Domain-Knowledge Driven Cognitive Degradation Modeling for Alzheimer's Disease

Ying Lin[1], Kaibo Liu[2], Eunshin Byon[3], Xiaoning Qian[4], Shuai Huang[5*]

ABSTRACT

Cognitive monitoring and screening holds great promises for early detection and intervention of Alzheimer's disease (AD). A critical enabler is the personalized degradation model to predict the cognitive status over time. However, estimating such a model using individual's data faces challenges due to the sparsity and fragmented nature of the cognitive data of each individual. To mitigate this problem, we propose novel methods, called the collaborative degradation model (CDM) together with its extended network regularized version, the NCDM, which can incorporate useful domain knowledge into the degradation modeling. While NCDM results in a difficult optimization problem, we are inspired by existing non-negative matrix factorization methods and develop an efficient algorithm to solve this problem and further provide theoretical results that ensure that the proposed algorithm can guarantee non-increasing property. Both simulation studies and the real-world application to AD are conducted across different degradation models and sampling schemes, which demonstrate the superiority of the proposed methods over existing methods.

1 INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disorder that is characterized by progressive memory loss and other cognitive impairments. Convergent evidences have shown that disease-modifying therapies will be most effective (e.g., for maintaining cognitive health or delaying disease progression) at earliest stages of the disease. With the rapidly developing biomarker measurement technologies, such as the MRI images, the PiB (Pittsburgh Compound B) PET scan, and CSF measurements, the delineation of the early states of AD has emerged as a major scientific priority in the hope to discover therapeutic interventions that can delay progression to AD. For example, one major disease-modifying therapy that holds great promises of preventing AD is anti-amyloid preventative treatment [14, 23].

Despite the promises offered by the biomarkers, the identification of the pre-symptomatic individuals who are going to experience abnormal cognitive decline poses a great challenge in terms of feasibility and cost. For example, although the PiB-PET scan offers a non-invasive in vivo method to detect and quantify brain amyloid deposition, this approach for pre-symptomatic detection is economically challenging for routine use given the current cost and restrictions on reimbursement [27]. Similarly, the clinical use of other useful biomarkers such as $A\beta_{1\text{-}42}$ and phosphorylated tau in cerebral spinal fluid (CSF) is also limited, since lumbar puncture carries risks and is met with resistance in elderly subjects. Furthermore it is unlikely to be used in primary health care centers to routinely screen a large number of participants. Given the cost and limited availability of these brain amyloid measurement techniques, more cost-effective first-line approaches for screening participants at risk of AD are needed [3].

Recently, there is increasing awareness in the AD research community regarding the importance of cognitive monitoring and screening as a cost-effective tool for early detection of AD. The primary objective of the cognitive monitoring and screening for elders is to identify subjects who may have unrecognized cognitive impairment or undiagnosed dementia. After decades of developments on the instruments that can be used to measure cognitive status, the Mini Mental State Examination (MMSE) and the Alzheimer's Disease Assessment Scale Cognitive subscale (ADAS-cog) have been the gold standard for clinical diagnosis of AD and drug effect evaluation in AD drug trials. Recent studies have also revealed the correlation of these cognitive measurements with the underlying biomarker measurements [19, 38]. While those instruments are undergoing continual refinement and development, to operationalize the idea of cognitive monitoring and screening in primary care or community setting, there is a

[1] Industrial & System Engineering, University of Washington, linyeliana.ie@gmail.com
[2] Industrial & System Engineering, University of Wisconsin-Madison, kliu8@wisc.edu
[3] Industrial and Operations Engineering, University of Michigan, ebyon@umich.edu
[4] Department of Electrical & Computer Engineering, Texas A&M University, xqian@ece.tamu.edu
[5] Industrial & System Engineering, University of Washington, shuai.huang.ie@gmail.com, correspondence author

pressing need for developing a personalized degradation model to predict the cognitive status over time. These personalized cognitive degradation models will lay the foundation for optimally allocating the limited healthcare resources to effectively detect the subjects who may experience abnormal cognitive decline [15, 16]. For instance, such a prediction model will enable the design of the repeated measurements over time to be able to detect trajectories of cognitive change, or it will enable the selection of at-risk subjects for early screening, etc.

In this paper, we investigate the statistical challenges and the corresponding modeling solutions that lay in the development of the personalized cognitive degradation models for routine cognitive monitoring and screening. Particularly, we recognize that the major challenge is the sparsity and fragmented nature of the cognitive data of each individual. This is not uncommon in many healthcare studies where it is difficult to collect an abundance of longitudinal data that can cover the whole spectrum of disease progression [12, 41]. Considering the significant complexity of the progression trajectory of the cognitive status that has been reported in the literature [5, 33], without enough data points that cover the whole spectrum of disease progression, it will result in significant bias when predicting in the area with sparse or no training data. An example is shown in Figure 1. Thus, one common approach to mitigate this problem is to merge all the individual data together and estimate a group-level prediction model [26]. However, this model reflects only average effect, failing to capture the personalized variation on the progression trajectory. In this paper, our solution is to incorporate the domain knowledge into the cognitive degradation modeling in the hope of reducing the bias. Particularly, we will focus on two forms of domain knowledge, including 1) the latent cluster structure: it has been discovered in the AD research community that there is a latent cluster structure in the AD population [15, 9, 40]. A rough definition of the cluster structure consists of three groups: the diseased group, the normal aging group (NC), and the mild cognitive impairment group (MCI). Previous studies demonstrated that these three groups have exhibited distinct progression rates on cognitive deterioration [40]. Recent research findings have further revealed that there are more subgroups in the MCI group [9, 34], suggesting a more refined cluster structure. While the subjects between groups may follow significantly different progression trajectories, it is reasonable to believe that the subjects within the same group may follow similar progression trajectories. Therefore, to utilize the cluster structure, we develop a novel model, that is called the collaborative degradation model (CDM). CDM assumes that there are K latent subgroups, while each subgroup has a distinct cognitive degradation model. Since it is usually unknown that which subgroup a subject may belong to, CDM assigns a membership vector (with K elements) to each subject, e.g., denoted as $\boldsymbol{c}_i$ for subject i, while $\boldsymbol{c}_{ij}$ represents the probability of subject i belongs to subgroup j. Then, the degradation model of subject i can be characterized by a weighted combination of the degradation models of the K latent subgroups. 2) The similarity information between subjects: we also extend the CDM formulation to incorporate more domain knowledge, represented as similarity information between subjects. This domain knowledge is common, e.g., it is usually available in AD research to have the individual measurements of some risk factors, including the demographic, social-economical, and even genetic and imaging information, etc. Thus, the similarity between two subjects can be quantified by comparing their profiles on these risk factors. More details of how to obtaining the similarity information based on risk factor measurements will be provided in Section 3.

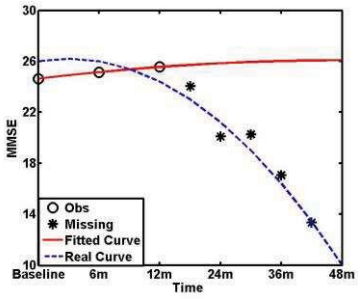


Figure 1: An illustration of the model estimation challenge due to the sparsity of observations

Thus, by incorporating these two forms of domain knowledge together, we expect that the proposed CDM can more effectively utilize the available sparse cognitive degradation data for better modeling. While CDM provides nice intuitive interpretations, it presents a challenging constrained optimization problem. Inspired by existing non-negative matrix factorization methods [6, 10], we then develop corresponding computational algorithm to solve this optimization problem, and we further provide theoretical results that ensure that the proposed algorithm can guarantee the non-increasing property. This paper is organized as follows: in Section 2, we will provide details of the proposed CDM method; in Section 3, we will derive the corresponding computational algorithm that solves the optimization problem of CDM, and provide theoretical results that ensure that the proposed algorithm can guarantee non-increasing property; in Section 4, we will conduct comprehensive simulation studies to demonstrate the efficacy and superiority of the CDM method over a number of existing methods, and demonstrate the CDM method on a real-world dataset on AD; in Section 5, we will provide a conclusion and brief discussion of future work.

2 PROBLEM STATEMENT AND MODEL FORMULATION

In this paper, we concern the problem of building the prediction model for each of the N subjects who are participants in a cognitive monitoring and screening program. We will propose the novel model, CDM, which can exploit the latent structure in the AD population and further incorporate the similarity information between subjects. As mentioned in Section 1, these prediction models are critical for operationalizing the idea of cognitive monitoring and screening in primary care or community setting. For instance, accurate prediction of cognitive status of healthy normal subjects may enable timely detection of the subjects that are subject to risk of cognitive decline. Also, accurate prediction model can help the health care provider to prioritize the screening efforts on the subjects.

To build these prediction models, specifically, with the cognitive status (e.g., measured by ADAS-cog or MMSE) as the outcome of interest, the prediction model aims to capture the predictive relationships between the p risk factors with the cognitive status. For each subject i, we assume that there are longitudinal measurements at n_i time points that include the longitudinal cognitive status measurements, denoted as $\boldsymbol{y}_i = [y_{i1}, \dots, y_{in_i}]^{\mathrm{T}} \in \mathbb{R}^{n_i \times 1}$, and the longitudinal measurements of the p risk factors, denoted as $\mathbf{X}_i = [\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in_i}]^{\mathrm{T}} \in \mathbb{R}^{n_i \times p}$. For simplicity, here, we focus our methodology development on linear models, e.g., we assume that the prediction model of subject i, denoted as $f^i(\boldsymbol{x})$, is a linear model on $\boldsymbol{x}$, i.e., $f^i(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}^i$, where $\boldsymbol{\beta}^i$ is the coefficient vector. Linear models have been found successful in characterizing the dynamic progression of cognitive status, such as the individual growth Curves (IGC) in [13]. Linear models can also be easily extended to capture the nonlinear dynamics by using nonlinear basis functions of $\boldsymbol{x}$ as demonstrated in these applications [21, 28].

Conventional approaches for building the prediction models individually have found difficult in this problem. As mentioned in Section 1 and demonstrated in Figure 1, the sparsity and fragmented nature of the cognitive measurements of individuals elevates the risk of being seriously biased in some areas where measurements are sparse or even null. Thus, the basic idea of CDM is to utilize the underlying cluster structure that has been discovered in many AD studies [7, 15, 20, 31, 37]. Thus, let $g^k(\boldsymbol{x})$ be the degradation model of the latent subgroup k and $\boldsymbol{q}_k$ be the corresponding regression parameters. Recall that each subject i has a membership vector $\boldsymbol{c}_i = [c_{i1}, \dots, c_{iK}]^{\mathrm{T}}$, where c_{ik} represents the probability of subject i belongs to subgroup k. To predict for subject i with any given risk profile, i.e., calculate for $f^i(\boldsymbol{x})$, it is reasonable to use the weighted combination of the predictions of the K degradation models $\sum_k c_{ik} g^k(\boldsymbol{x})$, i.e., $f^i(\boldsymbol{x}) = \sum_k c_{ik} g^k(\boldsymbol{x})$. Apparently, the more similar between subject i with subgroup k (i.e., larger c_{ik}), the more important the degradation model $g^k(\boldsymbol{x})$ will have in determining the value of $f^i(\boldsymbol{x})$. Furthermore, the following Theorem reveals that, on linear models, this modeling assumption also leads to an interesting property as the coefficient vector $\boldsymbol{\beta}^i$ of the model $f^i(\boldsymbol{x})$ can also be represented as a linear combination of the coefficient vectors of the K degradation models.

Theorem 1: With $f^i(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}^i$, $g^k(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{q}_k$, then $\boldsymbol{\beta}^i = \sum_k c_{ik}\boldsymbol{q}_k = \mathbf{Q}\boldsymbol{c}_i$ while $\mathbf{Q} = [\boldsymbol{q}_1, \dots, \boldsymbol{q}_K]$.

Proof: please see the Appendix for a detailed proof.

Based on this model, if we use the least square loss function to measure the goodness-of-fit of the model and consider the constraint that the $\boldsymbol{c}_i$ is normalized and has nonnegative elements, we will have the following optimization formulation of the CDM:

$$(2.1)\qquad \begin{aligned} &\min_{\boldsymbol{c}_i, \mathbf{Q}} \textstyle\sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2, \\ &\text{Subject to } c_{ik} \geq 0, \textstyle\sum_k c_{ik} = 1, \mathbf{X}_i\mathbf{Q} \geq \mathbf{0}, \\ &\forall\, i = 1, \dots, N \text{ and } k = 1, \dots, K. \end{aligned}$$

Here, $\mathbf{X}_i\mathbf{Q} \geq \mathbf{0}$ is imposed due to neuropsychological constraints that the predicted cognitive score should stay nonnegative. $c_{ik} \geq 0, \sum_k c_{ik} = 1$ is imposed due to its definition as a membership vector. We then extend this formulation to incorporate the domain knowledge that is represented as similarity information between subjects, i.e., denoted as w_{jl} for subjects j and l. The similarity w_{jl} can be calculated based on some commonly available measurements of some risk factors, such as the demographic, social-economical, and even genetic and imaging information, etc. More details will be provided in Section 3 for how to calculate w_{jl} using these information. Here, to incorporate the similarity knowledge, we add a regularization term, $\sum_{j,l=1}^{N} \left|\left|\boldsymbol{c}_j - \boldsymbol{c}_l\right|\right|^2 w_{jl}$, into the objective function of (2.1), which leads to the NCDM formulation:

$$(2.2)\qquad \begin{aligned} &\min_{\boldsymbol{c}_i, \mathbf{Q}} \textstyle\sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2 + \lambda \sum_{j,l} ||\boldsymbol{c}_j - \boldsymbol{c}_l||^2 w_{jl}, \\ &\text{Subject to } c_{ik} \geq 0, \textstyle\sum_k c_{ik} = 1, \mathbf{X}_i\mathbf{Q} \geq \mathbf{0}, \\ &\forall\, i = 1, \dots, N \text{ and } k = 1, \dots, K. \end{aligned}$$

Apparently, the regularization term is added to encourage the membership vectors more similar if the similarity between the two subjects is larger. λ is the tuning parameter that is used to control the degree to which the regularization term will affect the parameter

estimation. While solving (2.2) will lead to optimal parameter estimation, it is a constrained optimization problem with non-convex objective function, which has no closed form solution and is difficult to solve by regular gradient-based algorithms. We will present the details of our proposed algorithm in Section 3.

Remark 1: In the literature, a number of transfer learning and multitask learning approaches [25, 26, 45] have been proposed to jointly learn multiple prediction models by treating these models as related. Although the CDM can be generally categorized as a multitask learning approach, one distinct difference of CDM from them is that CDM explicitly exploits the cluster structure embedded in many applications. By allowing the prediction model of each individual to be a weighted combination of the degradation models of the K latent subgroups, CDM essentially enables automatic determination of the relatedness of the prediction models. Furthermore, as a byproduct, the CDM also reveals the clustering structure of the underlying problem and produces the subgroup-level degradation models of the K latent subgroups, leading to valuable domain insights not available in many existing multitask learning methods.

3 PARAMETER ESTIMATION

In this section, we present the proposed algorithm for estimating the parameters $\mathbf{Q}$ and $\mathbf{C} = [\boldsymbol{c}_1, \dots, \boldsymbol{c}_n]$ of NCDM (CDM is a special case of NCDM with $\lambda = 0$). First, we rewrite the formulation in (2.2) in matrix form:

$$\begin{aligned} (3.3)\quad & \min_{\boldsymbol{c}_i, \mathbf{Q}} \textstyle\sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2 + \lambda \mathrm{Tr}(\mathbf{C}^{\mathrm{T}}\mathbf{L}\mathbf{C}) \\ & \text{Subject to } c_{ik} \geq 0, \textstyle\sum_k c_{ik} = 1, \mathbf{X}_i\mathbf{Q} \geq \mathbf{0}, \\ & \forall\, i = 1, \dots, N \text{ and } k = 1, \dots, K. \end{aligned}$$

Here, we used the fact that $\sum_{j,l=1}^{N} \left\|\boldsymbol{c}_j - \boldsymbol{c}_l\right\|^2 w_{jl} = \sum_{j=1}^{N} \boldsymbol{c}_j{}^{\mathrm{T}}\boldsymbol{c}_j d_{jj} - \sum_{j,l=1}^{N} \boldsymbol{c}_j{}^{\mathrm{T}}\boldsymbol{c}_l w_{jl} = \mathrm{Tr}(\mathbf{C}^{\mathrm{T}}\mathbf{L}\mathbf{C})$, where $d_{jj} = \sum_l w_{jl}$, $\mathbf{D}$ is a diagonal matrix with entries $\{d_{jj}, j = 1,2, \dots, N\}$, and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ that is also called the graph Laplacian matrix [17]. We observe that by decoupling the estimation of $\mathbf{Q}$ and $\mathbf{C}$ as two sub-optimization problems, we can derive a simple but efficient iterative algorithm. This is because that the constraints are imposed on $\mathbf{C}$ and $\mathbf{Q}$ separately. Within each iteration, with the latest estimation of $\mathbf{C}$, we can derive a solution of optimizing problem with respect to $\mathbf{Q}$; while with the latest estimation of $\mathbf{Q}$, we can derive an efficient updating algorithm for $\mathbf{C}$.

1) Solve for $\boldsymbol{Q}$ at fixed parameters $\boldsymbol{C}^*$:
With given $\mathbf{C}^*$, the optimization problem (3.3) can be rewritten as

$$\begin{aligned} & \min_{\mathbf{Q}} \textstyle\sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2, \\ & \text{Subject to } \mathbf{X}_i\mathbf{Q} \geq \mathbf{0}\ \forall\, i = 1, \dots, N, \end{aligned}$$

which is essentially a constrained Least Square (LS) problem. Specifically, define $\mathbf{X}_i{}^*$ as

$$\mathbf{X}_i{}^*{}_{(n_i \times Kp)} = \mathbf{X}_i \tilde{\mathbf{C}}_i{}^*,$$

where

$$(3.4)\qquad \tilde{\mathbf{C}}_i{}^* = \begin{bmatrix} \boldsymbol{c}_i{}^{*\mathrm{T}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{c}_i{}^{*\mathrm{T}} \end{bmatrix}_{(p \times Kp)}.$$

The objective function of $\mathbf{Q}$ is

$$\min_{\boldsymbol{q}} \textstyle\sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2,$$

where $\boldsymbol{q}$ is a $Kp \times 1$ vector that is generated by concatenating the columns of the matrix $\mathbf{Q}$. The constraint $\mathbf{X}_i\mathbf{Q} \geq \mathbf{0}$ can be written as $\mathbf{B}_i\boldsymbol{q} \geq \mathbf{0}$ with the definition $\mathbf{B}_{i_{(Kp \times Kp)}} = diag(\mathbf{X}_i, \dots, \mathbf{X}_i)$. Then, the problem in (3.3) can be reformulated as

$$\begin{aligned} (3.5)\quad & \min_{\boldsymbol{q}} \textstyle\sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2, \\ & \text{Subject to } \mathbf{B}_i\boldsymbol{q} \geq \mathbf{0},\ \forall\, i = 1, \dots, N. \end{aligned}$$

This quadratic programming problem can be solved by existing algorithms [32, 39].

2) Solve for $\boldsymbol{C}$ at fixed latent models $\boldsymbol{Q}^*$:
The objective function is

$$\textstyle\sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2 + \lambda \mathrm{Tr}(\mathbf{C}^{\mathrm{T}}\mathbf{L}\mathbf{C}).$$

By introducing the Lagrange multiplier μ_i for constraint $\boldsymbol{c}_i{}^{\mathrm{T}}\mathbf{1} = 1$, the Lagrangian is

$$\begin{aligned} \mathrm{L} = & \textstyle\sum_{i=1}^{N} \boldsymbol{y}_i{}^{\mathrm{T}}\boldsymbol{y}_i - 2\sum_{i=1}^{N} \boldsymbol{y}_i{}^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^*\boldsymbol{c}_i + \\ & \textstyle\sum_{i=1}^{N} \boldsymbol{c}_i{}^{\mathrm{T}}\mathbf{Q}^{*\mathrm{T}}\mathbf{X}_i{}^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^*\boldsymbol{c}_i + \lambda \mathrm{Tr}(\mathbf{C}^{\mathrm{T}}\mathbf{L}\mathbf{C}) + \\ & \textstyle\sum_{i=1}^{N} \mu_i(\boldsymbol{c}_i{}^{\mathrm{T}}\mathbf{1} - 1). \end{aligned}$$

The partial derivative of L with respect to $\boldsymbol{c}_i$ is

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{c}_i} = & -2\mathbf{Q}^{*\mathrm{T}}\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{y}_i + 2\mathbf{Q}^{*\mathrm{T}}\mathbf{X}_i{}^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^*\boldsymbol{c}_i + 2(\lambda\mathbf{L}\mathbf{C})_i + \\ & \mu_i\mathbf{1}. \end{aligned}$$

Using the complementarity condition to enforce the nonnegativity of c_{ik}, $\frac{\partial L}{\partial c_{ik}} c_{ik} = 0$, we get the following equation for c_{ik}

$$(3.6)\qquad -\left(\mathbf{Q}^{*\mathrm{T}}\mathbf{X}_i{}^{\mathrm{T}}\boldsymbol{y}_i\right)_k c_{ik} + \left(\mathbf{Q}^{*\mathrm{T}}\mathbf{X}_i{}^{\mathrm{T}}\mathbf{X}_i\mathbf{Q}^*\boldsymbol{c}_i\right)_k c_{ik} +$$

$$(\lambda \mathbf{LC})_{ik} c_{ik} + \frac{1}{2}\mu_i c_{ik} = 0.$$

Summing order i and using the primary feasibility ${c_i}^T \mathbf{1} = 1$ and $\mathbf{L} = \mathbf{D} - \mathbf{W}$, we have

$$\frac{1}{2}\mu_i = \left(\mathbf{Q}^{*\mathrm{T}} {\mathbf{X}_i}^{\mathrm{T}} \boldsymbol{y}_i\right)^{\mathrm{T}} \boldsymbol{c}_i + \lambda(\mathbf{WC})_i^{\mathrm{T}} \boldsymbol{c}_i - (\mathbf{Q}^{*\mathrm{T}} {\mathbf{X}_i}^{\mathrm{T}} \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i)^{\mathrm{T}} \boldsymbol{c}_i - \lambda(\mathbf{DC})_i{}^{\mathrm{T}} \boldsymbol{c}_i. \tag{3.7}$$

Using the expression of multiplier μ_i in (3.7), (3.6) can be written as

$$\left[\left(\mathbf{Q}^{*\mathrm{T}} {\mathbf{X}_i}^{\mathrm{T}} \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i\right)_k + \lambda(\mathbf{DC})_{ik} + \left(\mathbf{Q}^{*\mathrm{T}} {\mathbf{X}_i}^{\mathrm{T}} \boldsymbol{y}_i\right)^{\mathrm{T}} \boldsymbol{c}_i + \lambda(\mathbf{WC})_i^{\mathrm{T}} \boldsymbol{c}_i\right] c_{ik} - \left[\left(\mathbf{Q}^{*\mathrm{T}} {\mathbf{X}_i}^{\mathrm{T}} \boldsymbol{y}_i\right)_k + \left(\mathbf{Q}^{*\mathrm{T}} {\mathbf{X}_i}^{\mathrm{T}} \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i\right)^{\mathrm{T}} \boldsymbol{c}_i + \lambda(\mathbf{WC})_{ik} + \lambda(\mathbf{DC})_i{}^{\mathrm{T}} \boldsymbol{c}_i\right] c_{ik} = 0.$$

This equation leads to the following updating rule:

$$c_{ik}{}^{m+1} = c_{ik}{}^{m} \frac{(\mathbf{Q}^{*\mathrm{T}} {\mathbf{X}_i}^{\mathrm{T}} \boldsymbol{y}_i + (\lambda \mathbf{WC}^m)_i)_k + (\mathbf{Q}^{*\mathrm{T}} {\mathbf{X}_i}^{\mathrm{T}} \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i{}^m)^{\mathrm{T}} \boldsymbol{c}_i{}^m + \lambda(\mathbf{DC}^m)_i{}^{\mathrm{T}} \boldsymbol{c}_i{}^m}{(\mathbf{Q}^{*\mathrm{T}} {\mathbf{X}_i}^{\mathrm{T}} \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i{}^m + (\lambda \mathbf{DC}^m)_i)_k + \left(\mathbf{Q}^{*\mathrm{T}} {\mathbf{X}_i}^{\mathrm{T}} \boldsymbol{y}_i\right)^{\mathrm{T}} \boldsymbol{c}_i{}^m + \lambda(\mathbf{WC}^m)_i^{\mathrm{T}} \boldsymbol{c}_i{}^m}. \tag{3.8}$$

Thus, we derive an algorithm that iteratively optimize for $\mathbf{C}$ and $\mathbf{Q}$. Figure 2 provides a summary of the overall algorithm that consists of the two steps.

As demonstrated in the numeric studies in Sections 4 and 5, the proposed algorithm is efficient and easy to converge. Besides this empirical evidence, the following Theorem 2 also shows that the objective function is non-increasing by using the proposed algorithm.

Theorem 2. The solution converges to a stationary point using the iterative algorithm in Figure 2.
Proof: Please see the Appendix for a detailed proof.

Note that we need to define a similarity matrix $\mathbf{W}$ to implement the algorithm presented in Figure 2. While sometimes it can be readily available through querying prior knowledge or expert opinion, here, we also adopt some existing approaches that have been found effective in the literature [30], including the 0-1 weighting, Heat Kernel Weighting and Dot-Product Weighting. In our numerical studies on both synthetic datasets and real-world datasets, we found that the heat kernel weighting method consistently lead to satisfactory results. One reason for the superior performance of the heat kernel weighting method is probably because that, comparing with the 0-1 weighting and dot-product weighting, the heat kernel weighting method allows optimal tuning of the similarity by introducing a scaling parameter σ^2, i.e., it is defined as $w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma^2}}$ for subjects i and j. Thus, the similarity between subjects can be automatically determined by model selection methods such as cross-validation, AIC and BIC. Other implementation issues include how to obtain initial values $\mathbf{C}^{(0)}$ and $\mathbf{Q}^{(0)}$. Empirical evidence in the simulation studies (Sec 4.1) and the real-world data analysis (Sec 4.2) shows that solutions from mixed effect models [1] can provide good initial values.

Input: measurements for risk factors and cognitive status on each subject, i.e., $\mathbf{X}_i$ and $\boldsymbol{y}_i$, $i = 1, \dots, N$; initial values for the parameters, $\mathbf{C}^{(0)}$ and $\mathbf{Q}^{(0)}$; regularization matrix, W; tuning parameter, λ; maximal iteration number, M

For $m = 0,1, \dots, M$

1. Transform $\boldsymbol{c}_i{}^m$ to $\tilde{\mathbf{C}}_i{}^m$ by (3.4).
2. Let $\mathbf{X}_i{}^{*m} = \mathbf{X}_i \tilde{\mathbf{C}}_i{}^m$, $\mathbf{B}_i = diag(\mathbf{X}_i, \dots, \mathbf{X}_i)$ and calculate $\boldsymbol{q}^{m+1}$ by solving the quadratic programming problem (3.5).
3. Transform $\boldsymbol{q}^{m+1}$ to $\mathbf{Q}^{m+1}$ by partitioning the $Kp \times 1$ vector to $p \times K$ matrix.
4. Calculate $\mathbf{C}^{m+1}$ by the updating rules in (3.8).

End for

Output: $\{\mathbf{Q}^{(M+1)}, \mathbf{C}^{(M+1)}\}$.

Figure 2: An overview of the proposed algorithm for solving (3.3)

4 NUMERICAL STUDIES

4.1 Simulation Studies We investigate the performances of the proposed CDM, the CDM with networked regularization term (NCDM), the mixed effect models (MEM) [1] (i.e., MEM assumes that $\boldsymbol{\beta}^i$ comes from a multivariate normal distribution), and the trivial method that builds the model for each individual separately (IGM) (i.e., estimate $\boldsymbol{\beta}^i$ independently). We compare these models across various settings of some parameters, such as the number of latent subgroups ($K = 3$ and $K = 5$), the sparsity of the samples (sparse sampling and dense sampling), and the types of degradation models (type 1 model and type 2 model), etc. Here, the sparsity of the samples controls how many samples we can collect in each individual. Throughout our simulation studies, for each individual, we always simulate longitudinal observations on 25 consecutive time points regardless of any other factors such as K or the type of degradation model, etc. Then, in the "dense sampling" scenario, we randomly select m observations (i.e., $m \sim Unif(15,20)$) from the first 20 observations as the training data for each individual; while in the "sparse sampling" scenario, we randomly select m observations (i.e., $m \sim Unif(4,8)$)

from the first 20 observations as the training data for each individual. The last 5 observations are used as testing data for each individual. For the types of degradation models, depending on what are the predictors, there has been a diverse type of degradation models that have been used in AD for predicting the cognitive degradation [42, 43]. Here, we focus our investigation on two typical types of degradation models. One is commonly referred as the disease trajectory model (type 1 model), which uses the age (or polynomial basis functions of the variable age) as the predictor (or predictors) [11, 18, 35]. In our simulation, we adopt the polynomial model. Another type of degradation models (type 2 model) uses risk factors such as the time-dependent biomarkers as the predictors, such as the Disease Progression Score (DPS) used in [4]. We adopt this model to generate the longitudinal measurements of biomarkers.

For any given K and the type of degradation model, we can generate the underlying model for each individual by the following procedure. For example, the underlying model for the individual i at time point t is $y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta}^i + \varepsilon_{it}$, where $\boldsymbol{x}_{it} = [1, t, t^2]$ for type 1 modeland $\boldsymbol{x}_{it} = \left[x_{i1t}, x_{i2t}, \dots, x_{ipt}\right]$ for type 2 model with x_{ijt} being the measurement of the j^{th} biomarker of subject i at time t. Since $\boldsymbol{\beta}^i = \mathbf{Q}\boldsymbol{c}_i$, we can randomly generate the matrix $\mathbf{Q}$ and $\boldsymbol{c}_i$. Specifically, to encourage the clustering structure between the subjects, we tend to generate the membership vector $\boldsymbol{c}_i$ that has a dominant element. Thus, for example, considering the case that there are three clusters. We design three multivariate normal distributions as below:

$$F_1(\boldsymbol{c}) \sim N\left(0, \begin{bmatrix}\sigma^2 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}\right),$$
$$F_2(\boldsymbol{c}) \sim N\left(0, \begin{bmatrix}1 & 0 & 0\\ 0 & \sigma^2 & 0\\ 0 & 0 & 1\end{bmatrix}\right),$$
$$F_3(\boldsymbol{c}) \sim N\left(0, \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & \sigma^2\end{bmatrix}\right).$$

For generating $\boldsymbol{c}_i$, we first use a multinomial distribution to randomly select which multivariate normal distribution will be used, and then, use the selected one to generate a random sample. This random sample can be further normalized to obtain $\boldsymbol{c}_i$. Apparently, the larger the magnitude of σ^2 in F_i, the more dominant the i^{th} element in $\boldsymbol{c}_i$. The results reported in what follows correspond to $\sigma^2 = 100$, while the simulation results are robust to other choices of σ^2. One example of the degradation models (for type 1 model) generated using our simulation procedure is illustrated in Figure 3.

We evaluate the performances of the models based on two major criteria, the parameter estimation and prediction accuracy. For parameter estimation, we calculate the difference between the estimated coefficients with the true coefficients, e.g., $\frac{\sum_{i,j}(\hat{\beta}_j^i - \beta_j^i)^2}{np}$. For prediction accuracy, we compare the normalized mean square error (nMSE) on the testing set, e.g., $\text{nMSE}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{\sum_{i=1}^{t} \|y_i - \hat{y}_i\|_2^2 / \sigma(y_i)}{\sum_{i=1}^{t} n_i}$, where β_j^i is the true coefficient of risk factor j and subject i, $\hat{\beta}_j^i$ is estimated value of β_j^i, $\boldsymbol{y}_i$ are the true cognitive measurements (including all the subjects in the testing dataset) at a single time point, $\hat{\boldsymbol{y}}_i$ are the corresponding predicted values of $\boldsymbol{y}_i$, and n_i is the number of observations at that time point.

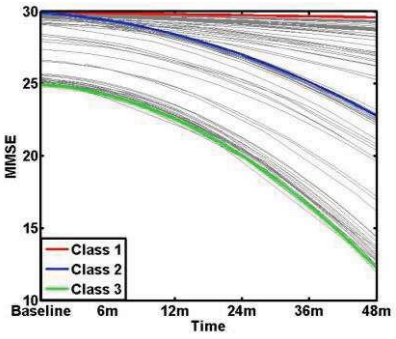


Figure 3: The curves of the randomly generated degradation models of 100 subjects

We summarize the simulation results in Table 1, which correspond to $K = 3$. Similar results can be obtained for other choices of K. Overall, we are able to draw the following observations: 1) the proposed CDM is effective on exploiting the latent structure as demonstrated by its better performance than IGM in terms of both parameter estimation and model prediction. This verifies our hypothesis as demonstrated in Figure 1 that CDM can effectively borrow strengths from subjects that tend to exhibit similar degradation patterns. 2) the NCDM performs best in all the models, which demonstrates that the NCDM can effectively incorporate the similarity information between the subjects to further enhance the model estimation. 3) the advantage of NCDM is larger in the sparse sampling scenario, indicating that the incorporation of prior knowledge will be more preferred when there is a lack of observations.

4.2 Application to Alzheimer`s Disease (AD) While simulation studies have demonstrated the efficacy and accuracy of the proposed CDM and NCDM, here, we demonstrate our proposed methods on a real-world dataset that is collected in the Alzheimer's Disease Neuroimaging Initiative (ADNI) [44]. In this study, we identified a set of 478 subjects who have longitudinal measurements of MMSE (collected at baseline, 12th month, 24th month, 36th month, 48th month and 60th

month), including 104 cognitively normal older individuals (NI), 261 patients with mild cognitive impairment (MCI), and 133 AD patients (AD). Figure 4 depicts the cognitive degradation curves of those 478 subjects, which clearly reveal the cluster structure of these subjects: the MMSE measurements of the NI subjects at different time points maintain at a high level with small fluctuations (blue line), while the MMSE measurements of AD patients degrade dramatically (red line), and the degradation of the MMSE measurements of MCI patients is faster than NI but slower than AD (black line). Among these subjects, 21 have 3 observations, 156 have 4 observations, 244 have 5 observations. Thus, we believe this dataset provides a good example for demonstrating our proposed methods. Particularly, we remove the group information before implementing the proposed methods to demonstrate that the CDM and NCDM can effectively recover the latent structure. We also use the measurements in the 48th month and 60th month as testing data, and the others as training data. In addition, besides the MMSE measurements, for each subject there are also some other baseline measurements of some risk factors. Specifically, we use the ApoE genotypes, the baseline MMSE score, and the baseline regional brain volume measurements extracted from MRI via FreeSurfer [8]. A total of 34 features are used. This information is used in our model to calculate the similarity between the subjects, using the heat kernel method mentioned in Section 3 by cross-validation.

Table 1: Comparison of CDM, MEM, NCDM, and IGM on simulated data when $\boldsymbol{K = 3}$

	Type 1 Model				**Type 2 Model**			
	IGM	CDM	MEM	NCDM	IGM	CDM	MEM	NCDM
Dense Sampling								
MSE	8.143	3.933	3.795	**3.221**	58.050	**37.647**	43.337	40.705
nMSE	0.0287	0.0282	0.020	**0.017**	0.996	0.064	0.131	**0.045**
wR	0.986	0.987	0.990	**0.992**	0.669	0.967	0.933	**0.976**
rMSE1-step	26.182	28.559	24.506	**23.453**	37.542	10.745	12.518	**10.044**
rMSE3-step	38.827	40.772	35.386	**32.293**	32.810	13.939	22.015	**12.235**
rMSE5-step	50.086	45.987	35.613	**33.667**	51.582	10.748	15.226	**7.720**
Sparse Sampling								
MSE	54.969	5.857	6.088	**5.391**	85.192	**57.607**	66.621	60.221
nMSE	20.979	0.233	0.348	**0.181**	2.841	0.696	0.626	**0.385**
wR	0.112	0.896	0.851	**0.917**	0.320	0.705	0.665	**0.809**
rMSE1-step	768.531	87.975	103.979	**78.835**	76.511	50.310	41.597	**33.147**
rMSE3-step	1028.041	111.641	137.560	**96.633**	87.744	42.914	42.914	**31.407**
rMSE5-step	1327.131	131.729	162.059	**115.804**	68.253	28.968	37.220	**18.193**

In our case study, we formulate the cognitive degradation process of subject i using the following model [22]:

$$y_{it} = \sum_k c_{ik}\, q_{k0} + \sum_k c_{ik} q_{k1}\, t + \sum_k c_{ik} q_{k2}\, t^2 + \varepsilon_{it},$$

where y_{it} is the MMSE measurement of subject i at time t. Then, we first investigate if the proposed method can automatically identify the latent structure. Thus, we employ the CDM model and use AIC to automatically select the best model. This model selection procedure is shown in Figure 5 (a), where the AIC value reaches minima when the number of latent model is 3, which is consistent with our prior knowledge of the AD dataset. Actually, the quality of this minimum is well as it is quite different from the AIC values of the neighbor values of K.

Besides the selection of K, we can also observe good performances on the selection of some other parameters in NCDM such as the λ and the scaling parameter in the heat kernel function by cross-validation. Furthermore, Figure 5 (b) also shows that the algorithm for NCDM converges quickly, i.e., with less than 50 iterations.

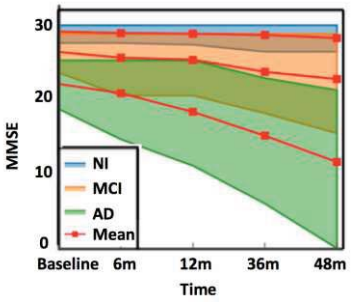
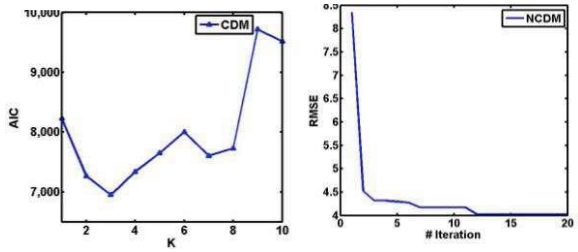


Figure 4: The MMSE curves of the 478 subjects

To compare the performance of the methods (IGM, CDM, MEM, and NCDM), besides the use of the normalized mean square error (nMSE) that has been used in our simulation study, we also use the weighted correlation coefficient (wR) as employed in the AD

literature [28, 43] that is defined as $\mathrm{wR}(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = \frac{\sum_{i=1}^{t} \mathrm{Corr}(\boldsymbol{y}_i, \widehat{\boldsymbol{y}_i})\, n_i}{\sum_{i=1}^{t} n_i}$.



(a) (b)

Figure 5: (a) The AIC values versus $\boldsymbol{K}$ for CDM; (b) convergence performance of the computational algorithm for NCDM

Table 2: The performances of the models (IGM, CDM, MEM, and NCDM) in terms of normalized mean square error (nMSE) and weighted correlation coefficient (wR)

	IGM	CDM	MEM	NCDM
Target: MMSE				
nMSE	1.799	0.936	0.755	**0.531**
wR	0.580	0.618	0.660	**0.716**
M48 rMSE	4.874	4.330	3.705	**3.651**
M60 rMSE	8.326	5.458	5.040	**3.777**

The prediction results are summarized in Table 2. It is clear that the NCDM is the best approach on all the performance evaluations, demonstrating that the proposed method is capable of providing more accurate cognitive prediction models. Also, we can observe that the CDM is better than MEM and IGM, indicating that the CDM is indeed effective on utilizing the cluster structure that is embedded in the AD dataset. It is also clear that the NCDM just slightly deteriorates on predicting on the 60th month, where the performance of the IGM drops dramatically. This shows that the proposed methods are particularly advantageous if long-term prediction/monitoring is desirable. Overall, both our simulation results and the real-world data analysis demonstrate that our methods have better prediction accuracy, and the gain of prediction accuracy has the potential of being translated into higher detection rate of early AD and more cost-effectiveness of the routine cognitive screening and monitoring programs that are currently launching in many primary health care centers.

5 CONCLUSION

In this paper, we propose CDM and NCDM to incorporate domain knowledge into the cognitive degradation modeling of AD. We also develop an efficient computational algorithm to estimate the models, and we further provide theoretical results that ensure that the proposed algorithm can guarantee the non-increasing property. Both simulation studies and AD cognitive data analysis show that the proposed methodology can lead to significantly better performance on both parameter estimation and prediction over existing methods.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation under Grant CMMI-1435584 and Grant CMMI-1435809.

REFERENCE

[1] A. GALECKI, AND T. BURZYKOWSKI, Linear Mixed-Effects Models using R. Springer Texts in Statistics, 2013.

[2] A. P. DEMPSTER, N. M. LAIRD AND D. B. RUBIN, Maximum Likelihood from Incomplete Data via the Em Algorithm, J. Royal Statistics Soc. Series B (Methodological), 39 1977), pp. 1-38.

[3] A. WIMO, B. WINBLAD, AND L. JONSSON, An estimate of the total worldwide societal costs of dementia in 2005, Alzheimer's and Dementia, 3 (2007), pp. 81–91.

[4] B. M. JEDYNAK, ET AL, A computational neurodegenerative disease progression scor, NeuroImage, 63 (2012), pp.1478–1486.

[5] C. A. REYNOLDS, D. FINKLE, N. L. PEDERSEN, Sources of influence on rate of cognitive change over time in Swedish twins, Experimental Aging Research, 28 (2002), pp. 407-433.

[6] C. DING, Y. ZHANG, T. LI, S. R. HOLBROOK, Biclustering protein complex interactions with a biclique finding algorithm, In: Data Mining, ICDM Sixth International Conference, 2006, pp. 178–187.

[7] C. HERTZOG, R. A. DIXON, D. F. HULTSCH AND S. W. MACDONALD, Latent change models of adult cognition, Psychology and Aging, 18 (2003), pp. 755– 769.

[8] C. JR. JACK, M. BERNSTEIN, N. FOX, ET AL., The Alzheimer`s disease neuroimaging initiative (ADNI): MRI methods, J. Magn. Regon. Imaging, 27 (2008), pp. 685-691.

[9] C. R. JR. JACK, D. S. KNOPMAN, W. J. JAGUST, ET AL., Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade, Lancet Neurol, 9 (2010), pp. 119–128.

[10] D. CAI, X. HE, J. HAN, AND T. S. HUANG, Graph regularized nonnegative matrix factorization for data representation, IEEE Transactions, 33 (2011), pp. 1548-1560.

[11] D. HEAD, R. L. BUCKNER, SHIMONY, et al, Differential vulnerability of anterior white matter in nondemented aging with minimal acceleration in dementia, Cerebral Cortex, 14 (2004), pp. 410–423.

[12] D. HEDEKER, AND R. D. GIBBONS, Application of random-effects pattern-mixture models for missing data in longitudinal studies, Psychological Methods, 2 (1997), pp. 64-78.

[13] D. J. FRANCIS, J. M. FLETCHER, K. K. STUEBING, K. C. DAVIDSON, AND N. M. TOMPSON, Analysis of change: Modeling individual growth, Journal of Consulting And Clinical Psychology, 59 (1991), pp. 27-37.

[14] D. J. SELKOE AND D. SCHENK, *Alzheimer's disease:* molecular understanding predicts amyloid-based therapeutics, Annu. Rev. Pharmacol. Toxicol, 43 (2003), pp. 545–584.

[15] D. R. ROYALL, R. PALMER, L. K. CHIODO, Normal rates of cognitive change in successful aging, Journal of Neuropsychological Society, 11 (2005), pp. 899–909.

[16] D. X. RASMUSSON, K. A. CARSON, R. BROOKMEYER, C. KAWAS, AND J. BRANDT, Predicting rate of cognitive decline in probable *Alzheimer's disease*, Brain and Cognition, 31 (1996), pp. 133–147.

[17] F. R. K. CHUNG, Spectral Graph Theory, 1997.

[18] G. BARTZOKIS, D. SULTZER, P. H. LU, K. H. NUECHTERLIEN, J. MINTZ, J. L. CUMMINGS, Heterogeneous age-related breakdown of white matter structural integrity, Neurobiol Aging, 25 (2004), pp. 843–851.

[19] G. CHETELAT, AND J. BARON, Early diagnosis of Alzheimer`s disease: contribution of structural neuroimaging, Neuroimage, 18 (2003), pp. 525-541.

[20] G. H. SUH, Y. S. JU, B. K. YEON, AND A. SHAH, A longitudinal *study of Alzheimer's disease: Rates of cognitive* decline, Journal of Geriatric Psychiatry, 19 (2004), pp. 817–824.

[21] J. ASHFORD, F. SCHMITT, Modeling the time-course of Alzheimer dementia, Current Psychiatry Report, 3 (2001), pp. 20-28.

[22] J. C. BIESANZ, N. DEEB-SOSSA, A. M. AUBRECHT, AND P. J. CURRAN, The role of coding time in estimating and interpreting growth curve models, Psychological Methods, 9 (2004), pp. 30 –52.

[23] J. HARDY, AND D. J. SELKOE, The amyloid hypothesis of *Alzheimer's disease: progress and* problems on the road to therapeutics, Science, 297 (2002), pp. 353–356.

[24] J. WISHART, Growth rate determination in nutrition studies, Biometrika, 30 (1938), pp. 16-28.

[25] J. ZHOU, J. CHEN AND J. YE, MALSAR: multi-task learning via structural regularization. Arizona State University, 2012. http://www.public.asu.edu/~jye02/Software/MALSAR.

[26] J. ZHOU, J. LIU, V. A. NARAYAN and J. YE, Modeling disease progression via multi-task learning, NeuroImage, 78 (2013), pp. 233–248.

[27] K. A. JOHNSON, S, MINOSHIMA, N. I. BOHNEN, ET AL., Appropriate use criteria for amyloid PET: A report of the Amyloid Imaging, Journal of Nuclear Medicine, 54 (2013), 1-16.

[28] K. ITO, ET AL., Disease progression model for cognitive deterioration from Alzheimer`s Disease Neuroimaging Initiative database, Alzheimers Dementia, 6 (2010), pp. 39-53.

[29] L. MOSCONI, M. BRYS, L. GLODZIK-SOBANSKA, S. D. SANTI, H. RUSINEK, M. J. DE LEON, Early detection of Alzheimer's disease using neuroimaging, Exp. Gerontol. 42 (2007), pp. 129–138.

[30] M. BELKIN, AND P. NIYOGI, Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, Advances in Neural Information Processing System, 14 (2001), pp. 585-591.

[31] M. J. SLIWINSKI, S. M. HOFER, ET AL., Modeling memory decline in older adults: The importance of preclinical dementia, Psychology and Aging, 18 (2003), pp. 658–671.

[32] M. L. FISHER, The Lagrangian relaxation method for solving integer programming problems, Management Science, 27 (1981), pp. 1-18.

[33] M. MCGUE, AND K. CHRISTENSEN, Heritability of level of and rate of change in cognitive functioning in Danish twins aged 70 years and older, Experimental Aging Research, 28 (2002), pp. 435– 451.

[34] M. S. ALBERT, S. T. DEKOSKY, D. DICKSON, ET AL., Recommendations from the National Institute on Aging-*Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease*, Alzheimers Dement, 7 (2011), pp. 270–9.

[35] N. RAZ, Aging of the brain and its impact on cognitive performance, In: Handbook of aging and cognition II (Craik FIM, Salthouse TA, eds), 2000, pp. 1–90.

[36] O. L. LOPEZ, W. J. JAGUST, S. T. DEKOSKY, ET AL., Prevalence and classification of mild cognitive impairment in the Cardiovascular Health Study Cognition Study, Arch Neurol, 60 (2003), pp. 1385-1389.

[37] P. A. WILKOSZ, H. J. SELTMAN, B. DEVLIN, E. A. WEAMER, O. L. LOPEZ, S. T. DEKOSKY, AND R. A. SWEET, Trajectories of cognitive decline in Alzheimer`s disease, Int. Psychogeriatric, 22 (2010), pp. 281-290.

[38] P. THOMPSOM, ET AL., Mapping hippocampal and ventricular change in AD, Neuroimage, 22 (2004), pp. 1754-1766.

[39] R. BRO, S. DE JONG, A fast non-negative-constrained least squares algorithm. J. Chemometrics, 11 (1997), pp. 393-401.

[40] R. C. PETERSEN, G. E. SMITH, S. C. WARING, ET AL., Mild cognitive impairment: clinical characterization and outcome. Archives of Neurology, 56 (1999), pp. 303-308.

[41] R. J. A. LITTLE, Modeling the dropout mechanism in repeated measures studies, Journal of American Statistical Association, 90 (1995), pp. 1112-1121.

[42] R. PEARSON, R. KINGAN, A. HOCHBERG, Disease progression modeling from historical databases, KDD 2005.

[43] S. DUCHESNE, A. CAROLI, C. GEROLDI, D. COLLINS, G. FRISONI, Relating one-year cognitive change in mild cognitive impairment to baseline MRI features, Neuroimage, 47 (2009), pp. 1363-1370.

[44] S. G. MUELLER, M. W. WEINER, L. J. THAL, ET AL., The Alzheimer's Disease Neuroimaging Initiative, Neuroimaging Clin North Am., 15 (2005), pp. 869–877.

[45] S. J. PAN, AND Q. YANG, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering, 22 (2010), pp. 1345-1359.