A Collaborative Learning Framework for Estimating Many Individualized Regression Models in a Heterogeneous Population

Ying Lin, Kaibo Liu, *Member, IEEE*, Eunshin Byon, Xiaoning Qian, Shan Liu, and Shuai Huang, *Member, IEEE*

***Abstract*—Mixed-effect models (MEMs) have been found very useful for modeling complex dataset where many similar individualized regression models should be estimated. Like many statistical models, the success of these models builds on the assumption that a central tendency can effectively establish the population-level characteristics and covariates are sufficient to characterize the individual variation as derivation from the center. In many real-world problems, however, the dataset is collected from a rather heterogeneous population, where each individual has a distinct model. To fill in this gap, we propose a collaborative learning framework that provides a generic methodology for estimating a heterogeneous population of individualized regression models by exploiting the idea of "canonical models" and model regularization. By using a set of canonical models to represent the heterogeneous population characteristics, the canonical models span the modeling space for the individuals' models, e.g., although each individual model is distinct, its model parameter vector can be represented by the parameter vectors of the canonical models. Theoretical analysis is also conducted to reveal a connection between the proposed method and the MEMs. Both simulation studies and applications on Alzheimer's disease and degradation modeling of turbofan engines demonstrate the efficacy of the proposed method.**

***Index Terms*—Collaborative learning, degradation modeling, regression, sparse and irregular measurements.**

I. Introduction

THIS study concerns the problem of estimating many individualized regression models in a heterogeneous population where each individual has a distinct regression model. Regression models have been popular tools in a wide range of reliability and prognostic tasks. A classic example is to model the relationship between demographical and/or intellectual variables with students' academic achievements in a school district where a number of regression models may be needed since each school may require a different model. The heterogeneous population is also observed in many applications, such as degradation modeling in many engineering systems and healthcare problems. Accurately predicting the health degradation on each individual has the potential to enable better decision making in clinical practice and engineering processes, including early prevention of disease onset and engine failure, as well as efficient monitoring and maintenance (treatment) strategy design.

One approach to model heterogeneous individuals is to estimate the regression model separately for each individual. This approach, however, could be less effective because the information from others is not exploited. Moreover, in many real-world studies, data on individuals could be unevenly distributed, i.e., some individuals may have many data while others' data are sparse. Consequently, measurements from an individual may not be sufficient to build an accurate prognostic model. Another common approach is to estimate a prediction model for all individuals by ignoring their individual's variations. However, this approach will only characterize the average effect, failing to capture the between-individual variations.

A more advanced treatment to mitigate the limitations of these two approaches is to use the mixed-effect model (MEM) [1], also known as hierarchical models [2] and multilevel models [3]. The MEM assumes that the regression parameters of these regression models are sampled from a distribution model, for example, a multivariate normal distribution. The MEM has been widely used to address the individual-to-individual (or unit-to-unit) variations in various applications [4]–[7]. The MEM approach could be effective when there is a central tendency and the heterogeneous patterns of individuals (i.e., the deviations from the central tendency) are randomly distributed. It, however, will be less effective where a mixture of central tendency exists that presents a complex heterogeneous structure. For instance, it is common in many degradation studies that a complex engineering system may degrade following a variety of degradation trajectories, corresponding to different failure modes. It is also common in many prognostic modeling studies in healthcare. For instance, in Alzheimer's disease research, it has been discovered that, roughly, there are three la-

Manuscript received April 16, 2017; revised July 2, 2017; accepted September 30, 2017. Date of publication December 7, 2017; date of current version March 1, 2018. This work was supported by the National Science Foundation under Grant CMMI-1536398 and Grant CMMI-1435809. Associate Editor: R. Kang. *(Corresponding author: Shuai Huang.)*

Y. Lin, S. Liu, and S. Huang are with the Department of Industrial and Systems Engineering, University of Washington, Seattle, WA 98133 USA (e-mail: liny90@uw.edu; liushan@uw.edu; shuaih@uw.edu).

K. Liu is with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI 53706 USA (e-mail: kliu8@wsic.edu).

E. Byon is with the Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: ebyon@umich.edu).

X. Qian is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: xqian@ece.tamu.edu).

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/TR.2017.2767941

tent phenotypes in the Alzheimer's disease population [8]–[11]. Each group shows a distinct degradation pattern on cognitive deterioration.

To efficiently estimate a heterogeneous population of degradation processes, we developed a general statistical learning framework, collaborative learning, and applied it to the contexts of cognitive decline in Alzheimer's disease and health degradation of turbofan engines. The basic idea of the proposed collaborative learning framework is to use a set of canonical models to represent the heterogeneous population characteristics [12]. The proposed collaborative model (CM) assumes that there are K canonical models, in which each degradation mechanism can be represented by a canonical model. In practice, it is usually unknown in prior which degradation mechanism an individual may follow, and some individuals could follow degradation pattern between multiple canonical models. Thus, mathematically, these canonical models span the modeling space for the individuals and provide a basis to characterize the individuals' variations on their own degradation mechanisms. This led us to create a membership vector $\boldsymbol{c}_i$ for individual i, where c_{ik} represents the degree to which the model of individual i resembles the canonical model k. With knowledge of the canonical models and the membership vector $\boldsymbol{c}_i$, the model of individual i can be derived.

The utilization of the canonical structure is just the first step to integrate the sparse data of multiple individuals. On top of CM, we further extend it to similarity-regularized CM (SCM) that can incorporate the similarity between individuals. Similarity between individuals have been long studied in the literature. In engineering applications for degradation modeling, the environmental factors and operational parameters of the individuals could be used to calculate their similarities. In healthcare applications, the demographic, social-economical, genetic and imaging information can be used for this purpose. Therefore, we believe that SCM could further enhance the estimation of the individual models with effective use of the canonical structure and the similarity information between individuals.

This paper is organized as follows. Section II reviews the related methods. Section III provides the details of the proposed collaborative learning framework. Section IV focuses on deriving an iterative updating algorithm that solves the constrained non-convex optimization problems. Section V presents comprehensive simulation studies and two case studies in Alzheimer's disease and degradation modeling of turbofan engines. Section VI provides a conclusion and discussion of future work.

II. Related Methods

As we have described in Section I, the proposed collaborative learning framework is different from the existing methods such as MEM models [1]–[3]. It is also different from the finite mixture regression models that have been developed in [15], [16]. The finite mixture regression models are also closely related to mixture regression model [17], latent class regression model [18], and clusterwise linear regression [19]. The finite mixture regression models assume that there are latent clusters, which is a different assumption from our proposed canonical structure. Our proposed collaborative learning approach is fundamentally different from these mixture regression models since our goal is to learn a regression model for each individual, rather than assigning individuals into clusters. Specifically, in those mixture regression models, a similar parameter as c_{ik} was also defined. However, in mixture regression models, it is assumed that the underlying model has each individual following one-and-only-one of the K canonical models, so the c_{ik} represents the probability that individual i follows the model k. However, here, a fundamental difference is that we assume that each individual i truly follows a linear combination of the K canonical models rather than one-and-only-one of the K canonical models. The model formulation is quite different, while the associated computational challenges are also different. The proposed method is also different from the reduced-rank regression models [20]–[22] that have been proposed for restricting the rank of the regression coefficient matrix of a regression model where multivariate outcomes are concerned simultaneously, and the dynamic weighted ensemble models [23], [24] which are developed to enhance the conventional framework of ensemble learning where a set of models are combined.

While many of the aforementioned methods could be shown as equivalent to the MEM model (but differ in the forms of the basis used in these regression models), the proposed collaborative learning method targets a different type of applications. The central assumption is that, in a heterogeneous population, although the regression model for each individual should differ from each other, the models can be characterized by a low-dimensional structure, e.g., considering the few types of degradation mechanisms of the phenotypes in a population. Individuals' models may be variants of these typical degradation mechanisms. Another unique merit of the proposed method is that it can address the data challenges such as sparse and irregular measurements, and explicitly utilize similarity information for modeling.

III. Collaborative Learning Framework

Since many degradation models take the form of regression, we use the degradation modeling as the context to illustrate the development of the collaborative learning framework. Particularly, we demonstrate the collaborative learning method using linear models in this study. Let the degradation model of individual i be $f_i(\boldsymbol{x})$ for $i = 1, \ldots, N$. We assume that $f_i(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}_i$, where $\boldsymbol{x}$ represents p predictors and $\boldsymbol{\beta}_i$ represents the corresponding regression parameters. Linear models have been found successful in characterizing a range of degradation models [25]–[28]. However, the proposed approach can incorporate nonlinear models by using nonlinear basis functions of predictors, e.g., polynomial basis functions are typical examples. Many nonlinear models such as Gaussian processes or kernel models can be represented as linear models using nonlinear basis functions or kernel tricks that map the original variables $\boldsymbol{x}(t)$ into the reproducing kernel Hilbert space defined by a certain kernel function [29], [30]. Here, we use the Gaussian process as an example for explanation.

Assuming the degradation model $f_i(\boldsymbol{x})$ takes the form as a Gaussian process, which can be defined as [31], [32]

$$f_i\left(\boldsymbol{x}\right) = \boldsymbol{x}\boldsymbol{\beta}_i + \sum_{j=1}^{n_i} \alpha_{ij} k\left(\boldsymbol{x}, \boldsymbol{x}_{ij}\right)$$

where $k(\boldsymbol{x}, \boldsymbol{x}_{ij})$ is the covariance function of the two vectors $\boldsymbol{x}$ and $\boldsymbol{x}_{ij}$, $\boldsymbol{\alpha}_i = [\alpha_{i1}, \ldots, \alpha_{in_i}]^T = (K_{\mathbf{y}_i,\mathbf{y}_i} + \sigma^2\mathbf{I})^{-1} (\boldsymbol{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)$, σ^2 is the variance parameter, $\mathbf{I}$ is the identical matrix, and $K_{\mathbf{y}_i,\mathbf{y}_i} = [k(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'})]_{n_i \times n_i}$ is called the covariance matrix of the observational data. The covariance function could take highly nonlinear forms such as the Gaussian covariance function or the polynomial covariance function. Therefore, a linear model provides a flexible framework for encompassing a wide range of models and can be easily extended to capture the nonlinear patterns.

Given the model formulation, the challenge is how to learn the models from data. The anticipated data for each individual, e.g., individual i, are longitudinal measurements at n_i time points. Denote these measurement as $\boldsymbol{y}_i = [y_{i1}, \ldots, y_{in_i}]^T \in \mathbb{R}^{n_i \times 1}$, and $\mathbf{X}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}]^T \in \mathbb{R}^{n_i \times p}$. The variability of the length of the longitudinal measurements for the individuals could be very large, resulting in a sparse and irregular data structure. In subsequent sections, we propose the CM and SCM to effectively learn $f_i(\boldsymbol{x})$, $i = 1, \ldots, N$.

A. Model Formulation

Let $g_k(\boldsymbol{x})$, $k = 1, \ldots, K$, be the degradation model of the kth canonical model such that $g_k(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{q}_k$, where $\boldsymbol{q}_k$ is the corresponding regression parameter vector. We assign a membership vector $\boldsymbol{c_i} = [c_{i1}, \ldots, c_{iK}]^T$ to each individual i, where c_{ik} represents the degree to which the individual i resembles the canonical model k, i.e., $f_i(\boldsymbol{x}) = \sum_k c_{ik} g_k(\boldsymbol{x})$. In linear models, i.e., $f_i(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}_i$, $g_k(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{q}_k$, this assumption could be further rewritten as $\boldsymbol{\beta}_i = \sum_k c_{ik}\boldsymbol{q}_k = \mathbf{Q}\boldsymbol{c}_i$ while $\mathbf{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K]$.

To learn the models, we propose the following formulation:

$$\min_{\boldsymbol{c}_i, \mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2$$
$$\text{subject to } c_{ik} \geq 0, \quad \sum_k c_{ik} = 1$$
$$\mathbf{X}_i\mathbf{Q} \geq \mathbf{0},\ \forall\, i = 1, \ldots, N, \text{ and } k = 1, \ldots, K. \tag{1}$$

Here, the objective function is the least square loss function to gauge the goodness-of-fit of the models. The last inequality $\mathbf{X}_i\mathbf{Q} \geq 0$ is imposed due to the constraint that the predicted health status, such as the cognitive status in AD, should stay nonnegative. The other two constraints, $c_{ik} \geq 0$ and $\sum_k c_{ik} = 1$, are imposed on $\boldsymbol{c}_i$ due to its definition as a membership vector. Then, by solving this optimization formulation, the K canonical models, encoded in $\mathbf{Q}$**,** and the membership vector for each individual, encoded in $\boldsymbol{c}_i$, can be estimated. Next, individuals' degradation models can be obtained by using $\boldsymbol{\beta}_i = \mathbf{Q}\boldsymbol{c}_i$.

The proposed collaborative learning framework is flexible and capable of fusing data and information from multiple sources. For instance, we could further extend the CM to incorporate the similarity information, denoted as w_{jl} for the similarity between individuals j and l. The similarity w_{jl} reflects how likely that the degradation models of the two individuals could be similar. Therefore, it is reasonable to assume that $\boldsymbol{c}_j$ and $\boldsymbol{c}_l$ are more similar with each other when w_{jl} is larger. w_{jl} can be obtained by various approaches that will be discussed in more details in Section IV.

To incorporate the similarity knowledge in the model formulation of CM, we add a regularization term, $\sum_{j,l} ||\boldsymbol{c}_j - \boldsymbol{c}_l||^2 w_{jl}$, into the objective function of (1), leading to the following SCM formulation:

$$\min_{\boldsymbol{c}_i, \mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2 + \frac{\lambda}{2}\sum_{j,l} ||\boldsymbol{c}_j - \boldsymbol{c}_l||^2 w_{jl}$$
$$\text{subject to } c_{ik} \geq 0, \quad \sum_k c_{ik} = 1, \quad \mathbf{X}_i\mathbf{Q} \geq \mathbf{0}$$
$$\forall\, i = 1, \ldots, N, \text{ and } k = 1, \ldots, K. \tag{2}$$

Here, λ is the parameter to control the effect of the regularization term on parameter estimation. A larger λ imposes more influence of the similarity regularization on the parameter estimation. The proposed formulation in (2) presents a challenging optimization task due to its non-convex nature and its constraint structure. We will present the details of the algorithm [12] in Section IV.

B. Relationship Between SCM and MEM

This section presents a connection between the proposed SCM with MEM. For linear models, MEM assumes that $\{\boldsymbol{\beta}_i, i = 1, 2, \ldots, N\}$ are independent identically distributed (i.i.d.), sampled from a multivariate normal distribution, i.e., $\boldsymbol{\beta}_i \sim N(0, \boldsymbol{G})$, where $\boldsymbol{G}$ denotes a covariance matrix. It can be shown that the objective function in (2) becomes equivalent to MEM under the specific conditions where $w_{jl} = 1$ for all j and l and $\boldsymbol{G} = \mathbf{Q}\mathbf{Q}^T$. Note that, here, $w_{jl} = 1$ for all j and l corresponds to the fact that MEM actually treats the individual models as identical samples from a distribution model.

Theorem 1: The objective function of the optimizing problem (2) is equivalent to the objective function of MEM when $\boldsymbol{W}$is a matrix with all the elements being one and $\boldsymbol{G} = \mathbf{Q}\mathbf{Q}^T$.

We include the detailed proof of the theorem in the Appendix. Theorem 1 provides a useful insight into the proposed collaborative learning approach's flexibility and unique capability of studying the heterogeneous models at a more fundamental and detailed level than MEM: (***a***) the proposed SCM provides more flexibility of incorporating information sources (w_{jl}) for capturing the similarity among individuals. However, the MEM is limited to $w_{jl} = 1$ for all j and l. These results suggest that the proposed approach should be more general than MEM; (***b***) the proposed model can characterize individual's heterogeneity by allocating different membership vectors, while the MEM treats individuals as identically distributed; (***c***) further, unlike the MEM that encapsulates the population heterogeneity into a variance-covariance matrix of random effects (e.g., encoded in $\mathbf{G}$), the proposed method can model the heterogeneity of the population by explicitly learning multiple canonical models in $\mathbf{Q}$.

Remark 1: While Theorem 1 suggests that the objective function of MEM can be considered as a special case of the objective function of SCM, it does not imply that MEM is a strictly special case of SCM. Because SCM employs the constraints in (2), it presents a more constrained version of MEM. In addition, the number of canonical models K plays an interesting role in defining an upper bound of the rank of the covariance matrix $\mathbf{Q}$. As such, SCM can be considered as a knowledge-driven MEM

```
The Learning Algorithm
Input: $\mathbf{X}_i$ and $\boldsymbol{y}_i$, $i = 1, \dots, N$; initial values $\mathbf{C}^{(0)}$ and $\mathbf{Q}^{(0)}$ ; $W$ ; $\lambda$ ; maximal iteration number, $MaxIter$
    For $r = 0,1,\dots,MaxIter$
        1. Convert $\boldsymbol{c}_i^r$ to $\tilde{\mathbf{C}}_i^r$using (4).
        2. Let $\mathbf{X}_i^* = \mathbf{X}_i\tilde{\mathbf{C}}_i^r, \mathbf{B}_i = diag(\mathbf{X}_i, \dots, \mathbf{X}_i)$. Calculate $\boldsymbol{q}^{r+1}$ by solving (5).
        3. Transform $\boldsymbol{q}^{r+1}$ to $\mathbf{Q}^{r+1}$ by partitioning the $Kp$ ×1 vector to the $p$×$K$ matrix.
        4. Calculate $\mathbf{C}^{r+1}$ by (8).
    End for
Output: $\{\mathbf{Q}^{(MaxIter+1)}, \mathbf{C}^{(MaxIter+1)}\}$
```

Fig. 1. Procedure of the proposed algorithm for solving SCM.

with an extra capability to incorporate the canonical structure of the random effects. On the other hand, SCM has more flexibility than MEM by further incorporating the similarity information between individuals as Theorem 1 indicates.

Remark 2: The proposed method is also different from the approaches [33]–[35] that have been proposed to jointly learn multiple regression models by treating these models as related. These models are commonly termed as transfer learning and multitask learning methods. Our proposed methods explicitly exploit the low-dimensional canonical structure and enable automatic determination of the relatedness of individual regression model to the canonical models. Finally, the proposed methods can reveal the low-dimensional canonical structure of the underlying problem (i.e., by using model selection methods to identify the number K) and produce the subgroup-level canonical models.

IV. Computational Implementation of the Proposed Methods

A. Derivation of the Computational Algorithm

In [12] we have developed the main body of this algorithm. Fig. 1 provides a summary of the overall algorithm. Our numerical studies in Section V suggest that the proposed algorithm is efficient and easy to converge. For completeness of the paper, we present main details of this algorithm. The optimization task, presented in (2), is to learn $\mathbf{Q}$ and $\mathbf{C}$. It is actually easier to adopt an iterative two-stage approach that solves for $\mathbf{Q}$ and $\mathbf{C}$ alternatively. To make this more clear, we transform (2) into (3)

$$\min_{\mathbf{C},\mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i\|^2 + \lambda \mathrm{Tr}\left(\mathbf{C}^T\mathbf{L}\mathbf{C}\right)$$
$$\text{subject to } c_{ik} \geq 0, \quad \sum_k c_{ik} = 1, \quad \mathbf{X}_i\mathbf{Q} \geq \mathbf{0}$$
$$\forall\, i = 1, \dots, N, \text{ and } k = 1, \dots, K. \tag{3}$$

Here $\frac{1}{2}\sum_{j,l}||\boldsymbol{c}_j - \boldsymbol{c}_l||^2 w_{jl} = \sum_{j=1}^N (\boldsymbol{c}_j)^T\boldsymbol{c}_j d_{jj} - \sum_{j,l=1}^N (\boldsymbol{c}_j)^T\boldsymbol{c}_l w_{jl} = \mathrm{Tr}(\mathbf{C}^T\mathbf{L}\mathbf{C})$, $d_{jj} = \sum_l w_{jl}$, $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix whose entries are $\{d_{jj}, j = 1, 2, \dots, N\}$ and $\mathbf{W}$ is the weight matrix for w_{jl}.

1) Estimation Step for $\mathbf{Q}$*:* Given $\boldsymbol{C}^r$, (3) is reduced to

$$\min_{\mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i\mathbf{Q}\boldsymbol{c}_i^r\|^2$$
$$\text{subject to } \mathbf{X}_i\mathbf{Q} \geq 0 \quad \forall\, i = 1, \dots, N.$$

Established methods could be used here to solve this problem. We define $\mathbf{X}_i^*$ as

$$\mathbf{X}_i^* = \mathbf{X}_i\tilde{\mathbf{C}}_i^r$$

where

$$\tilde{\mathbf{C}}_i^r = \begin{bmatrix} (\boldsymbol{c}_i^r)^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (\boldsymbol{c}_i^r)^T \end{bmatrix}_{(p\times Kp)} . \tag{4}$$

Then, the objective function of $\mathbf{Q}$ is

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

where $\boldsymbol{q}$ is a $Kp\ \times 1$ vector of the matrix $\mathbf{Q}$, i.e., by concatenating the columns of the matrix $\mathbf{Q}$. The constraint $\mathbf{X}_i\mathbf{Q} \geq \mathbf{0}$ can be written as $\mathbf{B}_i\boldsymbol{q} \geq \mathbf{0}$ where $\mathbf{B}_{i(Kp\times Kp)} = \mathrm{diag}(\mathbf{X}_i, \dots, \mathbf{X}_i)$. In this way, (3) can be rewritten as

$$\min_{\mathbf{Q}} \sum_i \|\boldsymbol{y}_i - \mathbf{X}_i^*\boldsymbol{q}\|^2$$
$$\text{subject to } \mathbf{B}_i\boldsymbol{q} \geq \mathbf{0} \quad \forall\, i = 1, \dots, N. \tag{5}$$

We solve this problem by a fast non-negative-constrained least square algorithm (FNNLS) [14], [36].

2) Estimation Step for $\mathbf{C}$*:* Given $\mathbf{Q}^r$, we can derive the Lagrangian of the formulation as

$$L = \sum_{i=1}^N (\boldsymbol{y}_i)^T\boldsymbol{y}_i - 2\sum_{i=1}^N (\boldsymbol{y}_i)^T\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i$$
$$+ \sum_{i=1}^N (\boldsymbol{c}_i)^T(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i + \lambda\mathrm{Tr}\left(\mathbf{C}^T\mathbf{L}\mathbf{C}\right)$$
$$+ \sum_{i=1}^N \mu_i\left[(\boldsymbol{c}_i)^T\mathbf{1} - 1\right].$$

Following the spirit in [13] and [14], we can derive an updating rule to solve for $\boldsymbol{c}_i$. It can be seen that

$$\frac{\partial L}{\partial \boldsymbol{c}_i} = -2(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\boldsymbol{y}_i + 2(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i$$
$$+\ 2(\lambda\mathbf{L}\mathbf{C})_i + \mu_i\mathbf{1}.$$

Since $\frac{\partial L}{\partial \boldsymbol{c}_i}$ should equal to zero, this establishes the first equation for solving for c_{ik}. On the other hand, we need to make sure that c_{ik} is nonnegative. We can use the complementarity condition to derive another equation for c_{ik}

$$
\begin{aligned}
&-\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\boldsymbol{y}_i\right]_k c_{ik}+\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i\right]_k c_{ik}\\
&+(\lambda\mathbf{L}\mathbf{C})_{ik}c_{ik}+\frac{1}{2}\mu_i c_{ik}=0. \qquad (6)
\end{aligned}
$$

Noting that $(\boldsymbol{c}_i)^T 1 = 1$ and $\mathbf{L} = \mathbf{D} - \mathbf{W}$, we have

$$
\begin{aligned}
\frac{1}{2}\mu_i &= \left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\boldsymbol{y}_i\right]^T\boldsymbol{c}_i+\lambda[(\mathbf{W}\mathbf{C})_i]^T\boldsymbol{c}_i\\
&\quad-\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i\right]^T\boldsymbol{c}_i-\lambda[(\mathbf{D}\mathbf{C})_i]^T\boldsymbol{c}_i. \qquad (7)
\end{aligned}
$$

Using (7), (6) can be written as

$$
\begin{aligned}
&\Bigg\{\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i\right]_k+\lambda(\mathbf{D}\mathbf{C})_{ik}\\
&+\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\boldsymbol{y}_i\right]^T\boldsymbol{c}_i\\
&+\lambda[(\mathbf{W}\mathbf{C})_i]^T\boldsymbol{c}_i\Bigg\}c_{ik}-\Bigg\{\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\boldsymbol{y}_i\right]_k\\
&+\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i\right]^T\boldsymbol{c}_i+\lambda(\mathbf{W}\mathbf{C})_{ik}\\
&+\lambda[(\mathbf{D}\mathbf{C})_i]^T\boldsymbol{c}_i\Bigg\}c_{ik}=0.
\end{aligned}
$$

Then we can derive the updating rule shown in (8) at the bottom of this page.

B. Convergence Properties of the Proposed Algorithm

Theorem 2 shows that by using the proposed algorithm, the objective function is non-increasing, converging to a stationary point.

Theorem 2: Stationary point convergence could be reached using the algorithm in Fig. 1.

Note that, stationary point is a necessary condition for optimality, but not a sufficient condition. Thus, Theorem 2 only implies that the algorithm will converge to a stationary point but not necessary the optimal point. However, empirically, we observe that it is usually the case that the algorithm will converge to the local optimal as well.

C. Empirical Issues of Implementing the Algorithm

Choosing the proper canonical models is critical for characterizing the degradation process. The form of canonical models can be determined based on expert opinion. For instance, the second-order polynomial model with respect to time is often used to describe the cognitive decline in Alzheimer's disease [47], [48] and the degradation of turbofan engine's health condition [25]. The optimal canonical model can also be found based on the empirical evidence from data. For example, the model selection techniques developed in the literatures [53], including adjusted R-square, AIC, BIC, likelihood ratio test, and cross validation, could be adopted to find the canonical models that give best performance.

In practices, sometimes $\mathbf{W}$ can be readily available through expert opinion. We could also quantify the similarity between individuals based on some covariates that reflect the characteristics of the individuals, e.g., such as the environmental factors and operational parameters in engineering applications for degradation modeling, or the demographic, social-economical, genetic and imaging information in many healthcare applications. For example, there have been many methods developed in the literature [37], [38] that can extract patient similarities from domain knowledge or medical records. Even when the covariates are not available, a heuristic approach could be used that treats the regression parameters of the individuals as the covariates. For instance, the MEM method can be used to learn the regression models of the individuals. The regression parameters estimated by the MEM represent the individual-to-individual variations. If the underlying low-dimensional canonical structure exists in the heterogeneous population, the SCM can further improve the estimation by extracting the similarity information from the regression parameters. To calculate the similarity, existing approaches, including the 0-1 weighting, heat kernel weighting, and dot-product weighting, could be used. Denote the covariates of individual j as $\boldsymbol{z}_j$. The heat kernel weighting defines the weights as $w_{ij} = \exp(-(\boldsymbol{z}_j - \boldsymbol{z}_l)^2/\sigma^2)$, where the scaling parameter σ^2 controls the similarity between individuals. The 0-1 weighting and dot-product weighting, on the other hand, do not have tuning parameter. For each individual, the 0-1 weighting finds its *k*-nearest neighborhoods and treats its neighborhoods equally similar, while the dot-product weighting measures the weights using cosine similarity between covariates, i.e., $w_{jl} = {\boldsymbol{z}_j}^T\boldsymbol{z}_l$. Therefore, the heat kernel weighting is more flexible to capture the similarity structure in real applications by allowing data-driven optimal tuning of similarities. We used the heat kernel weights in our numerical studies and it leads to satisfactory results on both synthetic datasets and real-world dataset.

How to obtain initial values of $\mathbf{C}^{(0)}$ and $\mathbf{Q}^{(0)}$ is also an important issue. We recommend using MEM [1] to initialize the parameters as we have gathered positive evidences from our simulation studies and real-world applications. Clustering algorithms such as k-means can be applied on the regression parameters that are learned by the MEM method. Then, the centroid vectors of the clusters that are learned by the k-means algorithm can be the initial values of $\mathbf{Q}^{(0)}$, and the similarity between the regression parameters of the individuals with the centroid vectors of the clusters can be used as the initial values of $\mathbf{C}^{(0)}$.

V. Numerical Studies

The proposed methods, CM [i.e., (1)] and SCM [i.e., (2)], will be compared with benchmark methods that include

$$
\boldsymbol{c}_{ik}^{m+1}=\boldsymbol{c}_{ik}^{m}\frac{\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\boldsymbol{y}_i+(\lambda\mathbf{W}\mathbf{C}^m)_i\right]_k+\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i^m\right]^T\boldsymbol{c}_i^m+\lambda[(\mathbf{D}\mathbf{C}^m)_i]^T\boldsymbol{c}_i^m}{\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\mathbf{X}_i\mathbf{Q}^r\boldsymbol{c}_i^m+(\lambda\mathbf{D}\mathbf{C}^m)_i\right]_k+\left[(\mathbf{Q}^r)^T(\mathbf{X}_i)^T\boldsymbol{y}_i\right]^T\boldsymbol{c}_i^m+\lambda[(\mathbf{W}\mathbf{C}^m)_i]^T\boldsymbol{c}_i^m} \qquad (8)
$$

TABLE I
COMPARISON OF CM, MEM, SCM, AND IGM ON SIMULATED DATASET WHEN $K = 3$

	Type 1 Model				Type 2 Model			
	IGM	CM	MEM	SCM	IGM	CM	MEM	SCM
Dense Sampling								
Running time (s)	**0.060**	28.260	75.000	95.580	**0.670**	203.740	248.050	260.120
MSE	8.143	3.933	3.795	**3.221**	58.050	**37.647**	43.337	40.705
nMSE	0.029	0.028	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								
Running time (s)	**0.450**	16.830	72.970	71.200	**0.590**	109.320	227.710	244.810
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**

COMPARISON OF CM, MEM, SCM, AND IGM ON SIMULATED DATASET WHEN $K = 5$

	Type 1 Model				Type 2 Model			
	IGM	CM	MEM	SCM	IGM	CM	MEM	SCM
Dense Sampling								
Running time (s)	**0.070**	34.210	76.290	76.750	**0.090**	285.090	364.560	381.900
MSE	11.753	6.589	7.758	**5.905**	52.606	40.004	47.166	**37.654**
nMSE	0.046	0.022	0.026	**0.018**	1.584	0.161	0.265	**0.154**
wR	0.978	0.991	0.988	**0.991**	0.580	0.921	0.768	**0.925**
rMSE1-step	39.828	30.620	35.149	**28.423**	56.393	15.408	21.029	**15.222**
rMSE3-step	56.066	38.667	41.008	**35.439**	50.239	16.521	18.817	**16.229**
rMSE5-step	74.589	50.243	52.345	**42.701**	58.343	16.865	21.341	**16.534**
Sparse Sampling								
Running time (s)	**0.240**	18.610	73.450	75.860	**0.510**	141.130	139.150	166.400
MSE	90.045	14.903	24.104	**10.784**	69.710	52.691	52.369	**52.088**
nMSE	9.473	0.358	0.114	**0.060**	1.703	0.569	0.543	**0.490**
wR	0.395	0.887	0.941	**0.971**	0.562	0.715	0.699	**0.752**
rMSE1-step	596.001	129.890	72.373	**54.688**	52.248	29.893	28.954	**28.551**
rMSE3-step	803.870	154.972	86.974	**64.305**	54.842	34.917	33.268	**31.080**
rMSE5-step	1050.278	191.821	108.631	**75.960**	68.883	31.403	32.560	**26.214**

the MEM and the model that estimates $\boldsymbol{\beta}_i$ independently (denote this model as IGM). Note the MEM estimates all the models together by assuming that $\boldsymbol{\beta}_i$ is sampled from a multivariate normal distribution. Performance of the models could be evaluated by the following criteria: parameter estimation (e.g., $\sum_{i,j} (\hat{\beta}_{ij} - \beta_{ij})^2 / (Np)$), the normalized mean square error (nMSE) on the testing set (e.g., $\mathrm{nMSE}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = [\sum_{t=1}^{T} \|\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t\|_2^2 / \sigma(\boldsymbol{y}_t)] / (\sum_{t=1}^{T} n_t)$), the weighted correlation coefficient (wR) (e.g., $\mathrm{wR}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = [\sum_{t=1}^{T} \mathrm{Corr}(\boldsymbol{y}_t, \widehat{\boldsymbol{y}_t}) n_t] / (\sum_{t=1}^{T} n_t)$), and the mean absolute error (MAE) (e.g., $\mathrm{MAE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{t=1}^{T} |\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t| / (\sum_{t=1}^{T} n_t)$). Here, $\widehat{\beta_{ij}}$ is estimated value of β_{ij}, $\hat{\boldsymbol{y}}_t$ is the predicted value of $\boldsymbol{y}_t$. We also evaluate the computational costs of these methods. Results could be found in Tables I–III.

TABLE II
PERFORMANCE OF THE MODELS (IGM, CM, MEM, AND SCM) ON COGNITIVE DECLINE PREDICTION.

	IGM	CM	MEM	SCM
Target: MMSE				
Running time (s)	**0.160**	6.220	8.050	92.090
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**

A. Simulation Studies

To enable a fair comparison, we consider multiple scenarios that consist of different combinations of several important dimensions: the number of canonical models, the sparsity of the data of each individual, and the types of degradation models. Particularly, given a degradation model and the randomly generated parameters for an individual, we randomly sample 25 time points from this model. To further mimic the sparsity of the data, we further randomly pick up M observations from the first 20 time points for model training. In the "dense sampling" scenario, $M \sim \mathrm{Unif}(15, 20)$; otherwise, $M \sim \mathrm{Unif}(4, 8)$ for "sparse sampling" scenario. The last five observations are always used for testing.

With given K and the type of degradation model, we can randomly generate $\mathbf{Q}$ and $\boldsymbol{c}_i$, and obtain $\boldsymbol{\beta}_i$ as $\boldsymbol{\beta}_i = \mathbf{Q}\boldsymbol{c}_i$. Specifically, to generate $\boldsymbol{c}_i$, we consider a mixture of distributions. Considering the case that there are three canonical models, we

TABLE III
PERFORMANCE OF THE MODELS (IGM, CM, MEM, AND SCM) ON ENGINE'S DEGRADATION PREDICTION.

	IGM	CM	MEM	SCM
Target: Health Index				
		No missing value		
Running time (s)	**0.359**	4.563	4.938	6.656
MAE	1.285	0.985	0.816	**0.741**
nMSE	7.757	3.075	2.362	**1.399**
wR	0.649	0.671	0.734	**0.741**
		20% missing value		
Running time (s)	**0.080**	3.906	4.210	5.938
MAE	1.340	1.151	0.915	**0.809**
nMSE	5.224	4.625	2.496	**1.592**
wR	0.580	0.689	0.718	**0.735**
		40% missing value		
Running time (s)	**0.110**	3.590	5.610	11.310
MAE	1.463	1.380	1.013	**0.902**
nMSE	7.465	4.736	3.909	**1.561**
wR	0.580	0.629	**0.695**	0.689
		80% missing value		
Running time (s)	**0.070**	0.906	3.320	3.234
MAE	2.558	1.465	1.246	**0.984**
nMSE	35.227	3.580	5.715	**1.040**
wR	0.442	0.491	**0.674**	0.658

could design three multivariate normal distributions as follows:

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

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

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

Each time, any of the three distribution could be randomly selected to generate a random sample. The resulting random sample is further normalized to obtain $\boldsymbol{c}_i$. Evidently, the larger the magnitude of v^2 in F_i, the more dominant the *i*th element in $\boldsymbol{c}_i$. Note that, here, v^2 controls the significance of the low-dimensional canonical structure, i.e., when v^2 is small, the difference between the canonical models becomes less significant. Thus, it is anticipated that when v^2 is large, SCM should outperform MEM; when v^2 is small, SCM and MEM should perform similarly. Our implementation results appear to be robust in a wide range of v^2 as long as the low-dimensional canonical structure is significant. In the sequel, we use $v^2 = 100$.

With $\boldsymbol{\beta}_i$, we can generate our training and testing samples, i.e., $y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta}_i + \varepsilon_{it}$. For example, for Type 1 model that has no predictor, $\boldsymbol{x}_{it} = [1, t, t^2]$; while for Type 2 model, $\boldsymbol{x}_{it} = [x_{i1t}, x_{i2t}, \ldots, x_{ipt}]$ assuming that predictors are available at each time epoch. Both types of models are popular in the literature of degradation modeling. We simulated $\boldsymbol{x}_{it}$ from the standard multivariate normal distribution.

After generating the data, we discover the optimal K, which is unknown in practice, by using the Akaike information criterion (AIC) [44]. To implement the SCM, here, the similarity between any two individuals is calculated based on the regression parameters estimated by MEM using the heat kernel function. Table I summarizes the results, which correspond to $K = 3$. Note that, in Table I, rMSE1-step means that the model is used to predict the degradation in the next time point; rMSE3-step means that the model is used to predict the degradation in the third time point; rMSE5-step means that the model is used to predict the degradation in the fifth time point. Our overall observations include:

1) When the low-dimensional canonical structure is significant (v^2 is large), the proposed CM and SCM is better than IGM, showing its efficacy in utilizing the low-dimensional structure.
2) Overall SCM outperforms other models, since SCM can effectively incorporate the similarity information between individuals to enhance model estimation.
3) The advantage of SCM is generally larger in the sparse sampling scenario, indicating that the incorporation of the structure of the heterogeneity of the population will be more preferred when there is a lack of observations.

B. Application to Alzheimer's Disease

This section demonstrates the performance of our proposed methods on a real-world dataset of Alzheimer's disease that was collected in [45]. A total of 478 subjects whose longitudinal measurements of Mini-Mental State Estimation (MMSE)—for measuring cognitive degradation—were collected at the baseline, 12th month, 24th month, 36th month, 48th month, and 60th month. These 478 subjects include 104, 261, and 113 individuals in the normal aging (NC), mild cognitive impairment (MCI), and Alzheimer's disease (AD) groups, respectively. Before we apply any model, it can be seen that the three groups show different degradation patterns: NC group is stable and slowly declines, AD group declines more rapidly, and MCI is in the between. The dataset is also sparse and irregular, i.e., 21, 156, and 244 individuals have only 3, 4, and 5 observations, respectively.

When we apply the CM and SCM models on this dataset, we didn't impose the knowledge that 3 groups exist in the dataset. Rather, as we will show later, it seems that our method can identify the optimal number of canonical models by data-driven approach. The last two measurements of the individuals are used as testing data, and the other measurements are used for training. Similarity information can be obtained by using the heat kernel weighting on baseline information of individuals such as ApoE genotypes, the baseline MMSE score, and the baseline regional brain volume measurements extracted from MRI via FreeSurfer [46]. To model the degradation of the MMSE score, the second-order polynomial model is used [47], [48].

To see if the proposed methods (CM and SCM) can automatically identify the low-dimensional canonical structure, AIC is used to select the best K in the CM and SCM models (together with other parameters such as λ and the scaling parameter in the heat kernel function). The results show that both methods can identify the low-dimensional canonical structure, e.g., the AIC results of SCM in Fig. 2 clearly show that the AIC value reaches minima when $K = 3$. Fig. 3 also shows that the algorithm converges quickly.

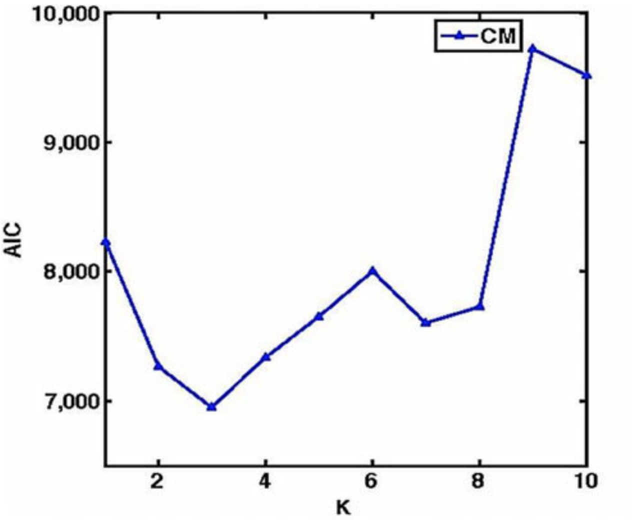

Fig. 2. AIC values versus $\boldsymbol{K}$ for CM.

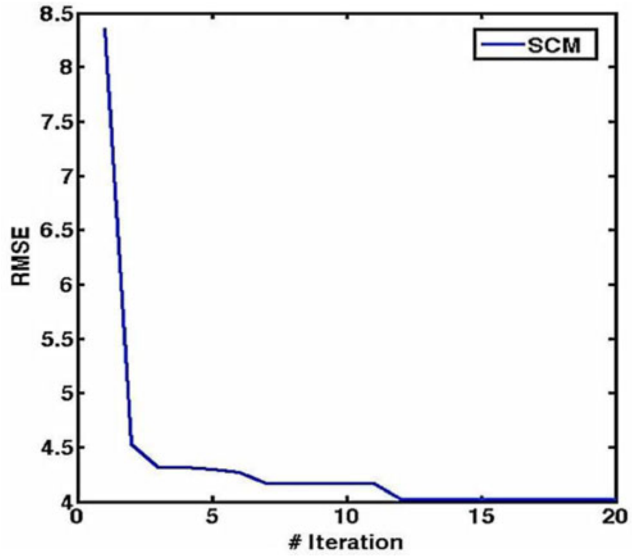

Fig. 3. Convergence performance of the computational algorithm for SCM.

We also investigate the canonical models discovered in Alzheimer's disease population, and show their cognitive degradation patterns in Fig. 4. It can be observed that three patterns of cognitive decline are discovered. These patterns represent the cognitive degradation trajectories of NC, MCI, and AD patients, respectively.

Table II summarizes the prediction results. Clearly, SCM outperforms other methods on all the performance metrics. We can also observe that CM is better than MEM and IGM. As a result, explicitly exploiting the heterogeneity of the population coupled with the low-dimensional canonical structure in CM and SCM appears to be better than only considering the variations among individuals in MEM. It is also clear that the performance deterioration of CM and SCM on predicting the 60th month (a long-term prediction) than the 48th month is much smaller than the other methods. Thus, long-term prediction/monitoring can be much improved by the CM and SCM methods.

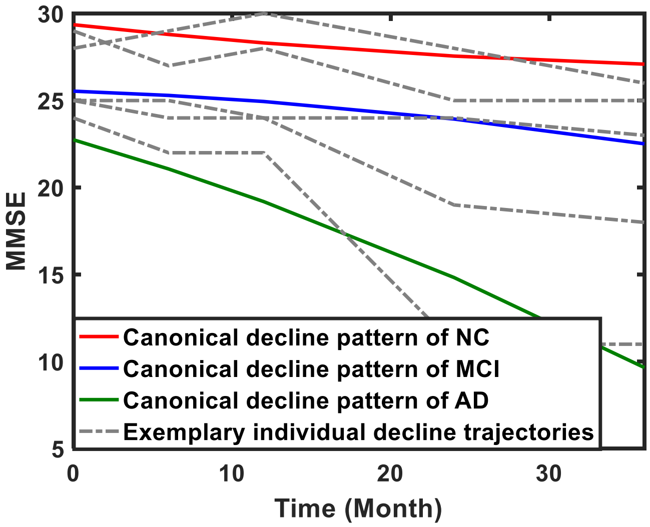

Fig. 4. Three canonical models discovered in Alzheimer's disease population using SCM.

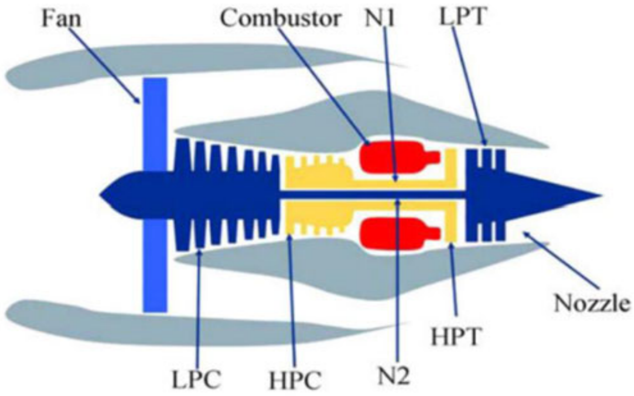

Fig. 5. Simplified engine diagram simulated in C-MAPSS [49].

C. Application to Degradation Modeling of Engines

We further investigate the performance of proposed methods on modeling the degradation processes of turbofan engines. We use a dataset that has been a benchmark dataset in the literature of degradation modeling, which was generated by a commercial modular aero-propulsion system simulation (C-MAPSS) testbed. The main purpose of this dataset is to mimic the degradation performance in large commercial turbofan engine (a schematic diagram is shown in Fig. 5). It has been known that there were two failure modes in our dataset, e.g., failure may occur at either the high-pressure compressor (HPC) or the fan of the engine. The dataset consists of measurements from 100 engines (units). Each engine was continuously measured from initial cycle of running until the end-of-life point (a cycle is defined as one flight). The analytic task is to convert the sensor information into degradation modeling. We adopt the data fusion model proposed in [25] as our degradation model. As suggested in [25], the degradation model of the health index of the engine is $f_i(t) = \theta_i^{(0)} + \theta_i^{(1)} t + \theta_i^{(2)} t^2 + \varepsilon_{i,t}$, where $f_i(t)$ represents the health index for unit i at cycle t and $\boldsymbol{\theta}_i = [\theta_i^{(0)}, \theta_i^{(1)}, \theta_i^{(2)}]$

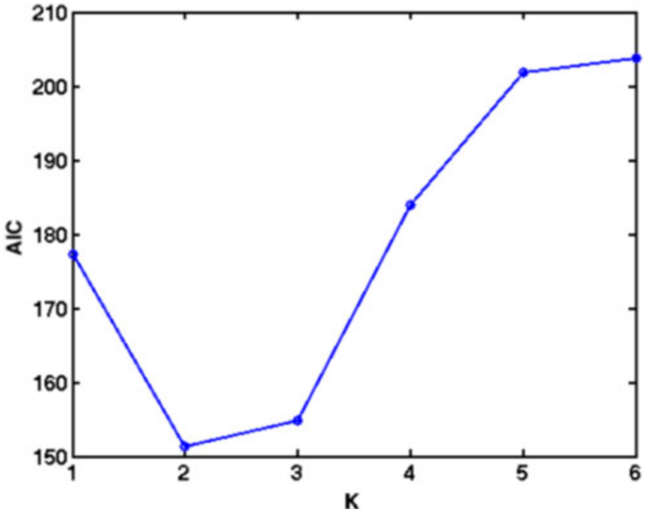

Fig. 6. AIC values versus $\boldsymbol{K}$ for CM under 80% missing value.

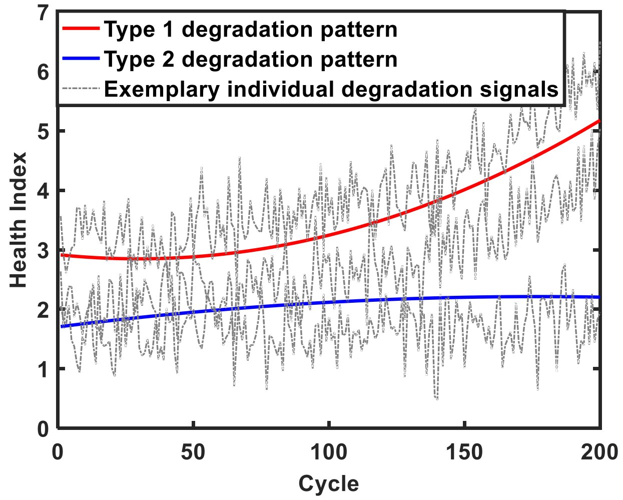

Fig. 7. Two canonical models discovered in turbofan engines' degradation using SCM.

represents the coefficients in the degradation model of unit i. For each unit, we use the first 150 health index observations as the training data (about 70% of the total data) to learn the degradation model and leave the remaining observations as testing data. Further, considering the commonly encountered measurement sparsity and irregularity in reality, we further trim the training data by randomly omitting some measurements, i.e., we consider different levels of missing value such as no missing value, 20%, 40%, and 80%. To obtain the similarity information of the units, we first fit a set of degradation models using MEM, and measure the similarities between units by applying the heat kernel weighting function on the regression coefficients of the units.

We first investigate if the proposed method can recover the underlying canonical structure, i.e., the number of failure modes which has been known to be two. As shown in Fig. 6, the AIC value obtained in the training data reaches minima when $K = 2$. Then we exploit the two types of degradation patterns characterized by canonical models in Fig. 7. As shown in Fig. 7, both failure modes are associated with increase of health index. One

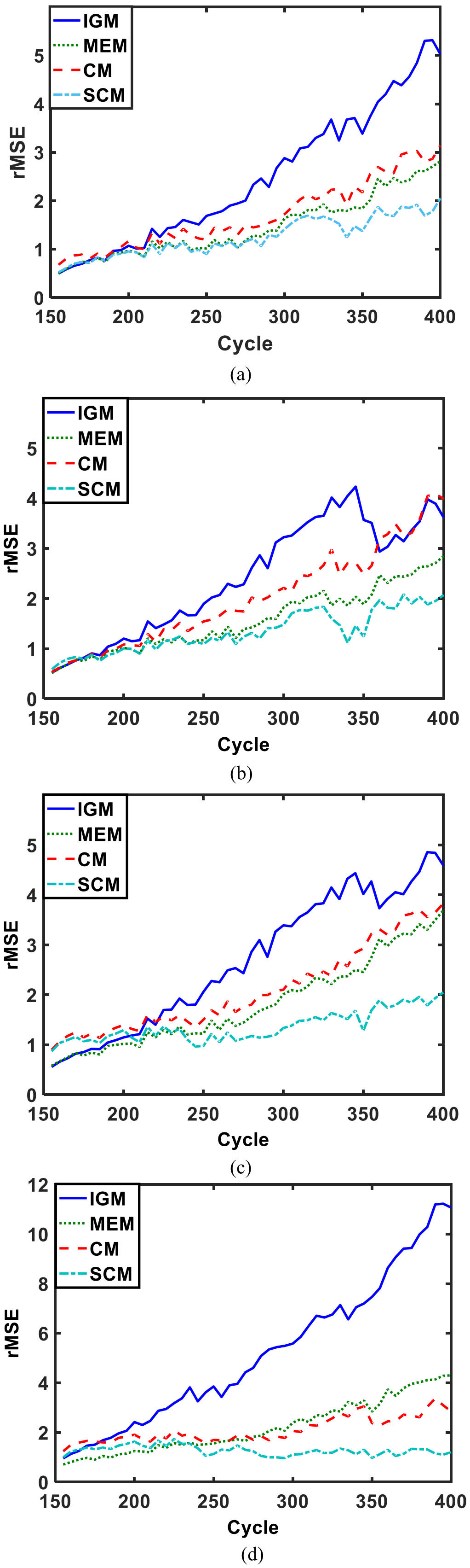

Fig. 8. Performance of the models (IGM, CM, MEM, and SCM) in terms of root of mean square error (rMSE) under: (a) no missing value, (b) 20% missing value, (c) 40% missing value, and (d) 80% missing value.

type of degradation pattern begins with high level of health index and grows dramatically whereas the other type of degradation pattern is characterized by relatively low health index with mild increase. The units following the fast degradation pattern (Type 1) tend to have shorter life cycles (distributed within 250 cycles). We further compare the prediction accuracy of the proposed methods, CM and SCM, with IGM and MEM on the testing data. Table III summarizes the overall prediction accuracy of these models, and the prediction accuracy under different amount of missing value is presented in Fig. 8. It can be observed that more observations in the training data will enhance the model learning process of all methods, leading to their improved prediction accuracy. The improvement of SCM model is more significant in the short-term prediction while predicting the long-term degradation with limit number of training cycles is still challenging. SCM outperforms the other models in both dense and sparse data, and the advantage is more obvious in the long-term prediction. While the IGM and MEM which assume the units are homogeneous or identically distributed focus on modeling the fast degradation process, the proposed methods which explicitly exploit the heterogeneity between units have significant advantage on characterizing the engines with longer life cycles leading to more obvious advantage in predicting the progression of health status after 250 cycles. In addition, by explicitly exploiting the similarity between units' degradation processes, CM and SCM are more powerful in predicting the progression of the health status of engines under sparse observations than IGM and MEM. This further shows the advantages of the proposed methods in monitoring the systems with limited sensing capacities or observations.

D. Computational Cost

The computational cost shown in Tables I to III indicates the following: 1) exploiting the correlation between individuals in the CM, MEM, and SCM models significantly improves the prediction accuracy and requires higher computational cost; 2) the CM is more computationally efficient than the MEM method and the SCM has a slightly higher computational cost than MEM; and 3) the computational cost of SCM is more sensitive to the number of individuals in the population compared to the CM model by explicitly exploiting the similarity between individuals. Therefore, the CM and SCM methods are more preferred under sparse and noisy data, and the CM model is more scalable to large population.

VI. Conclusion

In this paper, we propose a novel collaborative learning framework to estimate a heterogeneous population of regression models. It is motivated by the fact that existing models, such as the MEM, impose the homogeneity assumption that assumes the parameters of these prediction models are sampled from a distribution which is usually a multivariate normal distribution, so the mean vector can characterize the mean tendency of these parameters and the covariance matrix can characterize the dispersion of the models. While MEM is not preferred for applications in a heterogeneous population, where considerable heterogeneity of the models exists. To mitigate such heterogeneity, we propose the collaborative learning framework by exploiting the idea of "canonical models" and model regularization. First, to characterize the heterogeneity of the population, it uses a set of canonical models to represent the population characteristics. Then, the model of each individual resembles these canonical models to different degrees, in which the individual variety is characterized by membership vectors that can be learned from data. To enhance the estimation of the individual models, the similarity between the individuals can also be used by imposing a similarity-regularized term in the learning framework. Such a collaborative learning framework is applied in the context of degradation modeling, which leads to the development of CM and SCM. Both simulation studies and real-world applications show the superior performance of the proposed model. In addition, theoretical analysis is conducted to reveal the connection between the proposed methods with MEM.

In the future, we plan to apply the proposed collaborative learning framework to other applications and integrate the framework with other degradation models including nonlinear regression models or other dynamic degradation models (such as Markov models). We will also extend the proposed framework to incorporate more complicated canonical structures, e.g., a hierarchical canonical structure. Another research direction is to develop optimal sensing and monitoring strategies based on the collaborative learning framework.

Appendix

Proof of Theorem 1: In what follows, we will show the equivalence between the objective function of the MEM model with the objective function of our SCM model, given number of latent classes K and penalty parameter λ. First, we consider the objective function of MEM. The MEM uses fixed effects $\boldsymbol{b}_0^{\mathrm{MEM}}$ and random effects $\boldsymbol{b}_{ri}^{\mathrm{MEM}}$ to model the degradation of measures on each subject. It also assumes the random effects are correlated which can be formulated as

$$\begin{aligned} \boldsymbol{y}_i &= \mathbf{X}_i \boldsymbol{b}_0^{\mathrm{MEM}} + \mathbf{X}_i \boldsymbol{b}_{ri}^{\mathrm{MEM}} + \boldsymbol{\varepsilon} \\ \boldsymbol{b}_{ri}^{\mathrm{MEM}} &\sim N\left(\mathbf{0}, \mathbf{G}\right),\ \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \omega^2 \mathbf{I}\right). \end{aligned}$$

The conditional distribution of $\boldsymbol{y}_i$ is

$$\boldsymbol{y}_i | \boldsymbol{b}_{ri}^{\mathrm{MEM}} \sim N\left(\mathbf{X}_i \boldsymbol{b}_0^{\mathrm{MEM}} + \mathbf{X}_i \boldsymbol{b}_{ri}^{\mathrm{MEM}},\ \omega^2 \mathbf{I}\right).$$

Based on the conditional distribution of $\boldsymbol{y}_i$, we can derive the log-likelihood function as

$$\begin{aligned} ML_{\mathrm{MEM}} = & \sum_i \left\| \boldsymbol{y}_i - \mathbf{X}_i \boldsymbol{b}_0^{\mathrm{MEM}} - \mathbf{X}_i \boldsymbol{b}_{ri}^{\mathrm{MEM}} \right\|^2 \\ & + \omega^2 \sum_i \left(\boldsymbol{b}_{ri}^{\mathrm{MEM}}\right)^T \mathbf{G}^{-1} \boldsymbol{b}_{ri}^{\mathrm{MEM}}. \end{aligned}$$

Use $\boldsymbol{b}_i^{\mathrm{MEM}} = \boldsymbol{b}_0^{\mathrm{MEM}} + \boldsymbol{b}_{ri}^{\mathrm{MEM}}$ to replace $\boldsymbol{b}_{ri}^{\mathrm{MEM}}$ in ML_{MEM}, we have

$$\begin{aligned} ML_{\mathrm{MEM}} &= \sum_i \left\|\boldsymbol{y}_i - \mathbf{X}_i \boldsymbol{b}_i^{\mathrm{MEM}}\right\|^2 + \omega^2 \sum_i \left[\left(\boldsymbol{b}_i^{\mathrm{MEM}} \right.\right. \\ &\quad \left.\left. - \boldsymbol{b}_0^{\mathrm{MEM}} \right)^{\mathrm{T}} \mathbf{G}^{-1} \left(\boldsymbol{b}_i^{\mathrm{MEM}} - \boldsymbol{b}_0^{\mathrm{MEM}}\right) \right] \\ &= \sum_i \left\|\boldsymbol{y}_i - \mathbf{X}_i \boldsymbol{b}_i^{\mathrm{MEM}}\right\|^2 + \omega^2 \sum_i \left[\left(\boldsymbol{b}_i^{\mathrm{MEM}}\right)^{\mathrm{T}} \right. \\ &\quad \mathbf{G}^{-1} \boldsymbol{b}_i^{\mathrm{MEM}} - 2\left(\boldsymbol{b}_0^{\mathrm{MEM}}\right)^{\mathrm{T}} \mathbf{G}^{-1} \boldsymbol{b}_i^{\mathrm{MEM}} + \left(\boldsymbol{b}_0^{\mathrm{MEM}}\right)^{\mathrm{T}} \\ &\quad \left. \mathbf{G}^{-1} \boldsymbol{b}_0^{\mathrm{MEM}} \right]. \end{aligned}$$

Second, given K and λ, we consider the objective function of SCM. The objective function of SCM is

$$\begin{aligned} ML_{\mathrm{SCM}} = \sum_i ||\boldsymbol{y}_i - \mathbf{X}_i \mathbf{Q} \boldsymbol{c}_i||^2 + \lambda &\left[\sum_i (\boldsymbol{c}_i)^T \boldsymbol{c}_i \mathbf{D}_{ii} \right. \\ &\left. - \sum_{i,j} (\boldsymbol{c}_i)^T \boldsymbol{c}_j w_{ij} \right]. \end{aligned}$$

Using $\boldsymbol{b}_i^{\mathrm{SCM}} = \mathbf{Q}\boldsymbol{c}_i$ to replace $\boldsymbol{c}_i$ in the objective function ML_{SCM}, we have $\boldsymbol{c}_i = (\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T \boldsymbol{b}_i^{\mathrm{SCM}} = \mathbf{Q}^+ \boldsymbol{b}_i^{\mathrm{SCM}}$. $\mathbf{Q}^+$ is the pseudoinverse of $\mathbf{Q}$, which means $\mathbf{Q}\mathbf{Q}^+\mathbf{Q} = \mathbf{Q}$. Then ML_{SCM} can be rewritten as

$$ML_{\mathrm{SCM}} = \sum_i ||\boldsymbol{y}_i - \mathbf{X}_i \boldsymbol{b}_i^{\mathrm{SCM}}||^2 + \lambda$$

where $\boldsymbol{\Lambda}^{-1} = (\mathbf{Q}^+)^T \mathbf{Q}^+$, which leads to $\boldsymbol{\Lambda} = \mathbf{Q}\mathbf{Q}^T$.

When $\mathbf{W}$ is $\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$, e.g., $\mathbf{W}_{ij} = 1, \mathbf{D}_{ii} = N$, we have

$$\begin{aligned} ML_{\mathrm{SCM}} = \sum_i ||\boldsymbol{y}_i - \mathbf{X}_i \boldsymbol{b}_i^{\mathrm{SCM}}||^2 + \lambda N \sum_i &\left[\left(\boldsymbol{b}_i^{\mathrm{SCM}}\right)^{\mathrm{T}} \right. \\ \boldsymbol{\Lambda}^{-1} \boldsymbol{b}_i^{\mathrm{SCM}} - 2\left(\boldsymbol{b}_0^{\mathrm{SCM}}\right)^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \boldsymbol{b}_i^{\mathrm{SCM}} &\left. + \left(\boldsymbol{b}_0^{\mathrm{SCM}}\right)^{\mathrm{T}} \boldsymbol{\Lambda}^{-1} \boldsymbol{b}_0^{\mathrm{SCM}} \right] \end{aligned}$$

where $\boldsymbol{b}_0^{\mathrm{SCM}} = \frac{\sum_i \boldsymbol{b}_i^{\mathrm{SCM}}}{N}$.

Comparing the formulation of ML_{SCM} and ML_{MEM}, we observe that, if $\lambda = \frac{\omega}{N}$, the objective function of MEM is equivalent to the objective function of SCM, with a constraint that $\mathrm{rank}(\mathbf{G}) = \mathrm{rank}(\boldsymbol{\Lambda}) \leq \mathrm{rank}(\mathbf{Q}) = K$.

Proof of Theorem 2: Our proof follows similar ideas used by [13] and [51]. It's obvious that the objective function in (3) is bounded from below by zero, so the Lagrangian L is also bounded from below. The solution will converge if the Lagrangian L is monotonically nonincreasing. To prove Theorem 2, first, we need to show that the Lagrangian is nonincreasing in each step of the iterative algorithm, and then, prove the iterative updates converge to a stationary point. Since in each iteration, the estimated $\mathbf{Q}$ minimizes the objective function**,** we only need to prove that the Lagrangian will be nonincreasing under the updating rule of $\mathbf{C}$ in (8).

Since the updating rule (8) is essentially element wise, it is sufficient to show that each $L_{ik}(c) = L(c, \mathbf{C}^m_{\backslash i \backslash k})$ is monotonically nonincreasing under the update step of (8). Note that $L_{ik}(c)$ only depends on c_{ik} when other values in $\mathbf{C}$ are given. To prove this, we need to use the concept of auxiliary function, which is similar to that used in the expectation-maximization algorithm [52].

A function $G(c', c)$ is an auxiliary function of $L_{ik}(c)$ if

$$G\left(c', c\right) \geq \; L_{ik}\left(c'\right); \; G\left(c, c\right) = L_{ik}\left(c\right). \tag{9}$$

By constructing $G(c', c)$, we define

$$c^{m+1} = \arg\min_c G\left(c, c^m\right). \tag{10}$$

Thus, we have $L_{ik}(c^m) = G(c^m, c^m) \geq G(c^{m+1}, c^m) \geq L_{ik}(c^{m+1})$. This leads to the monotonicity of L_{ik} under the iterative updating rule of (10).

To construct an auxiliary function, we write $L_{ik}(c)$ using the Taylor expression

$$\begin{aligned} L_{ik}\left(c\right) = L_{ik}\left(c_{ik}^m\right) &+ L'_{ik}\left(c_{ik}^m\right)\left(c - c_{ik}^m\right) \\ &+ \frac{1}{2} L''_{ik}(c_{ik}^m)(c - c_{ik}^m)^2 \end{aligned}$$

where

$$\begin{aligned} L'_{ik} &= \frac{\partial L}{\partial c_{ik}} = \left[-2\mathbf{Q}^T (\mathbf{X}_i)^T \boldsymbol{y}_i + 2\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q} \boldsymbol{c}_i^m \right]_k \\ &\quad + 2\lambda(\mathbf{L}\mathbf{C}^m)_{ki} + \mu_i \\ L''_{ik} &= \frac{\partial L'}{\partial c_{ik}} = 2\left[\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q} \right]_{kk} + 2\lambda \mathbf{D}_{kk} \\ L_{ik}^{(t)} &= 0, \qquad \text{for } t > 2. \end{aligned}$$

Then we have

$$\begin{aligned} \frac{\left[\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q} \boldsymbol{c}_i^m \right]_k + (\lambda \mathbf{D}\mathbf{C}^m)_{ki}}{c_{ik}^m} &\geq \left[\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q} \right]_{kk} \\ + \lambda \mathbf{D}_{kk} = \frac{1}{2} L''_{ik}\left(c_{ik}^m\right) \end{aligned}$$

because

$$\begin{aligned} \left[\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q} \boldsymbol{c}_i^m \right]_k &= \sum_j \left[\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q} \right]_{kj} c_{ij}^m \\ &\geq \left[\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q} \right]_{kk} c_{ik}^m \end{aligned}$$

and

$$(\lambda \mathbf{D}\mathbf{C}^m)_{ki} = \sum_j \lambda \mathbf{D}_{kj} c_{ij}^m \geq \lambda \mathbf{D}_{kk} c_{ik}^m$$

given that $\mathbf{X}_i\mathbf{Q} \geq 0$, $\boldsymbol{c}_i \geq 0$.

This leads to the first unnumbered equation at the bottom of the next page.

Therefore, we can show the function shown in the second unnumbered equation at the bottom of the next page is an auxiliary function of $L_{ik}(c)$.

This is because $G(\cdot,\cdot)$ satisfies the conditions in (9): the last term in $G(c, c_{ik}^m) \geq L_{ik}(c)$; and the equality holds when $c = c_{ik}^m$.

The minimum for (10) can be obtained by setting the gradient to zero, i.e. see the third unnumbered equation at the bottom of this page.

This leads to the solution shown in the unnumbered equation at the top of the next page.

By substituting the Lagrange multiplier μ_i, as shown in (7) in the aforementioned equation, we recover the updating rule (8).

Next, we prove the iterative updates converge to a stationary point that satisfies the Karush Kuhn Tucker conditions.

Lemma 1: Starting from an arbitrary feasible nonzero point $\mathbf{C}^0$and $\mathbf{Q}^0$, the iterative procedure based on updating rule in (8) converge to a point that satisfies the KKT conditions for the optimization problem

$$\min_{\boldsymbol{c}_i,\, i=1,\ldots,\, k} \sum_i ||\boldsymbol{y}_i - \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i||_F^2 + \lambda \mathrm{Tr}\left(\mathbf{C}^T \mathbf{L} \mathbf{C}\right)$$

$$\text{subject to: } \boldsymbol{c}_i \geq 0,\ \boldsymbol{c}_i{}^T 1 = 1 \quad \forall\, i = 1,\, \ldots, N. \tag{11}$$

Proof of Lemma 1: We first write the KKT conditions for the optimization problem in (11)

1) Stationarity:

$$-\,2(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i + 2(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i + \, 2(\lambda \mathbf{L}\mathbf{C})_i + \boldsymbol{\varphi}_i + \mu_i \mathbf{1} = 0,\ \forall\, i = 1,\, \ldots, N.$$

2) Primal feasibility:

$$\boldsymbol{c}_i \geq \mathbf{0} \quad \forall\, i = 1,\, \ldots, N$$

$$\boldsymbol{c}_i^T \mathbf{1} = 1 \quad \forall\, i = 1,\, \ldots, N.$$

3) Dual feasibility:

$$\boldsymbol{\varphi}_i \geq \mathbf{0} \quad \forall\, i = 1,\, \ldots, N.$$

4) Complementary slackness:

$$\varphi_{ik} c_{ik} = 0 \quad \forall\, i = 1, \ldots,\, N,\ k = 1, \ldots,\, K.$$

It is straightforward to observe that, starting with nonnegative nonzero $\mathbf{C}^0$, the updating rule in (8) always keeps $\boldsymbol{c}_i^m$ nonnegative nonzero since the nonnegative assumption of the cognitive measures, $\boldsymbol{y}_i$ and $\mathbf{X}_i \mathbf{Q}^*$.

Assuming $\boldsymbol{c}_i^m$ converge to $\boldsymbol{c}_i^*$, we have (12) shown at the top of the next page.

Equation (12) implies (13) shown at the top of the next page.

Therefore, we obtain the second unnumbered equation shown at the top of the next page.
and

$$\left\{\left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i^*\right]^{\mathbf{T}} \boldsymbol{c}_i^* + \lambda[(\mathbf{D}\mathbf{C}^*)_i]^T \boldsymbol{c}_i^* + \left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i\right]^T \boldsymbol{c}_i^* + \lambda[(\mathbf{W}\mathbf{C}^*)_i]^T \boldsymbol{c}_i^*\right\}\left(\sum_k c_{ik}^* - 1\right) = 0.$$

This leads to $\sum_k c_{ik}^* - 1 = 0$, because $\boldsymbol{c}_i^*$ is nonnegative nonzero under the updating rule. Consequently, the first term in the equation is nonzero. The updates $\boldsymbol{c}_i^m$ converge to a feasible solution.

By (7), it is known that

$$\frac{1}{2}\mu_i = \left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i\right]^T \boldsymbol{c}_i^* + \lambda[(\mathbf{W}\mathbf{C}^*)_i]^T \boldsymbol{c}_i^* - \left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i^*\right]^{\mathbf{T}} \boldsymbol{c}_i^* - \lambda[(\mathbf{D}\mathbf{C}^*)_i]^T \boldsymbol{c}_i^*.$$

$$\frac{\left[\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q} \boldsymbol{c}_i^m\right]_k + (\lambda \mathbf{D}\mathbf{C}^m)_{ki} + \left[\mathbf{Q}^T (\mathbf{X}_i)^T \boldsymbol{y}_i\right]^T \boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^m)_i]^T \boldsymbol{c}_i^m}{c_{ik}^m} \geq \left[\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}\right]_{kk} + \lambda \mathbf{D}_{kk}$$

$$G\left(c, c_{ik}^m\right) = L_{ik}\left(c_{ik}^m\right) + L'_{ik}\left(c_{ik}^m\right)\left(c - c_{ik}^m\right) + \frac{\left[\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q} \boldsymbol{c}_i^m\right]_k + (\lambda \mathbf{D}\mathbf{C}^m)_{ki} + \left[\mathbf{Q}^T (\mathbf{X}_i)^T \boldsymbol{y}_i\right]^T \boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^m)_i]^T \boldsymbol{c}_i^m}{c_{ik}^m}\left(c - c_{ik}^m\right)^2$$

$$\begin{aligned}\frac{\partial G\left(c, c_{ik}^m\right)}{\partial c} &= L'_{ik}\left(c_{ik}^m\right) + 2\frac{\left[\mathbf{Q}^{\mathrm{T}} (\mathbf{X}_i)^{\mathrm{T}} \mathbf{X}_i \mathbf{Q} \boldsymbol{c}_i^m\right]_k + (\lambda \mathbf{D}\mathbf{C}^m)_{ki} + \left[\mathbf{Q}^{\mathrm{T}} (\mathbf{X}_i)^{\mathrm{T}} \boldsymbol{y}_i\right]^{\mathrm{T}} \boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^{\mathrm{m}})_i]^{\mathrm{T}} \boldsymbol{c}_i^m}{c_{ik}^m}\left(c - c_{ik}^m\right) \\ &= 2\left[\left[-\mathbf{Q}^{\mathrm{T}} (\mathbf{X}_i)^{\mathrm{T}} \boldsymbol{y}_i\right]_k - \lambda(\mathbf{W}\mathbf{C}^m)_{ki} + \frac{1}{2}\mu_i - \left[\mathbf{Q}^{\mathrm{T}} (\mathbf{X}_i)^{\mathrm{T}} \boldsymbol{y}_i\right]^{\mathrm{T}} \boldsymbol{c}_i^m - \lambda[(\mathbf{W}\mathbf{C}^{\mathrm{m}})_i]^{\mathrm{T}} \boldsymbol{c}_i^m \right. \\ &\quad \left. + \frac{\left[\mathbf{Q}^{\mathrm{T}} (\mathbf{X}_i)^{\mathrm{T}} \mathbf{X}_i \mathbf{Q} \boldsymbol{c}_i^m\right]_k + (\lambda \mathbf{D}\mathbf{C}^{\mathrm{m}})_{ki} + \left[\mathbf{Q}^{\mathrm{T}} (\mathbf{X}_i)^{\mathrm{T}} \boldsymbol{y}_i\right]^{\mathrm{T}} \boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^{\mathrm{m}})_i]^{\mathrm{T}} \boldsymbol{c}_i^m}{c_{ik}^m} c\right] = 0\end{aligned}$$

$$c = c_{ik}^{m} \frac{\left[\mathbf{Q}^T (\mathbf{X}_i)^T \boldsymbol{y}_i\right]_k + \lambda(\mathbf{W}\mathbf{C}^m)_{ki} + \left[\mathbf{Q}^T (\mathbf{X}_i)^T \boldsymbol{y}_i\right]^T \boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^m)_i]^T \boldsymbol{c}_i^m - \frac{1}{2}\mu_i}{\left[\mathbf{Q}^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q} \boldsymbol{c}_i^m\right]_k + (\lambda \mathbf{D}\mathbf{C}^m)_{ki} + \left[\mathbf{Q}^T (\mathbf{X}_i)^T \boldsymbol{y}_i\right]^T \boldsymbol{c}_i^m + \lambda[(\mathbf{W}\mathbf{C}^m)_i]^T \boldsymbol{c}_i^m}$$

$$c_{ik}^{*} = c_{ik}^{*} \frac{\left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i + (\lambda \mathbf{W}\mathbf{C}^*)_i\right]_k + \left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i^*\right]^{\mathbf{T}} \boldsymbol{c}_i^* + \lambda[(\mathbf{D}\mathbf{C}^*)_i]^T \boldsymbol{c}_i^*}{\left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i^* + (\lambda \mathbf{D}\mathbf{C}^*)_i\right]_k + \left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i\right]^T \boldsymbol{c}_i^* + \lambda[(\mathbf{W}\mathbf{C}^*)_i]^T \boldsymbol{c}_i^*} \tag{12}$$

$$c_{ik}^{*} \left\{ \frac{\left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i + (\lambda \mathbf{W}\mathbf{C}^*)_i\right]_k + \left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i^*\right]^{\mathbf{T}} \boldsymbol{c}_i^* + \lambda[(\mathbf{D}\mathbf{C}^*)_i]^T \boldsymbol{c}_i^*}{\left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i^* + (\lambda \mathbf{D}\mathbf{C}^*)_i\right]_k + \left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i\right]^T \boldsymbol{c}_i^* + \lambda[(\mathbf{W}\mathbf{C}^*)_i]^T \boldsymbol{c}_i^*} - 1 \right\} = 0 \tag{13}$$

$$\sum_k \left\{ c_{ik}^{*} \left\{ \frac{\left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i + (\lambda \mathbf{W}\mathbf{C}^*)_i\right]_k + \left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i^*\right]^{\mathbf{T}} \boldsymbol{c}_i^* + \lambda[(\mathbf{D}\mathbf{C}^*)_i]^T \boldsymbol{c}_i^*}{\left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i^* + (\lambda \mathbf{D}\mathbf{C}^*)_i\right]_k + \left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i\right]^T \boldsymbol{c}_i^* + \lambda[(\mathbf{W}\mathbf{C}^*)_i]^T \boldsymbol{c}_i^*} - 1 \right\} \right\} = 0$$

With (13), we have

$$c_{ik}^{*} \left[\left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i + (\lambda \mathbf{W}\mathbf{C}^*)_i\right]_k - \left[(\mathbf{Q}^*)^T \ (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i^* \right.\right.$$

$$\left.\left. + \ (\lambda \mathbf{D}\mathbf{C}^*)_i \right]_k - \frac{1}{2}\mu_i \right] = 0$$

with

$$\varphi_{ik}^{*} = 2\left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \boldsymbol{y}_i + (\lambda \mathbf{W}\mathbf{C}^*)_i\right]_k$$

$$- \ 2\left[(\mathbf{Q}^*)^T (\mathbf{X}_i)^T \mathbf{X}_i \mathbf{Q}^* \boldsymbol{c}_i^* + (\lambda \mathbf{D}\mathbf{C}^*)_i\right]_k - \mu_i.$$

Showing that both the complementary slackness and stationarity conditions are satisfied. Because c_{ik}^* is nonnegative nonzero, φ_{ik}^* is necessary to be zero. Thus, all the KKT conditions are satisfied in $\boldsymbol{c}_i^*$ based on our updating rules.

References

[1] A. Galecki and T. Burzykowski, "Linear mixed-effects models using R," *Springer Texts in Statistics*. New York, NY, USA: Springer, 2013.

[2] A. S. Bryk and S. Raudenbush, "Application of hierarchical linear models to accessing change," *Psychol. Bull.*, vol. 101, pp. 147–158, 1987.

[3] H. Goldstein, *Multilevel Models in Education and Social Research*. New York, NY, USA: Oxford Univ. Press, 1987.

[4] C. Villegas, G. A. Paula, and V. Leiva, "Birnbaum–Saunders mixed models for censored reliability data analysis," *IEEE Trans. Rel.*, vol. 60, no. 4, pp. 748–758, Dec. 2011.

[5] J. Son *et al.*, "Evaluation and comparison of mixed effects model based prognosis for hard failure," *IEEE Trans. Rel.*, vol. 62, no. 2, pp. 379–394, Jun. 2013.

[6] P. Qiu, C. Zou, and Z. Wang, "Nonparametric profile monitoring by mixed effects modeling," *Technometrics*, vol. 52, pp. 265–277, 2012.

[7] Z. Ye *et al.*, "Degradation data analysis using Wiener processes with measurement errors," *IEEE Trans. Rel.*, vol. 62, no. 4, pp. 772–780, Dec. 2013.

[8] C. Hertzog, R. A. Dixon, D. F. Hultsch, and S. W. Macdonald, "Latent change models of adult cognition," *Psychol. Aging*, vol. 18, pp. 755–769, 2003.

[9] C. R. J. Jack, D. S. Knopman, W. J. Jagust *et al.*, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," *Lancet Neurol.*, vol. 9, pp. 119–128, 2010.

[10] D. R. Royall, R. Palmer, and L. K. Chiodo, "Normal rates of cognitive change in successful aging," *J. Neuropsychol. Soc.*, vol. 11, pp. 899–909, 2005.

[11] R. C. Petersen, G. E. Smith, S. C. Waring *et al.*, "Mild cognitive impairment: Clinical characterization and outcome," *Arch. Neurol.*, vol. 56, pp. 303–308, 1999.

[12] Y. Lin, K. Liu, E. Byon, X. Qian, and S. Huang, "Domain knowledge driven cognitive degradation modeling for Alzheimer's disease," in *Proc. Soc. Ind. Appl. Math. Publ. SIAM Int. Conf. Data Mining*, 2015, pp. 721–729.

[13] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[14] R. Bro and S. D. Jong, "A fast non-negative-constrained least squares algorithm," *J. Chemometr.*, vol. 11, pp. 393–401, 1997.

[15] G. McLachlan and D. Peel, *Finite Mixture Models*. New York, NY, USA: Wiley, 2004.

[16] F. Leisch, "FlexMix: A general framework for finite mixture models and latent class regression in R," *J. Statist. Softw.*, vol. 11, no. 8, pp. 1–18, 2004.

[17] T. Hoshikawa, "Mixture regression for observational data, with application to functional regression models," 2013. [Online]. Available: http://arxiv.org/abs/1307.0170

[18] J. K. Vermunt and J. Magidson, "Latent class cluster analysis," *Applied Latent Class Analysis*, Jacques A. Hagenaars and Allan L. McCutcheon, eds. Cambridge, U.K.: Cambridge Univ. Press, ch. 3, pp. 89–106, 2002.

[19] W. S. DeSarbo and W. L. Cron, "A maximum likelihood methodology for clusterwise linear regression," *J. Classif.*, vol. 5, no. 2, pp. 249–282, 1999.

[20] T. W. Anderson, "Estimating linear restrictions on regression coefficients for multivariate normal distributions," *Ann. Math. Statist.*, vol. 22, pp. 327–351, 1951.

[21] A. J. Izenman, "Reduced-rank regression for the multivariate linear model," *J. Multivar. Anal.*, vol. 5, no. 2, pp. 248–264, 1975.

[22] G. C. Reinsel and R. P. Velu, *Multivariate Reduced Rank Regression: Theory and Applications*. New York, NY, USA*:* Springer, 1998.

[23] Z. Q. Shen and F. S. Kong, "Dynamically weighted ensemble neural networks for regression problems," in *Proc. IEEE Int. Conf. Mach. Learn. Cybern.*, 2004*,* vol. 6, pp. 3492–3496.

[24] J. Liu, V. Vitelli, E. Zio, and R. Seraoui, "A novel dynamic-weighted probabilistic support vector regression-based ensemble for prognostics of time series data," *IEEE Trans. Rel.*, vol. 64, no. 4, pp. 1203–1213, Dec. 2015.

[25] K. Liu and S. Huang, "Integration of data fusion methodology and degradation modeling process to improve prognostics," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 1, pp. 344–354, Jan. 2016.

[26] K. Liu, N. Gebraeel, and J. Shi, "A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis," *IEEE Trans. Autom. Sci. Eng.*, vol. 10, no. 3, pp. 652–664, Jul. 2013.

[27] M. J. Zuo, R. Jiang, and R. C. M. Yam, "Approaches for reliability modeling of continuous-state devices," *IEEE Trans. Rel.*, vol. 48, no. 1, pp. 9–18, Jul. 2013.

[28] N. Gorjian, L. Ma, M. Mittinty *et al.*, "A review on degradation models in reliability analysis," in *Engineering Asset Lifecycle Management*. London, U.K.: Springer, 2010, pp. 369–384.

[29] A. J. Smola, A. Gretton, L. Song, and B. Scholkopf, "A hilbert space embedding for distributions," *Lecture Notes Comput. Sci.*, vol. 4757, pp. 13–31, 2007.

[30] B. Scholkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA, USA: MIT Press, 2002.

[31] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[32] K. A. Doksum and S. L. T. Normand, "Gaussian models for degradation processes—Part I: Methods for the analysis of biomarker data," *Lifetime Data Anal.*, vol. 1, no. 2, pp. 131–144, 1995.

[33] J. Zhou, J. Chen, and J. Ye, "MALSAR: Multi-task learning via structural regularization," Arizona State Univ., 2012. [Online]. Available: http://www.public.asu.edu/∼jye02/Software/MALSAR

[34] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via multi-task learning," *NeuroImage*, vol. 78, pp. 233–248, 2013.

[35] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[36] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems.* Englewood Cliffs, NJ, USA: Prentice-Hall, 1947.

[37] F. Wang, J. Sun, J. Hu, and S. Ebadollahi, "Imet: Interactive metric learning in healthcare applications," in *Proc. SIAM Data Mining Conf.*, pp. 944–55, 2011.

[38] J. Sun, D. Sow, J. Hu, and S. Ebradollahi, "Localized supervised metric learning on temporal physiological data," in *Proc. 20th Int. Conf. Pattern Recog. (ICPR)*, pp. 4149–4152, 2010.

[39] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Adv. Neural Inf. Process. Syst.*, vol. 14, pp. 585–591, 2001.

[40] S. Duchesne, A. Caroli, C. Geroldi *et al.*, "Relating one-year cognitive change in mild cognitive impairment to baseline MRI features," *NeuroImage*, vol. 47, pp. 1363–1370, 2009.

[41] D. Head, R. L. Buckner, J. S. Shimony *et al.*, "Differential vulnerability of anterior white matter in nondemented aging with minimal acceleration in dementia," *Cerebral Cortex*, vol. 14, pp. 410–423, 2004.

[42] G. Bartzokis, D. Sultzer, P. H. Lu *et al.*, "Heterogeneous age-related breakdown of white matter structural integrity," *Neurobiol. Aging*, vol. 25, pp. 843–851, 2004.

[43] B. M. Jedynak *et al.*, "A computational neurodegenerative disease progression score," *NeuroImage*, vol. 63, pp. 1478–1486, 2012.

[44] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[45] S. G. Mueller, M. W. Weiner, L. J. Thal *et al.*, "The Alzheimer's disease neuroimaging initiative," *Neuroimaging Clin. North Amer.*, vol. 15, pp. 869–877, 2005.

[46] C. J. Jack, M. Bernstein, N. Fox *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Recog. Imaging*, vol. 27, pp. 685–691, 2008.

[47] 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," *Psychol. Methods*, vol. 9, pp. 30–52, 2004.

[48] M. J. Sliwinski *et al.*, "Modeling memory decline in older adults: The importance of preclinical dementia," *Psychol. Aging*, vol. 18, pp. 658–671, 2003.

[49] A. Saxena, K. Goebel, D. Simon, and N. Eklund, "Damage propagation modeling for aircraft engine run-to-failure simulation," in *Proc. 1st Int. Conf. Prognostics Health Manage.*, Denver, CO, USA, Oct. 2008, pp. 1–9.

[50] S. Sarkar, X. Jin, and A. Ray, "Data-driven fault detection in aircraft engines with noisy sensor measurements," *ASME J. Eng. Gas Turbines Power*, vol. 133, p. 081602, 2011.

[51] C. Ding, Y. Zhang, T. Li, and S. R. Holbrook, "Biclustering protein complex interactions with a biclique finding algorithm, In: Data Mining," in *Proc. 6th Int. Conf. Data Mining*, 2006, pp. 178–187.

[52] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[53] J. B. Kadane and N. A. Lazar, "Methods and criteria for model selection," *J. Amer. Statist. Assoc.*, vol. 99, no. 465, pp. 279–290, 2004.

Ying Lin received the B.S. degree in statistics from the University of Science and Technology of China, Hefei, China, in 2012, the M.S. degree in industrial and management system engineering from the University of South Florida, Tampa, FL, USA, in 2014, and the Ph.D. degree in industrial and systems engineering from the University of Washington, Seattle, WA, USA, in 2017.

She is currently an Assistant Professor in the Department of Industrial Engineering, University of Houston, Houston, TX, USA. Her research interests include process modeling and sensing strategy design for large-scale population.

Prof. Lin is a member of INFORMS, IISE, and SMDM.

Kaibo Liu (M'14) received the B.S. degree in industrial engineering and engineering management from the Hong Kong University of Science and Technology, Hong Kong, China, in 2009, and the M.S. degree in statistics and the Ph.D. degree in industrial engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2011 and 2013, respectively.

He is currently an Assistant Professor in the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA. His current research interests include data fusion for process modeling, monitoring, diagnosis, and prognostics.

Prof. Liu is a member of INFORMS, IISE, and ASQ.

Eunshin Byon received the B.S. and M.S. degrees in industrial and systems engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejon, South Korea, and the Ph.D. degree in industrial and systems engineering from Texas A&M University, College Station, TX, USA.

She is currently an Assistant Professor in the Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, USA. Her current research interests include optimizing operations and management of power systems, data analytics, quality and reliability engineering, and simulations.

Prof. Byon is a member of the IISE and INFORMS.

Xiaoning Qian received the Ph.D. degree in electrical engineering from Yale University, New Haven, CT, USA, in 2005.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. His current research interests include computational network biology, genomic signal processing, and biomedical image analysis.

Shan Liu received the B.S. degree in electrical engineering from the University of Texas at Austin, Austin, TX, USA, in 2006, the M.S. degree in technology and policy from Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008, and the Ph.D. degree in management science and engineering from Stanford University, Stanford, CA, USA, in 2013.

She is currently an Assistant Professor in the Department of Industrial and Systems Engineering, University of Washington, Seattle, WA, USA. Her current research interests include applied operations research methods in medical decision making and health policy modeling.

Prof. Liu is a member of INFORMS, IISE, and SMDM.

Shuai Huang (M'12) received the B.S. degree in statistics from the University of Science and Technology of China, Hefei, China, in 2007, and the Ph.D. degree in industrial engineering from the Arizona State University, Tempe, AZ, USA, in 2012.

He is currently an Assistant Professor in the Department of Industrial and Systems Engineering, University of Washington, Seattle, WA, USA. His current research interests include statistical learning and data mining with applications in healthcare and manufacturing.

Prof. Huang is a member of INFORMS, IISE, and ASQ.