rate adjuster to monitor and adjust the learning rate α. This adjuster is triggered on each epoch. It will shrink the learning rate if the objective goes up. The idea is that in this case, the learning algorithm is overshooting the bottom of the objective function. On the other hand, the adjuster will increase the learning rate if the objective decreases too slowly. This makes our learning rate parameter less important to the initialized value. The initial value is set as 0.01. Though it is not a very mathematically principled approach, it works well in practice.

V. EXPERIMENTS

A. Experimental Settings

In our experiments, for the given subjects, we predict their future health statuses in terms of MMSE and ADAS-Cog scores using various sources at the baseline. MMSE and ADAS-Cog have been shown to be correlated with the underlying AD pathology and a progressive deterioration of functional ability [2]. A larger MMSE, a lower ADAS-Cog, or both indicate a better health status. Fig. 2(b) and (c) shows the evolution of cognitive scores in terms of MMSE and ADAS-Cog, respectively. For ADAS-Cog cognitive progression, the AD curve grows up very fast, while the MCI curve slightly rises within four years. When it comes to the NL curve, it is relatively stable. A similar pattern can be observed for the MMSE. Table II shows the number of subjects available for MMSE and ADAS-Cog at different time points. In this paper, we do not predict the health status at time point M18 and M48 because of the severe missing data problem. To compared with other prevailing approaches, we measured the regression performance by normalized mean-squared error (nMSE) [43], which is the mean-squared error divided by the variance of the target

$$\text{nMSE} = \frac{\sum_{i=1}(p_i - r_i)^2}{\sum_{i=1}(r_i - \bar{r})^2} \tag{22}$$

where p_i is the predicted value, r_i is the target value, and $\bar{r}$ is the average target value. In addition, we verify the competitors utilizing the correlation coefficient (R-value) between the predicted values and the ground truth [44]

$$R\text{-value} = \frac{\sum_{i=1}(p_i - \bar{p})(r_i - \bar{r})}{\sqrt{\sum_{i=1}(p_i - \bar{p})^2}\sqrt{\sum_{i=1}(r_i - \bar{r})^2}} \tag{23}$$

where $\bar{p}$ is the average predicted value. R-value always takes a value of between -1 and 1, with 1 or -1 indicating perfect correlation. A correlation value close to 0 indicates no association between the variables. A good regression model has a high R-value and low nMSE value.

The results reported in this paper were based on the tenfold cross validation due to the small sample size. It is worth highlighting that subjects with missing labels in the testing set were not utilized to assess our model, even their labels were estimated by our proposed data completion method. Take M36 as an example. 10-cross validation assigns the testing set with approximately 81 subjects. However, only about 45 subjects with real labels were taken for testing.

B. Comparison Among Progression Models

To examine the efficacy of the proposed disease progression model, we comparatively verified the following state-of-the-arts regression models in disease progression domain.

1) *RR:* Ridge regression (RR) is a simple approach to estimate the future health statuses by modeling the tasks at different time points separately [45]. It also assumes that the sources are independent. We can write the ridge constraint as the penalized residual sum of squares, $(\mathbf{y^t}-\mathbf{X}\mathbf{w^t})^2+\delta \parallel \mathbf{w^t} \parallel_F^2$. RR admits an analytical solution given by: $\mathbf{w^t} = (\mathbf{X}^T\mathbf{X} + \delta\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$.
2) *TGL:* Zhou *et al.* [7] proposed a TGL model to predict the disease progression. It captured the intrinsic relatedness among different tasks at different time points by a TGL regularizer. The regularizer consists of two components, including $l_{2,1}$-norm penalty on the regression weight vectors and a temporal smoothness term, which ensures a small deviation between two regression models at successive time points.
3) *cFSGL:* A novel convex fused sparse group lasso (cFSGL) formulation was proposed in [8]. It simultaneously selects task-shared and task-specific features using the sparse group lasso penalty. Meanwhile, it incorporates the temporal smoothness using the fused lasso penalty. The proximal operator associated with the optimization problem exhibits a certain decomposition property and, thus, can be solved effectively.
4) *nFGL:* To reduce the shrinkage bias inherent in the convex formulation, a nonconvex fused group lasso (nFGL) model was introduced to model the disease progression in [8, eq. (17)]. It is a composite $l_{0.5,1}$-like penalty.

TABLE III

COMPARISON OF OUR PROPOSED MODEL AND THE EXISTING STATE-OF-THE-ART BENCHMARKS ON LONGITUDINAL MMSE AND ADAS-COG PREDICTION USING VARIOUS SOURCES. EXPERIMENTAL RESULTS ON OVERALL TASKS AND INDIVIDUAL TASK OF EACH TIME POINT ARE REPORTED IN TERMS OF nMSE AND *R*-VALUE

			RR	**TGL**	**cFSGL**	**nFGL**	**MSMT**
MMSE	All Tasks	R-value	0.7982	0.8464	0.8523	0.8557	**0.8794**
		nMSE	0.1577	0.1416	0.1403	0.1382	**0.1233**
	Individual Task (nMSE)	M06	0.1800	0.1610	0.1586	0.1581	**0.1421**
		M12	0.1706	0.1485	0.1479	0.1462	**0.1270**
		M24	0.1173	0.1072	0.1073	0.1041	**0.0945**
		M36	0.1560	0.1459	0.1434	0.1394	**0.1258**
ADAS-Cog	All Tasks	R-value	0.8485	0.8793	0.8906	0.8928	**0.9094**
		nMSE	0.0433	0.0412	0.0383	0.0389	**0.0347**
	Individual Task (nMSE)	M06	0.0445	0.0408	0.0407	0.0406	**0.0367**
		M12	0.0401	0.0380	0.0352	0.0356	**0.0320**
		M24	0.0420	0.0412	0.0370	0.0377	**0.0336**
		M36	0.0481	0.0469	0.0409	0.0430	**0.0371**

The difference of a convex programming technique was employed to solve the nonconvex formulations.

5) *MSMT:* Our proposed MSMT learning model that jointly regularizes the source relatedness and task relatedness.

For each method mentioned above, there parameters were carefully tuned between 10^{-3} to 10^{3}, and the parameters with the best performance with respect to *R*-value were used to report the final results. In particular, we first equally split our data set into ten subsets. For each round of the tenfold cross validation, we utilized eight subsets as training set, one as the validation set and the rest as the testing set. We built our model on the training set, select the optimal model based on the validation set, and assess the performance of the selected model on the testing set.

The experimental results are shown in Table III. From this table, we can derive some interesting observations. First, nMSE on ADAS-Cog is much smaller as compared against that on MMSE. That is because the variance of ADAS-Cog is very large. ADAS-Cog scores range from 0–70, while MMSE scores range from 0–30. Second, the last four models outperform RR under the same time points and all tasks. This may be caused by the fact that all the models except RR consider the temporal relatedness among tasks. Multiple task learning effectively increases the number of samples by learning multiple related tasks simultaneously, while RR treats all tasks independently. Third, it can be observed that our proposed MSMT is consistently and significantly better than the current publicly disclosed disease progression models in terms of *R*-value and nMSE. The reason may be that none of the benchmark systems model the relatedness among sources. Instead they implicitly assume that the sources are independent. As a consequence, we can conclude that the consistent relatedness among sources is able to reinforce the descriptions of individual sources and, hence, enhance the modeling performance.

One unexpected result is that some learning models on some tasks, where the sizes of real labeled samples are relatively small, perform surprisingly well.[10] For example, the learning performance on M24 is greater than that on M06. After analyzing the testing subject distributions, we found it is reasonable. The percentage of NL is changed from 28.3% in M06 to 31.4% in M24, which results in a smaller variance and, hence, a possible larger nMSE.

In addition, we conducted the analysis of variance (popularly known as the ANOVA) with respect to *R*-value over all the tasks. In particular, we performed paired t-test between our model and each of the benchmarks based on tenfold cross validation. Such analyses were carried out on the results of MMSE and ADAS-Cog, respectively. We found that all the *p*-values are much smaller than 0.05, which shows that the improvements of our proposed disease progression model are statistically significant.

C. On Data Preprocessing

As aforementioned, the missing of sources and labels is not unusual in the data collection of chronic diseases. Tables I and II, respectively, uncover these two kinds of missing data in our studied AD data set. To evaluate the necessity, effectiveness, and efficiency of our proposed data preprocessing, i.e., fast data completion method, we compared the performance of our disease progression model under the following scenarios.

1) *DEL:* We simply eliminated the subjects with either missing sources or missing labels. According to our statistic, only 184 subjects have all the five sources, and 429 subjects have all the labels at M06, M12, M24, and M36. To make matters worse, only 93 subjects simultaneously have all the labels and five sources. In fact, this is insufficient to train effective models.
2) *ZERO:* We assigned zero value to any element that is missing. When the data set was first normalized to have

[10]Though we estimated the missing labels before the training, there exists a certain bias between the estimated labels and the real labels.

TABLE IV

PERFORMANCE COMPARISON AMONG DIFFERENT DATA COMPLETION METHODS. R-VALUE AND nMSE ARE REPORTED UNDER OUR PROPOSED DISEASE PROGRESSION MODEL. TIME REFERS TO THE COMPUTATION TIME OF DATA COMPLETION, WHICH IS MEASURED IN EPOCH

Method	Predicted by MSMT Model				Time
	MMSE		ADAS-Cog		
	R-value	nMSE	R-value	nMSE	
DEL	0.6527	0.1758	0.6893	0.0609	–
ZERO	0.7688	0.1642	0.7946	0.0516	–
KNN	0.8475	0.1429	0.8692	0.0425	–
MF	0.8794	0.1233	0.9094	0.0347	1174
fMF	0.8794	0.1233	0.9094	0.0347	32

zero mean and unit standard deviation, this is equivalent to mean value imputation [10].

3) *KNN:* The k-nearest neighbor (KNN) method replaced the missing value in the data matrix with the corresponding value from the nearest column based on META and MRI sources. Gaussian kernel function [12] was employed to estimate the pairwise similarity.
4) *MF:* The missing data was completed by our proposed MF method without adaptive learning rate adjuster. The learning rate was set as a sufficient small value of 0.00001 to avoid oscillation. This value was fixed across the training process.
5) *fMF:* Our proposed fast MF method with adaptive learning rate adjuster was adopted to infer the missing items.

To ensure fair comparison among various data completion methods, we utilized the same MSMT model after data completion. The comparative results are summarized in Table IV. From this table, it can be seen that DEL obtains the worse performance across different cognitive measures and evaluation criteria. The possible reason is that it ignores a vast amount of useful information and that substantially reduces the training size, while subjects with incomplete data cannot be investigated for disease progression. Moreover, with this approach, the resource and time devoted to those subjects with incomplete data are totally wasted. This set of experiments reflects that the data completion procedure is necessary.

Both of our proposed MF and the KNN methods are superior to ZERO. This is because ZERO does not leverage any self or neighbor content to infer the missing values. KNN does explore the local neighbor information, but overlooks how well the completed matrix globally approximates the original one. The MF methods, though complex, exhibit better performance than others. MF is competitive with fMF with respect to effectiveness. While in the case of efficiency, the former cannot work as faster as the latter. This is because MF is unable to adaptively adjust its learning rate and, thus, takes longer time to reach convergence.

D. On Source Combination

Even though we firmly believe that the integration of multiple sources can enhance the prediction performance of disease progression, it is still of vital importance to quantify how much it is improved as compared against single source via our proposed MSMT learning model. As discussed previously, each subject after data completion in this paper is described by five sources, namely, MRI, PET, CSF, PROT, and META. In other word, if we take every conceivable combination of data sources into consideration, each subject has up to 31 kinds of representation (five for using one source, ten for using two or three source combinations, five for four source combination, and one for integrating all sources). Rather than exhaustively examining all the source combinations, we fed the following representative combinations into our proposed model and validated their description power.

TABLE V

PERFORMANCE COMPARISON AMONG DIFFERENT SOURCE COMBINATION. R-VALUE AND nMSE ARE REPORTED UNDER OUR PROPOSED DISEASE PROGRESSION MODEL. p-VALUE REFERS TO THE SIGNIFICANCE TEST RESULTS (R-VALUE UNDER MMSE)

Sources	Predicted by MSMT Model				p-value
	MMSE		ADAS-Cog		
	R-value	nMSE	R-value	nMSE	
MRI	0.7574	0.1609	0.7688	0.0547	$2.7e$-15
PET	0.7369	0.1675	0.7548	0.0564	$1.4e$-15
CSF	0.7168	0.1694	0.7314	0.0572	$2.5e$-17
PROT	0.7859	0.1582	0.8042	0.0508	$1.1e$-12
META	0.8294	0.1466	0.8450	0.0437	$7.4e$-9
IMG	0.8348	0.1427	0.8482	0.0429	$6.6e$-8
nIMG	0.8687	0.1279	0.8864	0.0394	$2.9e$-3
ALL	**0.8794**	**0.1233**	**0.9094**	**0.0347**	–

1) *MRI, PET, CSF, PROT, and META:* We utilized five sources separately to train and test our proposed disease progression model.
2) *IMG:* Imaging sources comprise of MRI and PET.
3) *nIMG:* Nonimaging sources compose of CSF, PROT, and META.
4) *ALL:* We incorporated all the five sources into our model simultaneously.

Notably, when learning on individual source, our MSMT model degenerates to a multiple task learning model, which only considers the temporal smoothness. The experimental results are comparatively summarized in Table V. It is intuitive that ALL obtains the best results. In addition, IMG and nIMG are effectively ahead of other five individual sources. Such comparison reveals that the description superiority of the combined sources is over individual ones. Meanwhile, we noticed that the model trained on nIMG achieved better performance than that trained on IMG, and that META is the most descriptive source among the five.

The last column of Table V shows the ANOVA analysis. Based on tenfold cross validation, we compared ALL and each of the others in terms of R-value under MMSE. We can see that all the p-values are extremely less than 0.05. This clearly shows that the improvements of source combination are statistically significant.

E. On Time Complexity

The computational complexity of the training process scales as $O(N^2 D_0^3 T^3)$, where N, T and D_0, respectively, refers to

the number of subjects, the number of tasks, and the total feature dimensions over all the sources. Usually, we consider less than ten times points, and hence, T is very small; D_0 is in the order of a few hundreds; we studied 818 subjects. The process can be completed in less than 1 s if we do not take the feature extraction part into account on system with (3.4 GHz and 16-G memory).

VI. Conclusion

This paper studied the progression modeling of chronic diseases by exploring imaging and nonimaging sources at the baseline time. In particular, we formulated the progression prediction as an MSMT regression problem by jointly optimizing: source consistency and temporal smoothness. We theoretically proved that our proposed model is a linear model and empirically demonstrated its efficiency. Before training our model, the MF technique was adopted to alleviate the data missing problem, where the learning rate was tuned adaptively. In addition, we successfully applied our model to the real-world AD sufferers, and it showed superiority over other state-of-the-art approaches.

In this paper, we did not analyze the source confidences and the feature representativeness within each source, which is able to facilitate the discriminant biomarker identification. We plan to explore these two lines of research in the near future.

References

[1] S. Duchesne, A. Caroli, C. Geroldi, D. L. Collins, and G. B. Frisoni, "Relating one-year cognitive change in mild cognitive impairment to baseline MRI features," *NeuroImage*, vol. 47, no. 4, pp. 1363–1370, 2009.

[2] J. R. Petrella, R. E. Coleman, and P. M. Doraiswamy, "Neuroimaging and early diagnosis of Alzheimer disease: A look to the future," *Radiology*, vol. 226, no. 2, pp. 315–336, 2003.

[3] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack, Jr., J. Ashburner, and R. S. Frackowiak, "Predicting clinical scores from magnetic resonance scans in Alzheimer's disease," *NeuroImage*, vol. 51, no. 4, pp. 1405–1413, 2010.

[4] R. K. Pearson, R. J. Kingan, and A. Hochberg, "Disease progression modeling from historical clinical databases," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2005, pp. 788–793.

[5] K. Ito *et al.*, "Disease progression model for cognitive deterioration from Alzheimer's disease neuroimaging initiative database," *Alzheimer's Dementia*, vol. 7, no. 2, pp. 151–160, 2010.

[6] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua, "Disease inference from health-related questions via sparse deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2107–2119, Aug. 2015.

[7] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 814–822.

[8] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1095–1103.

[9] J. Ye *et al.*, "Heterogeneous data fusion for Alzheimer's disease study," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 1025–1033.

[10] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye, "Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1149–1157.

[11] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye, "Multi-source learning with block-wise missing data for Alzheimer's disease prediction," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 185–193.

[12] J. Zhang and J. Huan, "Inductive multi-task learning with multiple view data," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 543–551.

[13] X. Jin, F. Zhuang, S. Wang, Q. He, and Z. Shi, "Shared structure learning for multiple tasks with multiple views," in *Machine Learning and Knowledge Discovery in Databases*, vol. 8189. Heidelberg, Germany: Springer, 2013, pp. 353–368.

[14] S.-H. Yang, R. W. White, and E. Horvitz, "Pursuing insights about healthcare utilization via geocoded search queries," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2013, pp. 993–996.

[15] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, "Bridging the vocabulary gap between health seekers and healthcare knowledge," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 396–409, Feb. 2015.

[16] D. Zhu and B. Carterette, "An adaptive evidence weighting method for medical record search," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2013, pp. 1025–1028.

[17] F. Farfan, V. Hristidis, A. Ranganathan, and M. Weiner, "XOntoRank: Ontology-aware search of electronic medical records," in *Proc. IEEE 25th Int. Conf. Data Eng.*, Mar./Apr. 2013, pp. 820–831.

[18] R. W. White and E. Horvitz, "Captions and biases in diagnostic search," *ACM Trans. Web*, vol. 7, no. 4, 2013, Art. ID 23.

[19] G. Luo, C. Tang, H. Yang, and X. Wei, "MedSearch: A specialized search engine for medical information retrieval," in *Proc. 17th ACM Conf. Inf. Knowl. Manage.*, 2008, pp. 143–152.

[20] L. Nie, T. Li, M. Akbari, J. Shen, and T.-S. Chua, "WenZher: Comprehensive vertical search for healthcare domain," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2014, pp. 1245–1246.

[21] P. Yan, W. Zhang, B. Turkbey, P. L. Choyke, and X. Li, "Global structure constrained local shape prior estimation for medical image segmentation," *Comput. Vis. Image Understand.*, vol. 117, no. 9, pp. 1017–1026, 2013.

[22] C. Misra, Y. Fan, and C. Davatzikos, "Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI," *NeuroImage*, vol. 44, no. 4, pp. 1415–1422, 2009.

[23] R. S. Desikan *et al.*, "Automated MRI measures predict progression to Alzheimer's disease," *Neurobiol. Aging*, vol. 31, no. 8, pp. 1364–1374, 2010.

[24] C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos, "Multitask classification hypothesis space with improved generalization bounds," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1468–1479, Jul. 2015.

[25] C. Li, M. Georgiopoulos, and G. C. Anagnostopoulos, "A unifying framework for typical multitask multiple kernel learning problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1287–1297, Jul. 2014.

[26] A. Rakotomamonjy, R. Flamary, G. Gasso, and S. Canu, "$\ell_p - \ell_q$ penalty for sparse linear and sparse multiple kernel multitask learning," *IEEE Trans. Neural Netw.*, vol. 22, no. 8, pp. 1307–1320, Aug. 2011.

[27] Y. Pan, R. Xia, J. Yin, and N. Liu, "A divide-and-conquer method for scalable robust multitask learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 3163–3175, Dec. 2015.

[28] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1860–1870, Dec. 2013.

[29] J. Liu, Y. Jiang, Z. Li, Z.-H. Zhou, and H. Lu, "Partially shared latent factor learning with multiview data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1233–1246, Jun. 2015.

[30] W. Yang, Y. Gao, Y. Shi, and L. Cao, "MRM-lasso: A sparse multiview feature selection method via low-rank analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 11, pp. 2801–2815, Nov. 2015.

[31] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, and Y. Wen, "Multiview vector-valued manifold regularization for multilabel image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 709–722, May 2013.

[32] L. Nie, L. Zhang, Y. Yang, M. Wang, R. Hong, and T.-S. Chua, "Beyond doctors: Future health prediction from multimedia and multimodal observations," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 591–600.

[33] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of clinical scores in Alzheimer's disease," *Multimodal Brain Image Anal.*, vol. 7012, pp. 60–67, 2011.

[34] X. Mei, Z. Hong, D. Prokhorov, and D. Tao, "Robust multitask multiview tracking in videos," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 11, pp. 2874–2890, Nov. 2015.

[35] J. He and R. Lawrence, "A graph-based framework for multi-task multi-view learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 25–32.

[36] X. Jin, F. Zhuang, H. Xiong, C. Du, P. Luo, and Q. He, "Multi-task multi-view learning for heterogeneous tasks," in *Proc. 23rd ACM Int. Conf. Inf. Knowl. Manage.*, 2014, pp. 441–450.
[37] Alzheimer's Association, "2014 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 10, no. 2, pp. e47–e92, 2014.
[38] N. Tzourio-Mazoyer *et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.
[39] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2001, pp. 556–562.
[40] Y.-L. Zhao, L. Nie, X. Wang, and T.-S. Chua, "Personalized recommendations of locally interesting venues to tourists via cross-region community matching," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, 2014, Art. ID 50.
[41] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 287–296.
[42] K. Zhou, S.-H. Yang, and H. Zha, "Functional matrix factorizations for cold-start recommendation," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2011, pp. 315–324.
[43] Z. Yu and D.-Y. Yeung, "Multi-task learning using generalized *t* process," *J. Mach. Learn. Res.*, vol. 9, pp. 964–971, 2010.
[44] L. Nie, M. Wang, Z.-J. Zha, and T.-S. Chua, "Oracle in image search: A content-based approach to performance prediction," *ACM Trans. Inf. Syst.*, vol. 30, no. 2, 2012, Art. ID 13.
[45] M. Eliot, J. Ferguson, M. P. Reilly, and A. S. Foulkes, "Ridge regression for longitudinal biomarker data," *Int. J. Biostatist.*, vol. 7, no. 1, pp. 1–11, 2011.

Liqiang Nie received the B.E. degree from Xi'an Jiaotong University, Xi'an, China, in 2009, and the Ph.D. degree from the National University of Singapore, Singapore, in 2013.

He is currently a Professor with Shandong University, Jinan, China. Various parts of his work have been published in top forums, including the international ACM SIGIR conference on research and development in Information Retrieval, the international ACM conference on multimedia, international joint conference on artificial intelligence, ACM transactions on information systems (TOIS), ACM transactions on intelligent sand technology (TIST), and the IEEE TRANSACTIONS ON MULTIMEDIA. His current research interests include multiple social network learning and media search.

Dr. Nie serves as the Guest Editor of the ICMR, MMM, the IEEE TRANSACTIONS ON BIG DATA, *Neurocomputing*, and multimedia tools and applications, and a Reviewer of SIGIR, TOIS, MM, and TIST.

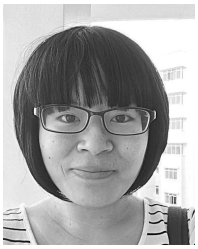
Luming Zhang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China.

He is currently a Professor with the Hefei University of Technology, Hefei, China. He has authored or co-authored several scientific articles at various top venues, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, IEEE Conference on Computer Vision and Pattern Recognition, and ACM MM. His current research interests include multimedia analysis, image enhancement, pattern recognition, weakly supervised learning, image enhancement, and multimedia applications.

Dr. Zhang served as a Guest Editor of *Neurocomputing*, *Signal Processing*, *Multimedia Tools and Applications*, *Multimedia Systems*, and the *Journal of Visual Communication and Image Representation*.

Lei Meng received the B.Eng. degree from the Department of Computer Science and Technology, Shandong University, Jinan, China, in 2010, and the Ph.D. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2015.

He is currently a Research Fellow with Nanyang Technological University. He has authored ten conference and journal papers in top venues, such as the ICMR, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. His current research interests include machine learning, data mining, and social media analytics.



Xuemeng Song received the bachelor's degree from the University of Science and Technology of China, Hefei, China, in 2012. She is currently pursuing the Ph.D. degree with the School of Computing, National University of Singapore, Singapore.

She has authored several papers in the top conferences and journals, such as SIGIR and TOIS. Her current research interests include information retrieval and social network analysis.

Xiaojun Chang is currently pursuing the Ph.D. degree with the University of Technology Sydney, Sydney, NSW, Australia.

His publications appear in proceedings of prestigious international conferences, such as International Conference on Machine Learning, ACM MM, AAAI Conference on Artificial Intelligence, and IJCAI. His current research interests include machine learning, data mining, and computer vision.

Xuelong Li (M'02–SM'07–F'12) is currently a Full Professor with the State Key Laboratory of Transient Optics and Photonics, Center for OPTical IMagery Analysis and Learning, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.