MULTIVARIATE FINITE MIXTURE LATENT TRAJECTORY MODELS WITH APPLICATION TO DEMENTIA STUDIES

Dongbing Lai

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biostatistics,
Indiana University

August 2015

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Sujuan Gao, Ph.D., Co-chair

Huiping Xu, Ph.D., Co-chair

Doctoral Committee

Tatiana Foroud, Ph.D.

July 2, 2015

Barry Katz, Ph.D.

Daniel Koller, Ph.D.

DEDICATION

To My family

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to Dr. Sujuan Gao and Dr. Huiping Xu, co-chairs of my thesis committee, for their immense knowledge, enormous amount of precious time and wonderful guidance. This work cannot be done without their brilliant ideas, patience and enthusiasm. They are tremendous mentors for me.

I would like to express my special appreciation to my committee members, Dr. Tatiana Foroud, Dr. Barry Katz and Dr. Daniel Koller for their expertise and time. Their motivations, insightful suggestions, and encouragements are key elements of this research.

I am grateful to the Department of the Biostatistics and the Department of Mathematics for this wonderful PhD program. Dedicated faculty and always ready-to-help staff provide a great academic environment and make studying a joyful experience. I also thank my fellow graduate students for their various supports.

Special thanks to Department of Medical and Molecular Genetics. My working experience of analyzing real data is a great help to my PhD study. I also take this opportunity to express my gratitude to my supervisors, Drs. Tatiana Foroud and Daniel Koller again for being supportive during my studies.

Dongbing Lai

MULTIVARIATE FINITE MIXTURE LATENT TRAJECTORY MODELS WITH APPLICATION TO DEMENTIA STUDIES

Dementia studies often collect multiple longitudinal neuropsychological measures in order to examine patients' decline across a number of cognitive domains. Dementia patients have shown considerable heterogeneities in individual trajectories of cognitive decline, with some patients showing rapid decline following diagnoses while others exhibiting slower decline or remain stable for several years. In the first part of this dissertation, a multivariate finite mixture latent trajectory model was proposed to identify longitudinal patterns of cognitive decline in multiple cognitive domains with multiple tests within each domain. The expectation-maximization (EM) algorithm was implemented for parameter estimation and posterior probabilities were estimated based on the model to predict latent class membership. Simulation studies demonstrated satisfactory performance of the proposed approach. In the second part, a simulation study was performed to compare the performance of information-based criteria on the selection of the number of latent classes. Commonly used model selection criteria including the Akaike information criterion (AIC), Bayesian information criterion (BIC), as well as consistent AIC (CAIC), sample adjusted BIC (SABIC) and the integrated classification likelihood criteria (ICLBIC) were included in the comparison. SABIC performed uniformly better in all simulation scenarios and hence was the preferred criterion for our proposed model. In the third part of the dissertation, the multivariate finite mixture latent trajectory model was extended to situations where the true latent class membership was known for a subset of patients. The proposed models were used to analyze data from the

Uniform Data Set (UDS) collected from Alzheimer's Disease Centers across the country to identify various cognitive decline patterns among patients with dementia.

Sujuan Gao, Ph.D., Co-chair

Huiping Xu, Ph.D., Co-chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Dementia is common in the elderly population and is characterized by the progressive decline of cognitive function leading to impairment in the ability to perform daily activities and eventually loss of independence. The leading cause of dementia is Alzheimer's disease (AD), followed by other disorders such as vascular dementia (VD), frontotemporal dementia (FTD), and Lewy body dementia (LBD). Many patients also have coexisting pathologies with two or more subtypes of dementia. Dementia patients show substantial heterogeneity in their individual trajectories of cognitive decline, with some patients showing rapid decline while others exhibiting slower decline or remaining stable [1]. In addition, the trajectory of cognitive decline also varies across cognitive domains, for example, patients with AD typically have more prominent memory deficits with additional deficits in language [2-4], whereas patients with FTD show greater impairment in language and less impairment in memory [5-9].

Most research on the identification of distinct longitudinal trajectories on the patterns of cognitive decline had focused on a single neuropsychological test using group-based trajectory models (GBTM) [10-12] or growth mixture models (GMM) [13-15]. Both types of models assume that subjects belong to one of several unobserved subpopulations/groups/latent classes (in the following chapters, these terms are used interchangeably) with each group characterized by a unique longitudinal trajectory. GBTM and GMM models have been widely used in many research areas such as sociology, psychology, and criminology [11, 14]. However, these models are not well suited for multivariate longitudinal data often encountered in dementia studies where multiple neuropsychological tests are typically performed in order to characterize patients'

level of cognition. To address this, Proust-Lima and colleagues developed a model that treats the multiple tests as measures of a single latent quantity, characterizes the latent process exhibiting distinct longitudinal patterns across subpopulations [16, 17]. This new model has the flexibility of handling multivariate cognitive outcomes simultaneously. However, it only allows one single latent quantity, which is a rather strong assumption because neuropsychological tests are from more than one cognitive domain.

In the first part of this dissertation, we extended the model proposed by Proust-Lima et al by allowing multiple latent quantities, one for each cognitive domain and measured by multiple neuropsychological tests from that domain. These latent quantities jointly identify subpopulations of patients who exhibit distinct longitudinal patterns. This model is aimed at identifying subpopulations that may share the same disease etiology and leading to better treatment outcomes.

One challenging issue in the area of finite mixture models is the determination of the number of subpopulations. There are many studies on this topic but no well-established approach thus far [11, 18-22]. The commonly used likelihood ratio test (LRT) cannot be used to compare models with different number of subpopulations due to the fact that the null hypothesis involves zero mixing proportions, hence violating the regularity condition [19]. Other likelihood-based approaches such as the Lo, Mendell and Rubin (LMR) test [23] and the bootstrap likelihood ratio test (BLRT) [19] have limited application due to the high computational burden, especially for complicated models. Most studies used information criteria (IC) based approaches such as Akaike's Information Criterion (AIC) [24] and Bayesian Information Criterion (BIC) [25]. However, several studies have shown that AIC tends to overestimate the number of

groups, especially when sample size is large [19, 20]. BIC has also been known to suffer from the overestimation of the number of subpopulations [26]. In addition, it has been observed in several studies that BIC may decrease monotonically as more groups were added [11]. In addition to AIC and BIC, there are other IC-based fit indices include consistent AIC (CAIC) [27]; sample adjusted BIC (SABIC) [28-31], integrated classification likelihood criterion (ICLBIC) [32]. They were proposed to augment the performance of AIC and BIC, and in several simulation studies, they showed promising results [18, 21, 22]. However, their performance was evaluated in different contexts such as latent class modeling and growth mixture modeling. In the second part of this dissertation, a simulation study was performed to evaluate the performance of aforementioned IC-based indices for the multivariate finite mixture latent trajectory model proposed in the first part of this dissertation under different conditions with varying number of subjects, number of observations per subject, and level of separation among latent classes.

In some dementia studies, although it is difficult to determine the exact dementia subtype for all subjects, there may be a subsample of patients who underwent autopsy and have known dementia subtypes. Such data are often called partially labelled data in latent class analysis literature [19, 33-36]. Incorporating the true dementia subtype for these patients could potentially improve the accuracy of inferring patients' unknown dementia subtype. Studies showed that even 10% of labelled data can improve the accuracy of classification [37, 38]. In addition to improved classification accuracy, the existence of labelled data can make model estimation more efficient with faster convergence [19]. In the third part of this dissertation, the multivariate finite mixture

latent trajectory model was further extended to the situation where partially labelled data are available. Simulation studies were performed to investigate how labelled data can improve the classification accuracy and estimation efficiency under several considerations. Then the same data set analyzed in the first part was re-analyzed and results were compared to see how the additional information can help identifying dementia subtypes.

CHAPTER 2. A MULTIVARIATE FINITE MIXTURE LATENT TRAJECTORY MODEL WITH APPLICATION TO DEMENTIA STUDIES

2.1 Summary

Dementia patients exhibit considerable heterogeneity in individual trajectories of cognitive decline, with some patients showing rapid decline following diagnoses while others exhibiting slower decline or remaining stable for several years. Dementia studies often collect longitudinal measures of multiple neuropsychological tests aimed to measure patients' decline across a number of cognitive domains. We propose a multivariate finite mixture latent trajectory model to identify distinct longitudinal patterns of cognitive decline simultaneously in multiple cognitive domains, each of which is measured by multiple neuropsychological tests. EM algorithm is used for parameter estimation and posterior probabilities are used to predict latent class membership. We present results of a simulation study demonstrating adequate performance of our proposed approach and apply our model to the Uniform Data Set (UDS) from the National Alzheimer's Coordinating Center (NACC) to identify cognitive decline patterns among dementia patients.

2.2 Introduction

Dementia is common in the elderly population with Alzheimer's disease (AD) the leading cause [39] and is characterized by the progressive decline of cognitive function leading to impairment in the ability to perform daily activities and consequently, loss of independence. There is considerable heterogeneity in the individual trajectories of cognitive decline among dementia patients, with some patients showing rapid decline

while others exhibiting slower decline or remaining stable [1]. The heterogeneity of the cognitive decline also varies across cognitive domains. Prior research has shown that patients with AD had more prominent memory deficits with additional deficits in language [2-4], whereas patients with FTD had greater impairment in language and less impairment in memory [5-9].

Research on the identification of distinct longitudinal trajectories of cognitive decline has focused on univariate cognitive outcomes, measured by a single neuropsychological test using group-based trajectory models (GBTM) [10-12] and growth mixture models (GMM) [13-15]. GBTM, also known as latent class growth analysis (LCGA) [14], was proposed by Nagin and colleagues [10-12] while GMM was developed by Muthén et al [13-15]. Both models assume that subjects belong to one of several subpopulations/groups/latent classes, each characterized by a unique longitudinal trajectory. A key difference between GBTM and GMM is that GBTM assumes conditional independence, i.e. longitudinal measures across time within a subject are independent, whereas the GMM allows correlations among longitudinal outcomes within a subject with the introduction of subject-specific random effects [14]. Therefore, GBTM can be considered as a special case of GMM [14].

Despite the successful application of the GBTM and GMM models in many research studies [11, 14], these models are not well suited for multivariate longitudinal data due to the restriction that each latent process can only be constructed from one test. When evaluating cognitive decline among dementia patients, data are collected across several cognitive domains with multiple neuropsychological tests in each domain. Tests within each domain are measures of same underlying latent construct from different

prospective. Proust-Lima and colleagues extended the GBTM and GMM models to multivariate longitudinal data by treating the multiple tests as measures of a single latent quantity with a latent process that exhibits distinct longitudinal patterns across subpopulations [16, 17]. Although this approach has the flexibility to handle multivariate cognitive outcomes, it only allows exploration of longitudinal patterns of a single latent quantity, a limitation that undermines the capability of the model. In dementia studies where many neuropsychological tests are used to measure different aspects of cognitive function including memory, language, and executive function, it will be more realistic to assume multiple latent quantities and identify longitudinal patterns associated with different cognitive domains.

In this chapter, we extend the model proposed by Proust-Lima et al by allowing more than one latent quantity, each of which can be measured by multiple tests, and identifying subpopulations of patients who exhibit distinct longitudinal patterns in these latent quantities. Our work was directly motivated by studies of cognitive decline among dementia patients. Our proposed approach is aimed at identifying longitudinal patterns of cognitive decline defined in multiple cognitive domains. The identified subpopulations share the similar cognitive decline patterns therefore may share the same disease etiology; therefore, it can help us to find better patient care and treatment. And these phenotypically homogeneous subgroups can be used to improve the ability of searching disease causing genes in genetic studies.

The remainder of the chapter is organized as follows. In Section 2.3, we introduce the multivariate finite mixture latent trajectory model. We discuss parameter estimation using the EM algorithm and standard error computation in Section 2.4. We present results

from simulation studies in Section 2.5. Section 2.6 includes results from the application of our model to the Uniform Data Set (UDS) from the National Alzheimer's Coordinating Center (NACC) [40]. We present a discussion and conclude the chapter in section 2.7.

2.3 The Multivariate Finite Mixture Latent Trajectory Model

Assume that the population consists of G subpopulations represented by G latent classes. For individual i, $i = 1, \dots, N$, we define a G-dimentional vector $\boldsymbol{\omega_i}$ denoting the latent class membership, with $\omega_{ig} = 1$ if individual i belongs to class g and 0 otherwise. Suppose there are K neuropsychological tests with continuous outcomes representing cognitive function in D cognitive domains. Let $\boldsymbol{y_i} = \left(\boldsymbol{y_{i1}^T}, \dots, \boldsymbol{y_{ik}^T}, \dots, \boldsymbol{y_{iK}^T}\right)^{\boldsymbol{T}}$ be the vector of all measurements for individual i, where $\boldsymbol{y_{ik}}$ is a vector of length n_{ik}, which denotes the number of longitudinal measurements for individual i and test k $(k = 1, \dots, K)$, hence the length of $\boldsymbol{y_i}$ is $\sum_{k=1}^{K} n_{ik}$. Let $\boldsymbol{X_{1i}(t)}$ and $\boldsymbol{Z_i(t)}$ be the matrices of covariates collected for individual i. $\boldsymbol{Z_i(t)}$ can have partial or all columns of $\boldsymbol{X_{1i}(t)}$ but contains at least one time variable. Then a measurement model if individual i is in latent class g is:

$$\boldsymbol{y_{i|\omega_{ig}=1}} = \boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)} + \boldsymbol{V_i c_i} + \boldsymbol{\varepsilon_i}, \tag{2.1}$$

Where the latent trajectory is defined as:

$$\boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)} = \boldsymbol{X_{1i}(t)\beta_g} + \boldsymbol{Z_i(t)b_{ig}}, \tag{2.2}$$

The length of latent process $\boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)}$ is also $\sum_{k=1}^{K} n_{ik}$. Note that for the tests that are in the same domain, they share the same latent process by having the same values in $\boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)}$. $\boldsymbol{\beta_g}$ is the vector of class-specific fixed effects from all cognitive domains in

latent class g. Its length is $P \times D$, where P is the number of covariates. $\boldsymbol{b_{ig}}$ is the class specific random effects for all domains in latent class g. Similar to $\boldsymbol{\beta_g}$, $\boldsymbol{b_{ig}}$ has length $q \times D$, where q is the number of random effects. We assume that $\boldsymbol{b_{ig}}$ has a multivariate normal distribution $N(\mathbf{0}, W_g^2 \boldsymbol{B})$ with $W_1^2 = 1$ and $\boldsymbol{B}$ is the covariance matrix of first latent class, similarly defined as in Proust et al [16]. $\boldsymbol{c_i}$ in (2.1) is the K-vector of test-specific random intercept. It introduces correlation among scores of the same test from the same individual. Here we assume $\boldsymbol{c_i}$ is distributed as $N(\mathbf{0}, \boldsymbol{\Sigma_c})$, where $\boldsymbol{\Sigma_c}$ is a diagonal matrix with σ_{ck}^2 in its diagonal. Design matrix $\boldsymbol{V_i}$ in (2.1) is a $\sum_{k=1}^{K} n_{ik} \times K$ block matrix with the following structure:

$$\boldsymbol{V}_i = \begin{bmatrix} \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}$$

where $\mathbf{1}$ is a column vector of 1s. In k^{th} column, the column vector of 1s has length n_{ik}. $\boldsymbol{\varepsilon_i}$ in (2.1) is an vector of random error with distribution $N(\mathbf{0}, \boldsymbol{\Sigma_\varepsilon})$, where $\boldsymbol{\Sigma_\varepsilon}$ is a block matrix with $\sigma_{\varepsilon k}^2 \boldsymbol{I}_{n_{ik}}$ at diagonal and all other entries are 0s.

Accordingly, covariate matrix $\boldsymbol{X_{1i}(t)}$ has the following structure:

$$\boldsymbol{X_{1i}(t)} = \begin{bmatrix} \boldsymbol{X_{1i1}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{X_{1iD}} \end{bmatrix}$$

Each $\boldsymbol{X_{1id}}$ has all covariates for all tests in domain $d, d = 1, \ldots, D$ with dimension $n_{ik} \times P$. Similarly the design matrix $\boldsymbol{Z_i(t)}$ has the following structure:

$$\boldsymbol{Z_i(t)} = \begin{bmatrix} \boldsymbol{Z_{i1}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{Z_{iD}} \end{bmatrix}$$

where $\boldsymbol{Z_{id}}$ is a matrix of time polynomial of degree $q - 1$ with dimension $n_{ik} \times q$. For example, if $n_{ik} = 3$, for a quadratic model, each Z_{id} has structure as following:

$$\boldsymbol{Z_{id}} = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 \\ 1 & t_{i2} & t_{i2}^2 \\ 1 & t_{i3} & t_{i3}^2 \end{bmatrix}$$

We assume that $\boldsymbol{b_{ig}}$, $\boldsymbol{c_i}$ and $\boldsymbol{\varepsilon_i}$ are mutually independent.

For individual $i,\ i = 1, \dots, N$, the probability that this individual belongs to a latent class $g, g = 1, \dots, G$, is π_{ig}, with $\sum_{g=1}^{G} \pi_{ig} = 1$. This can be modeled through a multinomial logistic regression as:

$$\pi_{ig} = P(\omega_{ig} = 1 | \boldsymbol{X_{2i}^T}) = \frac{\exp(\boldsymbol{X_{2i}^T \gamma_g})}{1+\sum_{h=1}^{G-1} \exp(\boldsymbol{X_{2i}^T \gamma_h})}, \tag{2.3}$$

where $\boldsymbol{\gamma_g}$ is the vector of the class-specific regression coefficients. For identifiability purpose, $\boldsymbol{\gamma_G}$ are set to 0s. Covariates $\boldsymbol{X_{2i}^T}$ used here can be the same or different from $\boldsymbol{X_{1i}(t)}$ in equation (2.2).

2.4 Parameter Estimation

Since the latent class memberships are unobserved and there are also multiple random effects, the expectation-maximization (EM) algorithm can be used for obtaining parameter estimates [19, 41-43]. Let

$$\boldsymbol{\Psi} = (\boldsymbol{\beta_1}, \dots, \boldsymbol{\beta_g}, \dots, \boldsymbol{\beta_G}, W_2^2, \dots, W_g^2, \dots, \boldsymbol{W_G^2}, \boldsymbol{B}, \boldsymbol{\Sigma_c}, \boldsymbol{\Sigma_\varepsilon}, \boldsymbol{\gamma_2}, \dots, \boldsymbol{\gamma_g}, \dots, \boldsymbol{\gamma_G})$$

be the parameters to be estimated, $f_{ig}(\boldsymbol{y_i})$ be the density function of $\boldsymbol{y_i}$ in latent class g, then the observed-data likelihood is:

$$L(\boldsymbol{\Psi}) = \prod_{i=1}^{N} \sum_{g=1}^{G} \pi_{ig} f_{ig}(\boldsymbol{y_i}) \tag{2.4}$$

$f_{ig}(\boldsymbol{y_i})$ has distribution $N(\boldsymbol{X_{1i}(t)\beta_g}, \boldsymbol{\Sigma_{ig}})$, where

$$\boldsymbol{\Sigma_{ig}} = \boldsymbol{Z_i(t)} W_g^2 \boldsymbol{B Z_i(t)^T} + \boldsymbol{V_i \Sigma_c V_i^T} + \boldsymbol{\Sigma_\varepsilon}.$$

Augmenting the observed data $\boldsymbol{y_i}$ with unobserved variables $\left(\boldsymbol{\omega_i}, \boldsymbol{b_{i1}} \dots, \boldsymbol{b_{ig}}, \dots, \boldsymbol{b_{iG}}, \boldsymbol{c_i}\right)$, the complete-data likelihood function is:

$$L^c(\boldsymbol{\Psi}) = \prod_{i=1}^{N} \prod_{g=1}^{G} \{\pi_{ig} f(\boldsymbol{y_i}|\boldsymbol{b_{ig}}, \boldsymbol{c_i}) f(\boldsymbol{b_{ig}}) f(\boldsymbol{c_i})\}^{\omega_{ig}}$$

The log-likelihood for the complete data is

$$\log\left(L^c(\boldsymbol{\Psi})\right) = \sum_{i=1}^{N}\sum_{g=1}^{G} \omega_{ig}\{\log(\pi_{ig}) + \log\left(f(\boldsymbol{y_i}|\boldsymbol{b_{ig}}, \boldsymbol{c_i})\right) + \log\left(f(\boldsymbol{b_{ig}})\right) + \log(f(\boldsymbol{c_i}))\}$$

$$= \sum_{i=1}^{N}\sum_{g=1}^{G} \omega_{ig}\log(\pi_{ig}) - \frac{\sum_{k=1}^{K} n_{ik} + l + K}{2} \sum_{i=1}^{N}\sum_{g=1}^{G} \omega_{ig} \log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G} \omega_{ig}\log|\boldsymbol{\Sigma_\varepsilon}|$$

$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G} \omega_{ig}\left(\boldsymbol{y_i} - \boldsymbol{X_{1i}}(\boldsymbol{t})\boldsymbol{\beta_g} - \boldsymbol{Z_i}(\boldsymbol{t})\boldsymbol{b_{ig}} - \boldsymbol{V_i}\boldsymbol{c_i}\right)^T \boldsymbol{\Sigma_\varepsilon^{-1}}\left(\boldsymbol{y_i} - \boldsymbol{X_{1i}}(\boldsymbol{t})\boldsymbol{\beta_g} - \boldsymbol{Z_i}(\boldsymbol{t})\boldsymbol{b_{ig}} - \boldsymbol{V_i}\boldsymbol{c_i}\right)$$

$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G} \omega_{ig}\log|\boldsymbol{B}| - \frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G} l * \omega_{ig} \log(W_g^2) - \frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G} \omega_{ig}\boldsymbol{b_{ig}^T}\left(W_g^2\boldsymbol{B}\right)^{-1}\boldsymbol{b_{ig}}$$

$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G} \omega_{ig}\log|\boldsymbol{\Sigma_c}| - \frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G} \omega_{ig}\boldsymbol{c_i^T}\boldsymbol{\Sigma_c^{-1}}\boldsymbol{c_i}, \tag{2.5}$$

where l is the dimension of square matrix $\boldsymbol{B}$.

2.4.1 The EM algorithm

The EM algorithm involves taking the conditional expectation of the complete-data log-likelihood and updating the parameters by maximizing the conditional

expectation. Based on (2.5), we can see that the E step at o^{th} iteration involves evaluating the following conditional expectations for each subject:

$$E_{\boldsymbol{\Psi}^{(o)}}(\omega_{ig}|\boldsymbol{y_i});\ E_{\boldsymbol{\Psi}^{(o)}}(\omega_{ig}\boldsymbol{b_{ig}}|\boldsymbol{y_i});\ E_{\boldsymbol{\Psi}^{(o)}}(\omega_{ig}\boldsymbol{b_{ig}}\boldsymbol{b_{ig}^T}|\boldsymbol{y_i});$$
$$E_{\boldsymbol{\Psi}^{(o)}}(\omega_{ig}\boldsymbol{c_i}|\boldsymbol{y_i});\ E_{\boldsymbol{\Psi}^{(o)}}(\omega_{ig}\boldsymbol{c_i}\boldsymbol{c_i^T}|\boldsymbol{y_i});\ E_{\boldsymbol{\Psi}^{(o)}}(\omega_{ig}\boldsymbol{c_i}\boldsymbol{b_{ig}^T}|\boldsymbol{y_i}).$$

Calculation of the first conditional expectation is straightforward:

$$\begin{aligned} E_{\boldsymbol{\Psi}^{(o)}}(\omega_{ig}|\boldsymbol{y_i}) &= \Pr(\omega_{ig}=1|\boldsymbol{y_i}) \\ &= \frac{\pi_{ig}f_{ig}(\boldsymbol{y_i})}{\sum_{h=1}^{G}\pi_{ih}f_{ih}(\boldsymbol{y_i})} = \tau_{ig}^{(o)}, \end{aligned} \tag{2.6}$$

which is the posterior probability of subject i belonging to latent class g at the current parameter estimate. In addition,

$$\begin{aligned} E_{\boldsymbol{\Psi}^{(o)}}(\omega_{ig}\boldsymbol{b_{ig}}|\boldsymbol{y_i}) &= E_{\boldsymbol{\Psi}^{(o)}}(\boldsymbol{b_{ig}}|\boldsymbol{y_i},\omega_{ig}=1) \\ &= E_{\boldsymbol{\Psi}^{(o)}}(\boldsymbol{b_{ig}}|\boldsymbol{y_i},\omega_{ig}=1)\Pr(\omega_{ig}=1|\boldsymbol{y_i}) \\ &= \tau_{ig}^{(o)}E_{\boldsymbol{\Psi}^{(o)}}(\boldsymbol{b_{ig}}|\boldsymbol{y_i},\omega_{ig}=1) \end{aligned}$$

Therefore, similarly, we only need to calculate:

$$E_{\boldsymbol{\Psi}^{(o)}}(\boldsymbol{b_{ig}}|\boldsymbol{y_i});\ E_{\boldsymbol{\Psi}^{(o)}}(\boldsymbol{b_{ig}}\boldsymbol{b_{ig}^T}|\boldsymbol{y_i});\ E_{\boldsymbol{\Psi}^{(o)}}(\boldsymbol{c_i}|\boldsymbol{y_i});\ E_{\boldsymbol{\Psi}^{(o)}}(\boldsymbol{c_i}\boldsymbol{c_i^T}|\boldsymbol{y_i});\ E_{\boldsymbol{\Psi}^{(o)}}(\boldsymbol{c_i}\boldsymbol{b_{ig}^T}|\boldsymbol{y_i}).$$

The joint distribution of $\left(\boldsymbol{y}_{\boldsymbol{i}}^{\boldsymbol{T}}, \boldsymbol{b}_{\boldsymbol{ig}}^{\boldsymbol{T}}, \boldsymbol{c}_{\boldsymbol{i}}^{\boldsymbol{T}}\right)^{\boldsymbol{T}}$ is a multivariate normal distribution with mean:

$$\begin{bmatrix} \boldsymbol{X_{1i}(t)\beta_g} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

and variance matrix:

$$\begin{bmatrix} \boldsymbol{Z_i(t)W_g^2BZ_i(t)^T + V_i\Sigma_C V_i^T + \Sigma_e} & \boldsymbol{Z_i(t)W_g^2B} & \boldsymbol{V_i\Sigma_c} \\ \boldsymbol{\left(Z_i(t)W_g^2B\right)^T} & \boldsymbol{W_g^2B} & \mathbf{0} \\ \boldsymbol{(V_i\Sigma_c)^T} & \mathbf{0} & \boldsymbol{\Sigma_c} \end{bmatrix}$$

Let $\boldsymbol{\Sigma_{ibcg}} = \begin{bmatrix} \boldsymbol{Z_i(t)W_g^2B} & \boldsymbol{V_i\Sigma_c} \end{bmatrix}$ and $\boldsymbol{\Sigma_{bcg}} = \begin{bmatrix} \boldsymbol{W_g^2B} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma_c} \end{bmatrix}$; therefore the joint distribution of $\boldsymbol{b_{ig,}c_i}$ condition on $\boldsymbol{y_i}$ is a multivariate normal distribution with mean

$$E_{\boldsymbol{\Psi^{(o)}}}\left(\boldsymbol{b_{ig}, c_i}\middle|\boldsymbol{y_i}\right) = \boldsymbol{\Sigma_{ibcg}^T\left(Z_i(t)W_g^2BZ_i(t)^T + V_i\Sigma_C V_i^T + \Sigma_e\right)^{-1}(y_i - X_{1i}(t)\beta_g)} \tag{2.7}$$

And variance-covariance matrix:

$$var_{\boldsymbol{\Psi^{(o)}}}\left(\boldsymbol{b_{ig}, c_i}\middle|\boldsymbol{y_i}\right) = \boldsymbol{\Sigma_{bcg} - \Sigma_{ibcg}^T\left(Z_i(t)W_g^2BZ_i(t)^T + V_i\Sigma_c V_i^T + \Sigma_\varepsilon\right)^{-1}\Sigma_{ibcg}}$$

$$= E_{\boldsymbol{\Psi^{(o)}}}\begin{bmatrix} \boldsymbol{b_{ig}b_{ig}^T}|\boldsymbol{y_i} & \boldsymbol{b_{ig}c_i^T}|\boldsymbol{y_i} \\ \boldsymbol{c_ib_{ig}^T}|\boldsymbol{y_i} & \boldsymbol{c_ic_i^T}|\boldsymbol{y_i} \end{bmatrix} - \begin{bmatrix} \boldsymbol{b_{ig}^{(o)}b_{ig}^{(o)^T}} & \boldsymbol{b_{ig}^{(o)}c_i^{(o)^T}} \\ \boldsymbol{c_i^{(o)}b_{ig}^{(o)^T}} & \boldsymbol{c_i^{(o)}c_i^{(o)^T}} \end{bmatrix}$$

(2.8)

From (2.8):

$$E_{\boldsymbol{\Psi}^{(o)}}\begin{bmatrix} \boldsymbol{b_{ig}b_{ig}^T}|\boldsymbol{y_i} & \boldsymbol{b_{ig}c_i^T}|\boldsymbol{y_i} \\ \boldsymbol{c_ib_{ig}^T}|\boldsymbol{y_i} & \boldsymbol{c_ic_i^T}|\boldsymbol{y_i} \end{bmatrix}$$

$$= \boldsymbol{\Sigma_{bcg}} - \boldsymbol{\Sigma_{ibcg}^T}\left(\boldsymbol{Z_i(t)}W_g^2\boldsymbol{BZ_i(t)^T} + \boldsymbol{V_i\Sigma_cV_i^T} + \boldsymbol{\Sigma_\varepsilon}\right)^{-1}\boldsymbol{\Sigma_{ibcg}}$$

$$+ \begin{bmatrix} \boldsymbol{b_{ig}^{(o)}b_{ig}^{(o)^T}} & \boldsymbol{b_{ig}^{(o)}c_i^{(o)^T}} \\ \boldsymbol{c_i^{(o)}b_{ig}^{(o)^T}} & \boldsymbol{c_i^{(o)}c_i^{(o)^T}} \end{bmatrix} \tag{2.9}$$

Thus, all conditional expectations can be obtained from (2.9).

Implementing the M-step is relatively trivial since there exist closed-form solutions to the maximization of the conditional expectation of the complete-data log-likelihood for the majority of the parameters except for $\gamma_g^{(o+1)}$ in the model for $\tau_{ig}^{(o)}$, which has to be updated numerically. For all other parameters, closed-form solutions are available and are given below:

$$\sigma_{\varepsilon k}^{2(o+1)} = \frac{\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{j=1}^{n_{ik}}\tau_{ig}^{(o)}\left(y_{ij} - \boldsymbol{X_{1ij}(t)\beta_g^{(o)}} - \boldsymbol{Z_{ij}(t)b_{ig}^{(o)}} - \boldsymbol{V_{ij}c_i^{(o)}}\right)^2}{\sum_{i=1}^{N}\sum_{g=1}^{G}\sum_{j=1}^{n_{ik}}\tau_{ig}^{(o)}}$$

$$\sigma_{ck}^{2(o+1)} = \frac{\sum_{i=1}^{N}\sum_{g=1}^{G}\tau_{ig}^{(o)}\left(c_{ik}^{(o)}\right)^2}{\sum_{i=1}^{N}\sum_{g=1}^{G}\tau_{ig}^{(o)}}$$

$$W_g^{2(o+1)} = \frac{\sum_{i=1}^{N}\tau_{ig}^{(o)}\boldsymbol{b_{ig}^{(o)^T}B^{(o)^{-1}}b_{ig}^{o}}}{\sum_{i=1}^{N} l * \tau_{ig}^{(o)}}$$

$$\boldsymbol{B^{(o+1)}} = \frac{\sum_{i=1}^{N}\sum_{i=1}^{G}\tau_{ig}^{(o)}\,\boldsymbol{b_{ig}^{(o)}b_{ig}^{(o)^T}}}{\sum_{i=1}^{N}\sum_{g=1}^{G}\tau_{ig}^{(o)}W_g^{2(o)}}$$

$$\boldsymbol{\beta}_{\boldsymbol{g}}^{(\boldsymbol{o+1})} = \left\{\sum_{i=1}^{N} \tau_{ig}^{(o)} \boldsymbol{X_{1i}(t)}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{(\boldsymbol{o})^{-1}} \boldsymbol{X_{1i}(t)}\right\}^{-1} \sum_{i=1}^{N} \tau_{ig}^{(o)} \boldsymbol{X_{1i}(t)}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{(\boldsymbol{o})^{-1}} (\boldsymbol{y_i} - \boldsymbol{X_{1i}(t)} \boldsymbol{b}_{\boldsymbol{ig}}^{(\boldsymbol{o})} - \boldsymbol{V_i} \boldsymbol{c}_{\boldsymbol{i}}^{(\boldsymbol{o})})$$

The E-step and M-step will be repeated until the difference of observed likelihood becomes smaller than a pre-specified threshold. For model fitting and parameter estimation we used SAS PROC IML. Initial parameter estimates used in the iterative program were obtained using SAS PROC NLMIXED without any random effects. To avoid local maxima, different initial parameter values around the estimates from PROC NLMIXED were used. Variance covariance matrix was calculated using the negative inversion of Hessian matrix by using the SAS function NLPFDD through the finite-differences method.

2.4.2 Posterior classification and model selection

Assignment of each subject into latent classes can be achieved by using the posterior probability defined in (2.6) and estimated based on the maximum likelihood estimates of the parameters. A subject is classified in the latent class for which he or she has the highest posterior probability. These posterior probabilities are then used to evaluate the degree to which the latent classes can be distinguished by the data [44]. Specifically, we will calculate a $G \times G$ classification table, with each row representing the average posterior probabilities for each latent class among subjects assigned to a given latent class [44]. High diagonal values close to 1 and low off-diagonal values close to 0 indicate good classification quality.

The number of latent classes for each data set is unknown and needs to be pre-specified before each model estimation procedure. Many model selection procedures can be used to select the "best" model when varying number of latent classes are used. As suggested by many studies, Bayesian information criterion (BIC) [25] will be used to select the number of latent classes due to its ease of implementation and superior performances [11, 12, 14].

2.5 Simulation Studies

For the simulation study, we focused on parameter estimation and latent class classification, i.e., whether we can identify the true trajectories and assign individuals into the correct latent classes. For each domain, linear trajectory was assumed. In addition to a time variable, a binary variable and a continuous variable were simulated as covariates for the domain specific fixed effects. For class specific random effects, both intercept and slope were assumed. To determine the latent class membership, one continuous variable was simulated. Since this model aimed at analyzing longitudinal data, for each sample, one to three observations at different time points were simulated.

Five scenarios were used with the assumed number of latent classes between 2 and 6. For each scenario, we generated 500 replications with each replication consists of 1500 subjects. Under each scenario we fitted a latent trajectory model with the true number of latent classes, e.g. for data that consists of 4 latent classes, only 4-class model was fitted.

Table 2.1 shows the results of parameter estimates for the 2 to 6-class model. It appears that our proposed method yields adequate parameter estimates and standard error

estimates for all model parameters. Table 2.2 includes the average coverage probabilities from 95% confidence intervals of all model parameters and misclassification rates. Coverage estimates are defined as the percentage of times that the 95% confidence interval of an estimated parameter contains the true parameter across all replications. Misclassification rate is calculated as the percentage of samples that are assigned into the wrong latent class according to the posterior probability. In our simulations, misclassification rates ranged from almost 0 to 13.97% with the trend that misclassification rates increase with the increase in the number of latent classes. This is expected since there is more room for classification error when there are more latent classes. In addition, for a fixed sample size, when the number of latent classes increases, the number of samples within each latent class decreases leading to increased standard errors estimates of class-specific parameters. Classification error hence increases with less well separated classes.

Table 2.1: Mean parameter estimates, asymptotic standard error (SE) and empirical standard error from 500 replications in simulations for 2 to 6-class model

A: 2-class model

Parameter	True Value	Mean estimates	Asymptotic SE	Empirical SE
γ_{11}	6.00	6.02	0.32	0.32
γ_{12}	-12.00	-12.05	0.61	0.62
β_{11}	5.60	5.58	0.13	0.13
β_{12}	1.60	1.57	0.04	0.05
β_{13}	1.60	1.60	0.12	0.12
β_{14}	1.60	1.60	0.19	0.20
β_{15}	1.20	1.18	0.13	0.14
β_{16}	1.20	1.16	0.04	0.06
β_{17}	1.20	1.20	0.13	0.13
β_{18}	1.20	1.19	0.20	0.22
β_{21}	-5.60	-5.58	0.13	0.13
β_{22}	-1.60	-1.56	0.04	0.05
β_{23}	-1.60	-1.60	0.12	0.12
β_{24}	-1.60	-1.60	0.20	0.20
β_{25}	-1.20	-1.18	0.13	0.13
β_{26}	-1.20	-1.17	0.04	0.05
β_{27}	-1.20	-1.20	0.12	0.12
β_{28}	-1.20	-1.18	0.21	0.21
$\sigma_{\varepsilon 1}$	0.80	0.80	0.02	0.02
$\sigma_{\varepsilon 2}$	1.20	1.20	0.02	0.02
$\sigma_{\varepsilon 3}$	0.90	0.90	0.02	0.02
$\sigma_{\varepsilon 4}$	1.10	1.10	0.02	0.02
σ_{c1}	1.10	1.10	0.04	0.04
σ_{c2}	0.90	0.90	0.05	0.05
σ_{c3}	1.20	1.20	0.04	0.04
σ_{c4}	0.80	0.79	0.06	0.06
B_1	1.20	1.19	0.05	0.05
B_2	1.40	1.40	0.03	0.03
B_3	1.30	1.29	0.05	0.05
B_4	1.50	1.50	0.03	0.03
B_{12}	0.65	0.65	0.04	0.03
B_{13}	0.15	0.15	0.05	0.05
B_{14}	0.15	0.15	0.04	0.04
B_{23}	0.15	0.15	0.04	0.04
B_{24}	0.15	0.15	0.03	0.03
B_{34}	0.58	0.58	0.03	0.03

W_2^2	0.90	0.90	0.05	0.05

B: 3-class model

Parameter	True Value	Mean estimates	Asymptotic SE	Empirical SE
γ_{11}	11.00	11.05	0.68	0.69
γ_{12}	-14.00	-14.13	0.74	0.72
γ_{21}	7.00	6.99	0.69	0.70
γ_{22}	-7.00	-6.99	0.63	0.63
β_{11}	5.60	5.57	0.13	0.14
β_{12}	1.60	1.57	0.04	0.06
β_{13}	1.60	1.60	0.13	0.13
β_{14}	1.60	1.60	0.22	0.22
β_{15}	1.20	1.17	0.15	0.15
β_{16}	1.20	1.16	0.04	0.07
β_{17}	1.20	1.20	0.13	0.14
β_{18}	1.20	1.22	0.24	0.24
β_{21}	-5.60	-5.57	0.14	0.15
β_{22}	-1.60	-1.57	0.05	0.06
β_{23}	-1.60	-1.60	0.14	0.14
β_{24}	-1.60	-1.62	0.23	0.24
β_{25}	-1.20	-1.17	0.16	0.16
β_{26}	-1.20	-1.17	0.05	0.07
β_{27}	-1.20	-1.20	0.14	0.15
β_{28}	-1.20	-1.19	0.26	0.25
β_{31}	6.00	5.98	0.15	0.15
β_{32}	2.00	1.97	0.05	0.06
β_{33}	2.00	2.00	0.14	0.13
β_{34}	2.00	2.00	0.25	0.23
β_{35}	1.60	1.58	0.15	0.15
β_{36}	1.60	1.57	0.05	0.06
β_{37}	1.60	1.59	0.15	0.14
β_{38}	1.60	1.61	0.24	0.24
$\sigma_{\varepsilon 1}$	0.80	0.80	0.01	0.01
$\sigma_{\varepsilon 2}$	1.20	1.20	0.02	0.02
$\sigma_{\varepsilon 3}$	0.90	0.90	0.02	0.01
$\sigma_{\varepsilon 4}$	1.10	1.10	0.02	0.02

σ_{c1}	1.10	1.10	0.04	0.04
σ_{c2}	0.90	0.89	0.04	0.04
σ_{c3}	1.20	1.20	0.04	0.04
σ_{c4}	0.80	0.80	0.05	0.05
B_1	1.20	1.19	0.05	0.05
B_2	1.40	1.40	0.03	0.04
B_3	1.30	1.29	0.05	0.05
B_4	1.50	1.50	0.04	0.04
B_{12}	0.65	0.65	0.03	0.03
B_{13}	0.58	0.58	0.03	0.03
B_{14}	0.15	0.16	0.05	0.05
B_{23}	0.15	0.15	0.04	0.04
B_{24}	0.15	0.15	0.04	0.03
B_{34}	0.15	0.15	0.03	0.03
W_2^2	0.90	0.90	0.05	0.05
W_3^2	0.80	0.80	0.05	0.05

C: 4-class model

Parameter	True Value	Mean estimates	Asymptotic SE	Empirical SE
γ_{11}	15.00	15.16	1.48	1.60
γ_{12}	-14.00	-14.17	1.10	1.19
γ_{21}	12.00	12.15	1.46	1.58
γ_{22}	-9.00	-9.13	1.03	1.12
γ_{31}	8.00	8.04	1.24	1.36
γ_{32}	-6.00	-6.03	0.79	0.88
β_{11}	5.60	5.60	0.16	0.17
β_{12}	1.60	1.60	0.04	0.07
β_{13}	1.60	1.60	0.15	0.15
β_{14}	1.60	1.60	0.26	0.26
β_{15}	1.50	1.49	0.17	0.18
β_{16}	1.50	1.50	0.04	0.08
β_{17}	1.50	1.50	0.16	0.16
β_{18}	1.50	1.51	0.27	0.28
β_{21}	7.60	7.60	0.16	0.17
β_{22}	-1.60	-1.60	0.04	0.07
β_{23}	-1.60	-1.60	0.15	0.15
β_{24}	-1.60	-1.60	0.24	0.26

β_{25}	-1.50	-1.51	0.17	0.18
β_{26}	-1.50	-1.50	0.04	0.08
β_{27}	-1.50	-1.50	0.17	0.16
β_{28}	-1.50	-1.49	0.28	0.28
β_{31}	4.00	3.99	0.27	0.27
β_{32}	1.20	1.20	0.06	0.11
β_{33}	1.20	1.20	0.23	0.24
β_{34}	1.20	1.22	0.42	0.42
β_{35}	1.10	1.11	0.28	0.29
β_{36}	1.10	1.10	0.06	0.12
β_{37}	1.10	1.10	0.27	0.26
β_{38}	1.10	1.09	0.45	0.46
β_{41}	6.00	6.00	0.16	0.16
β_{42}	-1.20	-1.20	0.04	0.06
β_{43}	-1.20	-1.19	0.15	0.15
β_{44}	-1.20	-1.20	0.25	0.25
β_{45}	-1.10	-1.10	0.17	0.17
β_{46}	-1.10	-1.10	0.04	0.06
β_{47}	-1.10	-1.10	0.15	0.16
β_{48}	-1.10	-1.12	0.27	0.27
$\sigma_{\varepsilon 1}$	0.80	0.80	0.01	0.01
$\sigma_{\varepsilon 2}$	1.20	1.20	0.02	0.02
$\sigma_{\varepsilon 3}$	0.90	0.90	0.01	0.01
$\sigma_{\varepsilon 4}$	1.10	1.10	0.02	0.02
σ_{c1}	1.10	1.10	0.04	0.04
σ_{c2}	0.90	0.90	0.04	0.04
σ_{c3}	1.20	1.20	0.04	0.04
σ_{c4}	0.80	0.80	0.05	0.05
B_1	1.20	1.19	0.05	0.06
B_2	1.40	1.40	0.04	0.04
B_3	1.30	1.29	0.06	0.06
B_4	1.50	1.50	0.05	0.04
B_{12}	0.65	0.65	0.04	0.04
B_{13}	0.58	0.59	0.03	0.03
B_{14}	0.15	0.15	0.06	0.06
B_{23}	0.15	0.15	0.04	0.04
B_{24}	0.15	0.15	0.04	0.04
B_{34}	0.15	0.15	0.03	0.03
W_2^2	0.90	0.90	0.07	0.06
W_3^2	0.80	0.80	0.08	0.08
W_4^2	0.60	0.60	0.05	0.04

D: 5-class model

Parameter	True Value	Mean estimates	Asymptotic SE	Empirical SE
γ_{11}	20.00	20.38	1.67	1.77
γ_{12}	-15.00	-15.28	1.23	1.31
γ_{21}	17.00	17.35	1.62	1.73
γ_{22}	-11.00	-11.23	1.12	1.19
γ_{31}	14.00	14.26	1.35	1.42
γ_{32}	-8.00	-8.15	0.74	0.77
γ_{41}	10.00	10.14	1.10	1.18
γ_{42}	-5.00	-5.07	0.54	0.58
β_{11}	5.60	5.61	0.18	0.19
β_{12}	1.60	1.60	0.05	0.08
β_{13}	1.60	1.60	0.17	0.17
β_{14}	1.60	1.60	0.28	0.28
β_{15}	1.50	1.51	0.18	0.19
β_{16}	1.50	1.50	0.05	0.08
β_{17}	1.50	1.51	0.18	0.18
β_{18}	1.50	1.48	0.31	0.30
β_{21}	7.60	7.60	0.26	0.26
β_{22}	-1.60	-1.60	0.06	0.11
β_{23}	-1.60	-1.61	0.25	0.23
β_{24}	-1.60	-1.60	0.39	0.41
β_{25}	-1.50	-1.51	0.27	0.27
β_{26}	-1.50	-1.50	0.06	0.12
β_{27}	-1.50	-1.49	0.25	0.25
β_{28}	-1.50	-1.49	0.45	0.44
β_{31}	4.00	4.00	0.23	0.25
β_{32}	1.20	1.20	0.06	0.10
β_{33}	1.20	1.21	0.22	0.22
β_{34}	1.20	1.18	0.37	0.38
β_{35}	1.10	1.09	0.26	0.27
β_{36}	1.10	1.10	0.06	0.11
β_{37}	1.10	1.09	0.25	0.24
β_{38}	1.10	1.13	0.40	0.42
β_{41}	6.00	6.02	0.19	0.20
β_{42}	-1.20	-1.20	0.05	0.08

β_{43}	-1.20	-1.19	0.18	0.18
β_{44}	-1.20	-1.23	0.30	0.30
β_{45}	-1.10	-1.09	0.19	0.20
β_{46}	-1.10	-1.10	0.05	0.08
β_{47}	-1.10	-1.10	0.19	0.19
β_{48}	-1.10	-1.12	0.32	0.32
β_{51}	3.00	3.01	0.21	0.22
β_{52}	1.40	1.40	0.05	0.11
β_{53}	1.40	1.40	0.20	0.20
β_{54}	1.40	1.39	0.34	0.34
β_{55}	-1.30	-1.29	0.22	0.23
β_{56}	-1.30	-1.30	0.05	0.10
β_{57}	-1.30	-1.31	0.22	0.21
β_{58}	-1.30	-1.30	0.37	0.36
$\sigma_{\varepsilon 1}$	0.80	0.80	0.01	0.01
$\sigma_{\varepsilon 2}$	1.20	1.20	0.02	0.02
$\sigma_{\varepsilon 3}$	0.90	0.90	0.01	0.01
$\sigma_{\varepsilon 4}$	1.10	1.10	0.02	0.02
σ_{c1}	1.10	1.10	0.04	0.04
σ_{c2}	0.90	0.90	0.05	0.05
σ_{c3}	1.20	1.20	0.04	0.04
σ_{c4}	0.80	0.80	0.05	0.05
B_1	1.20	1.19	0.06	0.06
B_2	1.40	1.40	0.04	0.04
B_3	1.30	1.29	0.06	0.06
B_4	1.50	1.50	0.05	0.05
B_{12}	0.65	0.66	0.04	0.04
B_{13}	0.58	0.59	0.03	0.03
B_{14}	0.15	0.15	0.06	0.06
B_{23}	0.15	0.15	0.04	0.04
B_{24}	0.15	0.15	0.04	0.04
B_{34}	0.15	0.15	0.03	0.03
W_2^2	0.90	0.90	0.08	0.08
W_3^2	0.80	0.79	0.08	0.08
W_4^2	0.60	0.60	0.05	0.05
W_5^2	1.10	1.10	0.09	0.09

E: 6-class model

Parameter	True Value	Mean estimates	Asymptotic SE	Empirical SE
γ_{11}	21.00	21.31	1.60	1.85
γ_{12}	-14.00	-14.26	1.04	1.12
γ_{21}	20.00	20.28	1.56	1.83
γ_{22}	-10.00	-10.18	0.81	0.93
γ_{31}	17.00	17.21	1.47	1.72
γ_{32}	-8.00	-8.11	0.67	0.78
γ_{41}	14.00	14.17	1.43	1.63
γ_{42}	-6.00	-6.07	0.59	0.68
γ_{51}	10.00	10.11	1.31	1.37
γ_{52}	-4.00	-4.04	0.51	0.53
β_{11}	5.60	5.61	0.31	0.32
β_{12}	1.60	1.60	0.07	0.13
β_{13}	1.60	1.58	0.29	0.28
β_{14}	1.60	1.60	0.49	0.50
β_{15}	1.50	1.50	0.31	0.34
β_{16}	1.50	1.50	0.08	0.14
β_{17}	1.50	1.50	0.30	0.30
β_{18}	1.50	1.49	0.51	0.55
β_{21}	7.60	7.60	0.17	0.18
β_{22}	-1.60	-1.61	0.05	0.07
β_{23}	-1.60	-1.60	0.15	0.16
β_{24}	-1.60	-1.59	0.26	0.27
β_{25}	-1.50	-1.51	0.17	0.18
β_{26}	-1.50	-1.50	0.05	0.08
β_{27}	-1.50	-1.50	0.18	0.17
β_{28}	-1.50	-1.49	0.28	0.29
β_{31}	4.00	3.99	0.26	0.30
β_{32}	1.20	1.20	0.06	0.12
β_{33}	1.20	1.19	0.24	0.25
β_{34}	1.20	1.21	0.42	0.46
β_{35}	1.10	1.09	0.29	0.33
β_{36}	1.10	1.09	0.07	0.13
β_{37}	1.10	1.10	0.27	0.27
β_{38}	1.10	1.11	0.48	0.52
β_{41}	6.00	6.03	0.29	0.32
β_{42}	-1.20	-1.20	0.07	0.12
β_{43}	-1.20	-1.20	0.26	0.27
β_{44}	-1.20	-1.23	0.45	0.49

β_{45}	-1.10	-1.08	0.28	0.32
β_{46}	-1.10	-1.10	0.06	0.12
β_{47}	-1.10	-1.10	0.28	0.29
β_{48}	-1.10	-1.13	0.46	0.54
β_{51}	3.00	3.00	0.24	0.25
β_{52}	1.40	1.40	0.06	0.13
β_{53}	1.40	1.41	0.22	0.23
β_{54}	1.40	1.39	0.39	0.40
β_{55}	-1.30	-1.31	0.25	0.28
β_{56}	-1.30	-1.30	0.07	0.12
β_{57}	-1.30	-1.30	0.25	0.25
β_{58}	-1.30	-1.30	0.42	0.45
β_{61}	6.00	6.01	0.24	0.24
β_{62}	-1.40	-1.40	0.06	0.11
β_{63}	-1.40	-1.40	0.22	0.22
β_{64}	-1.40	-1.41	0.37	0.38
β_{65}	1.30	1.29	0.25	0.27
β_{66}	1.30	1.30	0.06	0.12
β_{67}	1.30	1.31	0.24	0.24
β_{68}	1.30	1.30	0.42	0.42
$\sigma_{\varepsilon 1}$	0.80	0.80	0.01	0.01
$\sigma_{\varepsilon 2}$	1.20	1.20	0.02	0.02
$\sigma_{\varepsilon 3}$	0.90	0.90	0.01	0.01
$\sigma_{\varepsilon 4}$	1.10	1.10	0.02	0.02
σ_{c1}	1.10	1.10	0.04	0.04
σ_{c2}	0.90	0.90	0.05	0.05
σ_{c3}	1.20	1.20	0.04	0.04
σ_{c4}	0.80	0.80	0.06	0.05
B_1	1.20	1.18	0.07	0.07
B_2	1.40	1.39	0.07	0.07
B_3	1.30	1.28	0.08	0.08
B_4	1.50	1.49	0.07	0.07
B_{12}	0.65	0.66	0.04	0.04
B_{13}	0.58	0.59	0.03	0.03
B_{14}	0.15	0.15	0.06	0.06
B_{23}	0.15	0.15	0.04	0.04
B_{24}	0.15	0.15	0.04	0.04
B_{34}	0.15	0.15	0.03	0.03
W_2^2	0.90	0.91	0.10	0.10
W_3^2	0.80	0.80	0.11	0.11
W_4^2	0.60	0.60	0.09	0.09
W_5^2	1.10	1.12	0.12	0.13

W_6^2	1.05	1.06	0.12	0.12

Note: Asymptotic standard error is the average of SE from 500 replications; Empirical SE is standard error of estimates from 500 replications; $\gamma, \sigma_\varepsilon,\ \sigma_c,\ \beta, W^2, B$ are defined in section 2.3; B_1 to B_4 are the square roots of the diagonal elements of matrix B, and all other B parameters are correlation coefficients of the corresponding terms as indicated by numbers in the subscripts.

Table 2.2: Average coverage probabilities of 95% confidence intervals for all parameters in a given model and misclassification rates of simulation results

Number of classes	Average coverage (range)	Misclassification rate
2	94.89% (92.60%-98.80%)	0.001%
3	95.16% (92.80%-99.00%)	2.29%
4	95.61% (91.40%-99.80%)	8.60%
5	95.73% (93.00%-100.00%)	12.29%
6	96.15% (92.60%-100.00%)	13.97%

2.6 Application to the UDS data:

The proposed multivariate finite mixture latent trajectory model was applied to study longitudinal patterns of cognitive decline among dementia patients using data from the UDS in the NACC data repository. The UDS is an ongoing data collection that was implemented in 2005 at 34 past and present NIA-founded Alzheimer's Disease Centers (ADC) around the country [40]. Patients were recruited into the ADCs and followed annually to collect information relevant to aging and dementia, including performance

measures on neuropsychological tests in multiple cognitive domains such as memory and language [45, 46].

The sample used for the analysis included Caucasian patients with dementia who had at least four annual cognitive evaluations. We also restricted our analyses to those whose cognitive decline began after 60 years of age in order to exclude patients with early onset dementia. Tests from two cognitive domains were used: logical memory immediate and delayed recalls tests for the memory domain; Animal Fluency Test and the Boston Naming Test for the language domain. As indicated by Weintraub et al, age of onset, gender and education had significant effects on test scores and were included in both the class membership model and the latent trajectory model [46]. Final analysis data set included 30,004 observations from 1517 patients, of whom 52.74% were male with 15.07 mean years of education and 73.33 as mean age of onset. Since these four test scores have different ranges, all outcomes were rescaled to be between 0 and 10 to achieve computational efficiency. In addition, education (in number of years) and age of onset (in years) were rescaled to be between 0 and 1. The time variable, age, was measured by decades and centered on the mean age. We tested linear and quadratic trajectories with the assumed number of latent classes ranging from 2 to 6. We present the estimated log likelihood and BIC for all models in table 2.3.

Table 2.3: Estimated log likelihoods and BICs in the UDS data for various models assuming different numbers of latent classes

	Linear Trajectory			Quadratic Trajectory		
number of classes	number of parameters	Log Likelihood	BIC	number of parameters	Log Likelihood	BIC
2	43	-47711.57	95738.10	47	-47675.48	95695.22
3	58	-47394.82	95214.45	64	-47329.63	95128.02
4	73	-46992.81	94520.31	81	-46971.34	94535.97
5	88	-46871.30	94387.15	98	-46832.84	94383.48
6	103	-46791.45	94337.31	115	-46713.23	94268.78

It can be seen that the differences between linear and quadratic models are small relative to the complexity of the models. Therefore, we chose the linear model for its ease of interpretation. The decrease in BIC was more pronounced when the number of latent classes increased from 2 to 4, but the BIC values became relatively flat with 4 or more latent classes; thus, we chose the model with 4 latent classes as the final model following the recommended practice by several authors [11, 12, 14]. Parameters estimates in the multinomial model for latent class membership and for the fixed effects in the latent trajectory models are presented in table 2.4.

Table 2.4: Parameter estimates for the latent class memebership model and for the fixed effects in the latent trajactory model

		Class 1		Class 2		Class 3		Class 4	
Models	Parameter	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Multinomial									
	Intercept	3.08	0.63	4.71	0.63	5.53	0.61	0	Ref
	Sex	-0.20	0.28	-0.63	0.28	-0.20	0.32	0	Ref
	Education	-1.65	0.99	-3.99	0.97	-5.98	0.80	0	Ref
	Age of onset	-4.16	1.16	-1.98	1.19	-1.52	1.72	0	Ref
Trajectory									
Memory domain	Intercept	-2.28	0.26	-0.37	0.36	-0.14	0.10	-9.19	1.58
	Linear slope	-2.18	0.12	-1.90	0.16	-0.50	0.04	-8.42	0.39
	Sex	-0.08	0.08	-0.29	0.17	-0.03	0.03	-0.39	0.85
	Education	0.19	0.34	1.40	0.40	0.02	0.12	0.34	1.93
	Age of onset	8.50	0.54	6.30	0.89	1.92	0.21	34.32	2.01
Language domain	Intercept	-6.98	1.51	1.49	0.53	0.22	0.41	-0.73	0.37
	Linear slope	-8.17	0.17	-2.35	0.14	-2.92	0.11	-4.91	0.30
	Sex	-0.26	0.33	-0.38	0.16	-0.35	0.14	-0.40	0.54
	Education	0.31	1.87	2.28	0.66	2.25	0.57	-1.47	1.02
	Age of onset	31.10	0.97	7.41	0.70	9.67	0.65	19.03	1.63

Note: male is the reference for gender.

In figure 2.1, we present the predicted trajectories of male patients with 15 years of education and age of onset at 73 (chosen as the sample means) in four latent classes. Latent class 1 has the steepest decline in language but relatively flat in memory decline; latent class 4 has the fastest decline in memory and also the second fastest decline in

language; patients in latent classes 2 and 3 have less decline in both language and memory domains than those in latent classes 1 and 4.

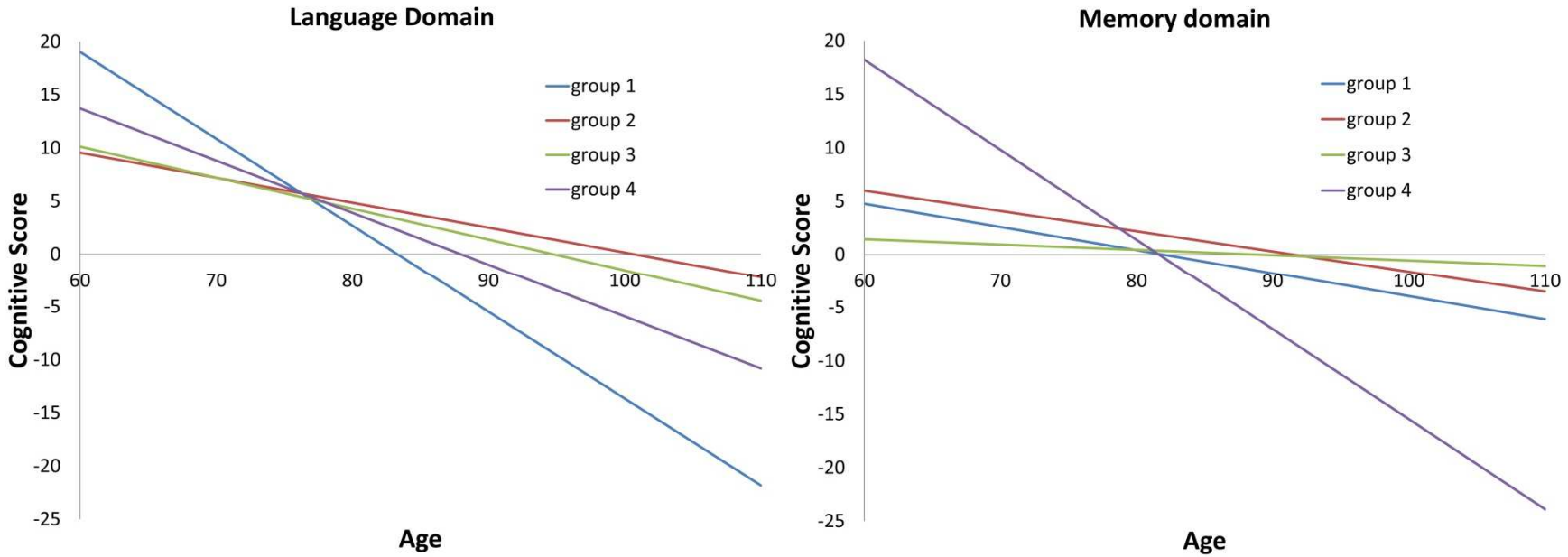

Figure 2.1: Estimated trajectories of language (left) and memory (right) decline for male dementia patients with education and age of onset at the sample means in four latent classes.

We further examined the association between patient characteristics and the four identified latent classes and present the results in Table 2.5. Since APOE e4 allele is an important risk factor for AD and about 60% of AD patients carry this allele [47-50], we also included percentages of samples having it in each latent class. Latent classes 3 and 2 captured the majority of patients followed by latent class 1 and 4. More than 70% of the patients in latent classes 1 and 3 were clinically diagnosed as probable AD only as defined by NINCDS-ADRDA criteria [51, 52] and no any other forms of dementia; while class 2 had the lowest percentage of probable AD only (41.37%) and the highest percentage of other dementia. Patients in latent class 1 also had the highest percentage of being an APOE e4 carrier compared to patients in the other classes. For latent classes 2 and 4, about half of samples have other types of dementia and not surprisingly, less than half of samples have APOE e4 allele. However, just as there were differences between latent classes 1 and 3, latent classes 2 and 4 also differ in gender composition, years of education, and age of onset pointing to potentially different etiologies.

Table 2.5: Patients characteristics by the four identified latent classes.

class	number of patients	male %	Average years of Education (SD)	Average age of onset (SD)	APOE e4 carrier (%)	Probable AD only (%)	Other Dementia (%)
1	300	54.00	15.91(2.89)	70.66(6.56)	72.69	76.00	22.67
2	510	59.80	15.18(3.30)	73.35(7.21)	44.57	41.37	55.88
3	560	46.96	14.16(3.24)	74.03(6.88)	57.79	71.96	27.14
4	147	47.62	16.38(2.10)	76.06(9.00)	47.06	51.02	42.86

To evaluate model fit, the average posterior probabilities for the linear model with 4 latent classes are presented in table 2.6. The diagonal values are all greater than 0.84, indicating a good separation of the four latent classes.

Table 2.6: Average posterior probabilities of 4 latent classes identified

Classified class	1	2	3	4
1	0.87	0.04	0.07	0.02
2	0.02	0.88	0.04	0.06
3	0.05	0.10	0.84	0.00
4	0.02	0.09	0.00	0.89

2.7 Discussion

In this chapter, we proposed a multivariate finite mixture latent trajectory model aiming at analyzing data that are often encountered in dementia studies. In these studies, there exist latent constructs of multiple domains, each of which may be measured by multiple neuropsychological tests. Our model is an extension of GBTM, GMM and the non-linear latent class model proposed by Proust-Lima et al, and can be used in studies where more than one test for the same underlying variable is used. We applied our method to the UDS data and identified four latent cognitive decline patterns.

Given that multiple cognitive measures are routinely collected in dementia and aging studies, appropriate statistical models with realistic assumptions for multiple tests in more than one domain is extremely important. The naïve method of analyzing data with multiple tests by modeling each test with a separate latent trajectory can lead to the

identification of many latent classes and without combining the information across different cognitive domains. There exist some methods that aim at reducing dimensionality by combining tests within the same domain using sum or weighted average test scores. However, as indicated by Gray and Brookmeyer, data reduction may cause loss of information and the results may be difficult to interpret [53]. In our method, by jointly modeling tests within the same domain, the numbers of model parameters is greatly reduced. By adding random test-specific effects, the difference and correlation among tests are accounted for. Furthermore, since these tests are measurements of the same underlying latent construct, joint modeling can improve our ability to identify the true latent construct.

The identified latent classes can be used for therapeutic and research purpose. Since patients in the same latent classes share similar cognitive decline patterns, this can help care providers and clinicians for better patients care and treatment. In addition, patients in each latent class may share the same disease etiology and may be caused by same genes; therefore the power in genetic studies that look for genes related dementia can be improved. For example, based on recent summary at ALzGene database, 695 genes related AD are found from 1395 studies, however, only a few of them are confirmed by multiple studies [47, 48]. The reason for this is, although patients are all clinically diagnosed as AD, their cognitive decline patters differ dramatically and this heterogeneity makes results from a given study hard to replicate thus the ability to find true genes is greatly reduced. Our method can be used to find samples that have similar cognitive decline patterns and genetics studies from these phenotypically homogenous sub groups will be more comparable.

In our model, we assumed normal distribution due to its tractability and ease of implementation. However, the normal distribution assumption may not apply for tests that have categorical or binary responses. In the future, we will extend our work to model non-normal variables and/or mixed types variables. Another limitation of using normal distribution lies in selecting the number of classes using information based criterion like BIC. It has been observed in this and many other studies that BIC is always decreasing as more classes are added [11, 19]. This problem is more pronounced when the sample size is large and sample sizes in each class are imbalanced. In these cases, the latent classes with larger sample sizes can be split into two or more latent classes [11] and currently the best way to address this problem is using background information to help model selection.

A common problem encountered in many dementia studies is floor or ceiling effects associated with some test scores. Proust et al proposed a transformation for the test scores using cumulative beta distribution and they demonstrated that the transformation fits the data well [16]. Jacqmin-Gadda et al also proposed a semi-parametric latent process model to address the problem of different sensitivities of tests at different dementia stages [54]. Future research will be needed to extend our models to handle these additional challenges.

CHAPTER 3. INFORMATION BASED CRITERIA FOR MODEL SELECTION IN FINITE MIXTURE LATENT TRAJECTORY MODELS: A SIMULATION STUDY

3.1 Summary

A challenge in finite mixture latent trajectory models is the selection of the number of classes. In this chapter, we performed a simulation study to compare the performance of information-based model selection criteria including the commonly used Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and other less commonly used information criteria such as consistent AIC (CAIC), sample adjusted BIC and integrated classification likelihood criteria (ICLBIC). These model selection criteria were compared across different scenarios with varying number of subjects, the number of observations for each subject, and the level of separation between latent classes. We found that the level of separation had substantial impact on the performances of model selection criteria. Sample adjusted BIC performed uniformly better in all scenarios and is therefore recommended for the selection of number of subpopulations for multivariate finite mixture latent trajectory models.

3.2 Introduction

In biomedical research, data are often collected from a heterogeneous population consisting of several unobserved subpopulations. For example, dementia patients exhibit considerable heterogeneity in their longitudinal trajectory of cognitive function, with some patients showing rapid decline following dementia diagnosis while others show slower decline or may even remain stable for several years [1]. Patients' cognitive trajectories also differ across cognitive domains, with some patients showing more rapid

decline in memory while others showing faster decline in language or executive function [2-9]. In our previous work, we have proposed a multivariate finite mixture latent trajectory model to identify patient subgroups with similar longitudinal patterns of cognitive decline using data from multiple cognitive domains. Both simulation studies and an application to a data set from dementia studies showed that the proposed model can accommodate the complexity of the data and has the capability to uncover the heterogeneity of the population. However, one unresolved issue in the application of the proposed model is the determination of the number of subpopulations.

Selecting the number of unobserved subpopulations is a critical but challenging issue for latent mixture models. Many studies have been devoted to this topic [11, 18-22]. However, there has been no well-established approach thus far. Information criteria (IC) based approaches have been commonly used in model selection for latent class models, with Akaike's Information Criterion (AIC) [24] and Bayesian Information Criterion (BIC) [25] being the most popular methods due to their ease of implementation. In the literature of finite mixture models, in particular for latent class modeling and growth mixture modeling, there were several simulation studies on the performance of IC-based fit indices [11, 19-22]. These studies have shown that AIC tends to overestimate the number of groups, especially when sample size is large [19, 20]. BIC is recommended by many researchers [12, 19, 20], and it yielded the best performance in several simulation studies [11, 20]. However, in some situations, BIC also selected larger numbers of groups than necessary [26]. In addition, it has been observed in several studies that BIC may decrease monotonically as more groups were added [11]. In addition to AIC and BIC, there are other IC-based fit indices including consistent AIC (CAIC) [27]; and sample adjusted

BIC (SABIC) [28-31]. They were proposed to augment the performance of AIC and BIC, and in several simulation studies, these modified indices showed considerable improvement, especially SABIC, which is the best in several recent studies [18, 21, 22].

In addition to above likelihood-based information criteria, classification-based information criteria have also been developed to measure the accuracy of classification, for example, the classification likelihood criterion [55], normalized entropy criterion [56], partition coefficient [57], and integrated classification likelihood criterion (ICLBIC) [32]. As reviewed by McLachlan and Peel, these methods either have restriction on the mixing proportions or unsatisfactory performance with the exception of ICLBIC [19, 58, 59]. ICLBIC was proposed by Biernacki et al and the goal is to correct the overestimation problem of BIC [32]. Therefore, it was referred to as ICLBIC in McLachlan and Peel's simulation studies [19]. ICLBIC was found to outperform all other information criteria including AIC and BIC in McLachlan and Peel's simulation study [19].

Another type of approach for determining the number of latent classes is likelihood-based approach. The commonly used likelihood ratio test (LRT) cannot be used to compare models with differing number of latent classes because the null hypothesis involves zero mixing proportions, resulting in parameters on the boundary of parameter space and hence violating the regularity condition [19]. Alternative likelihood-based approaches include the Lo, Mendell and Rubin (LMR) test [23] and the bootstrap likelihood ratio test (BLRT) [19]. Application of these likelihood-based tests has been limited due to the high computational burden. We therefore restrict our simulation study to IC-based approaches.

Although IC-based approaches have been widely used in many applications and studied in literature, their performance was evaluated in different contexts such as latent class modeling and growth mixture modeling. Performance of these indices for multivariate finite mixture latent trajectory models is not known. In this study, we will perform a simulation study to evaluate the performance of IC-based indices under different conditions with varying number of subjects, number of observations per subject, and level of separation among latent classes. The rest of this chapter is organized as follows. Section 3.3 presents an overview of the multivariate finite mixture latent trajectory model. Section 3.4 introduces IC-based fit indices considered in our study. In Section 3.5, we describe the simulation study and present the results. A brief discussion is presented in Section 3.6.

3.3 The Multivariate Finite Mixture Latent Trajectory Model

This model has been described in detail previously and will only be briefly introduced here. Our work was directly motivated by studies of cognitive decline among dementia patients, in which multiple neuropsychological tests are typically performed to characterize patients' level of cognition in several cognitive domains. Our model allows more than one latent quantity, each of which can be measured by multiple tests, and identifying subpopulations of patients who exhibit distinct longitudinal patterns in these latent quantities.

Assume that the population consists of G subpopulations represented by G latent classes. For individual $i,\ i = 1, \dots, N$, we define a G-dimentional vector $\boldsymbol{\omega_i}$ denoting the latent class membership, with $\omega_{ig} = 1$ if individual i belongs to class g and 0 otherwise.

Suppose there are K neuropsychological tests with continuous outcomes representing cognitive function in D cognitive domains. Let $\boldsymbol{y_i} = \left(\boldsymbol{y_{i1}^T}, \dots, \boldsymbol{y_{ik}^T}, \dots, \boldsymbol{y_{iK}^T}\right)^{\boldsymbol{T}}$ be the vector of all measurements for individual i, where $\boldsymbol{y_{ik}}$ is a vector of length n_{ik}, which denotes the number of longitudinal measurements for individual i and test k $(k = 1, \dots, K)$, hence the length of $\boldsymbol{y_i}$ is $\sum_{k=1}^{K} n_{ik}$. Let $\boldsymbol{X_{1i}(t)}$ and $\boldsymbol{Z_i(t)}$ be the matrices of covariates collected for individual i. $\boldsymbol{Z_i(t)}$ can have partial or all columns of $\boldsymbol{X_{1i}(t)}$ but contains at least one time variable. Then a measurement model if individual i is in latent class g, $g = 1, \dots, G,$ is:

$$\boldsymbol{y_{i|\omega_{ig}=1}} = \boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)} + \boldsymbol{V_i c_i} + \boldsymbol{\varepsilon_i}, \tag{3.1}$$

Where the latent trajectory is defined as:

$$\boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)} = \boldsymbol{X_{1i}(t)\beta_g} + \boldsymbol{Z_i(t)b_{ig}}, \tag{3.2}$$

The length of latent process $\boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)}$ is also $\sum_{k=1}^{K} n_{ik}$. Note that for tests that are in the same domain, they share the same latent process by having the same values in $\boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)}$. $\boldsymbol{\beta_g}$ is the vector of class-specific fixed effects from all cognitive domains in latent class g. $\boldsymbol{b_{ig}}$ is the class specific random effects for all domains in latent class g. We assume that $\boldsymbol{b_{ig}}$ has a multivariate normal distribution $N\left(\boldsymbol{0}, W_g^2 \boldsymbol{B}\right)$ with $W_1^2 = 1$ and $\boldsymbol{B}$ is the covariance matrix of first latent class, similarly defined as in Proust et al [16]. $\boldsymbol{c_i}$ in (3.1) is the K-vector of test-specific random intercept. It introduces correlation among scores of the same test from the same individual. Here we assume $\boldsymbol{c_i}$ is distributed as $N(\boldsymbol{0}, \boldsymbol{\Sigma_c})$, where $\boldsymbol{\Sigma_c}$ is a diagonal matrix with σ_{ck}^2 in its diagonal. $\boldsymbol{\varepsilon_i}$ in (3.1) is an vector of random error with distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma_\varepsilon})$, where $\boldsymbol{\Sigma_\varepsilon}$ is a block matrix with $\sigma_{\varepsilon k}^2 \boldsymbol{I}_{n_{ik}}$ at diagonal and all other entries are 0s.

For individual i, the probability that this individual belongs to a latent class g is π_{ig}, with $\sum_{g=1}^{G} \pi_{ig} = 1$. This can be modeled through a multinomial logistic regression as:

$$\pi_{ig} = P(\omega_{ig} = 1 | \boldsymbol{X_{2i}^T}) = \frac{\exp(\boldsymbol{X_{2i}^T \gamma_g})}{1+\sum_{h=1}^{G-1} \exp(\boldsymbol{X_{2i}^T \gamma_h})}, \tag{3.3}$$

where $\boldsymbol{\gamma_g}$ is the vector of the class-specific regression coefficients. For identifiability purpose, $\boldsymbol{\gamma_G}$ are set to 0s. Covariates $\boldsymbol{X_{2i}^T}$ used here can be the same or different from $\boldsymbol{X_{1i}(t)}$ in equation (3.2). Let

$$\boldsymbol{\Psi} = (\boldsymbol{\beta_1}, \dots, \boldsymbol{\beta_g}, \dots, \boldsymbol{\beta_G}, W_2^2, \dots, W_g^2, \dots, W_G^2, \boldsymbol{B}, \boldsymbol{\Sigma_c}, \boldsymbol{\Sigma_\varepsilon}, \boldsymbol{\gamma_2}, \dots, \boldsymbol{\gamma_g}, \dots, \boldsymbol{\gamma_G})$$

be the parameters to be estimated, $f_{ig}(\boldsymbol{y_i})$ be the density function of $\boldsymbol{y_i}$ in latent class g, then the observed-data likelihood is:

$$L(\boldsymbol{\Psi}) = \prod_{i=1}^{N} \sum_{g=1}^{G} \pi_{ig} f_{ig}(\boldsymbol{y_i})$$

$f_{ig}(\boldsymbol{y_i})$ has distribution $N(\boldsymbol{X_{1i}(t)\beta_g}, \boldsymbol{\Sigma_{ig}})$, where

$$\boldsymbol{\Sigma_{ig}} = \boldsymbol{Z_i(t)} W_g^2 \boldsymbol{B Z_i(t)^T} + \boldsymbol{V_i \Sigma_c V_i^T} + \boldsymbol{\Sigma_\varepsilon}$$

Since the latent class memberships are unobserved and there are also multiple random effects, the expectation-maximization (EM) algorithm can be used for obtaining

parameter estimates [19, 41-43]. The EM algorithm evaluates the conditional expectation of the complete-data log-likelihood and the parameters are updated by maximizing the conditional expectation. Closed-form solutions to the maximization of the conditional expectation of the complete-data log-likelihood for the majority of parameters are available except for γ_g in the model for τ_{ig}, which has to be updated numerically. The E-step and M-step will be repeated until the difference of observed likelihood becomes smaller than a pre-specified threshold and the likelihood at the last step is used for the calculation of information criteria.

3.4 Information Criteria Surveyed

In our study, we will consider the commonly used indices including AIC, BIC, as well as CAIC, SABIC, and ICLBIC. CAIC, SABIC and ICLBIC were chosen because they showed better results than AIC and/or BIC in previous simulation studies [19-21]. AIC is the first information criterion proposed and derived by Akaike in early 1970's by using Kullback-Leibler measure [24, 60]. It is still widely used today and defined as:

$$AIC = -2logL(\boldsymbol{\Psi}) + 2p$$

where p is the number of model parameters and $L(\boldsymbol{\Psi})$ is the likelihood as defined previously. It is not consistent in the sense that it only penalizes on the number of parameters no matter what the sample size is. To overcome this problem, Bozdogan proposed CAIC, in which $log(N) + 1$ was used as the multiplier instead of 2 [27].:

$$CAIC = -2logL(\boldsymbol{\Psi}) + p(log(N) + 1)$$

where N is the number of independent subjects.

BIC is proposed by Schwarz within the Baysian framework and it is the most used information criterion [25]. BIC is defined as:

$$BIC = -2logL(\boldsymbol{\Psi}) + p(\log(N))$$

Compared to AIC, BIC penalizes both numbers of parameters and samples. CAIC is similar to BIC but puts extra penalty on the number of parameters.

Using the minimum description length (MDL) principle in computational learning theory, Rissanen proposed an information criterion that is basically BIC but the sample size N is replaced by an adjusted sample size N^* [29-31], which is defined as:

$$N^* = (N + 2)/24$$

It is referred as adjusted BIC in Nylund et al study [20], ADBIC in Kim's study [18], SABIC in Tofighi and Enders study [21]. We will refer this index as SABIC in our study. For comparison purpose, Tofighi and Enders also used N^* in CAIC and called it SACAIC [21]. Although it is not the best overall index, since SACAIC had better performance than BIC in their study in several settings [21], we will also include it in our study.

Biernacki et al noticed that sometimes BIC chose more components than it should and to overcome this overestimation problem, they proposed ICLBIC [32]. It requires an additional penalty term called entropy. Let τ_{ig} be the posterior probability of individual i belonging to group $g, g = 1, \dots, G$, then entropy $EN(\tau)$ is defined as:

$$EN(\tau) = -\sum_{i=1}^{N}\sum_{g=1}^{G} \tau_{ig}\log(\tau_{ig})$$

and ICLBIC is [32]:

$$ICLBIC = -2logL(\boldsymbol{\Psi}) + 2EN(\tau) + plog(N)$$

Since the entropy term is always positive thus ICLBIC also has larger penalty than BIC. Similar to what Tofighi and Enders did in their study, it is of scientific interest to see ICLBIC's performance if we replaced N with N^*, therefore, we will also include it in our simulation study and refer it as SAICLBIC.

3.5 Simulation

The purpose of this simulation study is to evaluate the overall performance of the IC-based indices in terms of how often each index correctly identifies the true number of latent classes. We also investigated the performance of these indices under different factors, and in the case that a wrong model is chosen, what the direction of model misspecification is, i.e. whether more or less number of classes is chosen.

Data were simulated based on the multivariate finite mixture latent trajectory model described in section 3.3. In this simulation study, we assumed 4 tests from 2 different domains with each domain having 2 tests. Tests within each domain shared the same fixed effect but have different test specific random effects. In addition, there were also domain-specific random effects. Fixed effects in each domain include a linear trajectory over time, as well as a binary covariate and a continuous covariate. Domain-specific random effects include intercept and slope, while only random intercept was included for test specific random effect. The latent class membership was associated with one continuous covariate.

We tested performance of IC-based fit indices under considerations of three factors: number of subjects in each data set, number of observations for each subject, and level of separation between groups. In simulation I, we examined the performance of fit

indices assuming 1000, 1500 and 2000 subjects with each subject having1 to 3 observations. In simulation II, we examined the performance of fit indices assuming 1 to 3 observations per subject and 1 to 4 observations per subject with 1500 subjects in each data set. In Simulation III, we assumed 1500 subjects in each data set with 1 to 3 observations for each subject, and with two different levels of group separations. Since each latent class in our model is determined jointly by latent trajectory and multinomial model, we used expected misclassification rate to measure the class separation. It is defined as the percentage of samples that could be assigned into the wrong latent class according to the posterior probability. High expected misclassification rate indicates the latent classes are close to each other and it is difficult to distinguish them, therefore, it has low class separation. Two expected misclassification rates, 7.76% and 15.44%, were used for high and low class separation respectively. For simulations I and II, we also used high class separation. Since in our previous application of multivariate finite mixture latent trajectory model on dementia data, we identified four distinct cognitive decline classes, therefore, all our data were simulated under a 4-class model.

For every scenario, we simulated 500 data sets. For each data set, we fitted five models with 2, 3, 4, 5, and 6 latent classes and compared each fit index across these five models. The number of classes associated with the lowest index value will be selected for each index. If the model with the lowest value of the fit index is the true model (4-class model), then the fit index is said to have correctly identified the number of latent classes. For each simulation setting, we calculated the percentages each class number was selected using the IC-based criteria. The percentage of selecting four classes represents

the rate of correctly selecting the true model while the percentage of selecting other class numbers represents model selection errors (model misidentification or misspecification).

Simulation I, II and III results are presented in tables 3.1 to 3.3 respectively. Table 3.1 shows effects of different numbers of subjects on performance of these indices. Except AIC, which sample size N is not used, and both versions of ICLBIC, which almost never chose the correct numbers of classes, with the increase of subjects, the number of correct identification increased, especially CAIC and BIC when samples size changed from 1000 to 1500. For SACAIC and SABIC, since the number of correct identification was close to perfect, the effect of number of subjects was not obvious. The worse performance for CAIC and BIC with 1000 subjects is not a surprise and this agrees with the study of Tofighi and Enders [21]. They observed that BIC had bad performance in complicated models when sample size is small. For multivariate finite mixture latent trajectory models, the number of parameters is usually large, e.g. in our simulation, there are 59 parameters in the 4-class mode, therefore, although there were 1000 subjects, it still cannot be considered as a big sample.

The effects of number of observations are presented in table 3.2. Four indices are either already having perfect performance (SACAIC and SABIC) or 0 correct identification (both versions of ICLBIC), therefore, varying the number of observations had no effect on them. For all others, by increasing the number of observations, the performance is more or less better, and surprisingly, AIC has the biggest improvement. The small increase of performance of CAIC and BIC is expected because for longitudinal data, effective sample size is increased with the increase of number of observation,

although sample size is unchanged. Again, this agrees with the results of Tofighi and Enders, which also found numbers of observations have small impact [21].

The result of different class separation is in table 3.3. In our study, the level of class separation has the biggest effect on the performance of these indices. When class separation is low, the performance dropped dramatically. CAIC and BIC have 98.4% and 92.2% misidentification rates respectively; SACAIC only correctly identified a little bit more than half of the data sets. The only indices with acceptable performance are AIC and SABIC. The bad performance of BIC is contradicting with Nylund et al [20] but agrees with Tofighi and Enders [21]. The possible explanation for this is the model complexity. Both ours and Tofighi and Enders had more complicated models than Nylund et al.

The directions of misidentification are also presented in tables 3.1 to 3.3. Except the scenario when number of subjects is 1000, AIC overestimated the number of classes as widely observed. CAIC and SACAIC corrected this overestimation problem by penalizing on sample size; however, CAIC obviously over corrected it and now it has underestimation problem. Similarly BIC suffers a little bit over correction problem and SABIC seems perfect except when class separation is low. Both versions of ICLBIC penalized too much and most of time, only 2-class model, the simplest one we tested, was chosen, just as observed in Nagin's study [11].

As expected, AIC didn't have the best performance in every scenario; however, in all scenarios except low separation, it has correct identification rates >94%. In addition, its performance did not vary as dramatically as some other indices, such as BIC and CAIC. For low class separation AIC has more than 20% misidentification rate; however,

every index had bad performance in that scenario and AIC is almost as good as SABIC, which has the best performance. Its consistent version, CAIC has comparable performance for 1500 subjects, 2000 subjects and 1 to 4 observations with AIC but much worse performance for 1000 subjects and low class separation. Sample adjusted CAIC greatly improved the performance, with perfect or almost perfect in all scenarios except low class separation. BIC is slightly better than AIC in many scenarios but shares the similar pattern as CAIC due to the similarity of their formula: the performance in 1000 subjects and low group separation is unacceptable. However, compared with BIC, the extra penalty on number of parameters in CAIC obviously degraded its performance. Sample adjusted BIC, just as observed in several other simulations studies, is the winner of this study. It outperformed all indices in all scenarios. Surprisingly, the performance of ICLBIC, whether sample adjusted or not, misidentified almost all data sets in all scenarios. This is contradicting to the study of McLachlan and Peel, in which ICLBIC was the best [19] but agrees with Nagin's study [11].

Table 3.1: Percentage of the lowest value of indices in each model fit under different numbers of subjects

Criteria	1000 Subjects					1500 Subjects					2000 Subjects				
	2-g	3-g	4-g	5-g	6-g	2-g	3-g	4-g	5-g	6-g	2-g	3-g	4-g	5-g	6-g
AIC	0	0.4	**96.0**	1.6	2	0	0	**94.4**	2.2	3.4	0	0	**98.6**	0.6	0.8
CAIC	72.4	0	**27.6**	0	0	4.8	0	**95.2**	0	0	0.2	0	**99.8**	0	0
SACAIC	1.2	0	**98.8**	0	0	0	0	**100**	0	0	0	0	**100**	0	0
BIC	35.4	0	**64.6**	0	0	1.0	0	**99.0**	0	0	0	0	**100**	0	0
SABIC	0.4	0	**99.6**	0	0	0	0	**100**	0	0	0	0	**100**	0	0
ICLBIC	99.6	0	**0.4**	0	0	100	0	**0**	0	0	100	0	**0**	0	0
SAICLBIC	99.4	0	**0.6**	0	0	100	0	**0**	0	0	100	0	**0**	0	0

Note: different numbers of subjects under a 4-class model with each subject having 1-3 observations. The bold numbers are percentages of data sets that correct number of classes were chosen.

Table 3.2: Percentage of the lowest value of indices in each model fit under different numbers of observations for each subject.

Criteria	1-3 Observations					1-4 Observations				
	2-g	3-g	4-g	5-g	6-g	2-g	3-g	4-g	5-g	6-g
AIC	0	0	**94.4**	2.2	3.4	0	0	**99.4**	0.2	0.4
CAIC	4.8	0	**95.2**	0	0	2.2	0	**97.8**	0	0
SACAIC	0	0	**100**	0	0	0	0	**100**	0	0
BIC	1.0	0	**99.0**	0	0	0	0	**100**	0	0
SABIC	0	0	**100**	0	0	0	0	**100**	0	0
ICLBIC	100	0	**0**	0	0	100	0	**0**	0	0
SAICLBIC	100	0	**0**	0	0	100	0	**0**	0	0

Note: different numbers of observations for each subject under a 4-classs model with 1500 subjects in each data set. The bold numbers are percentages of data sets that correct number of classes were chosen.

Table 3.3: Percentage of the lowest value of indices in each model fit under high and low class separation

Criteria	High Separation					Low Separation				
	2-g	3-g	4-g	5-g	6-g	2-g	3-g	4-g	5-g	6-g
AIC	0	0	**94.4**	2.2	3.4	0	0	**78.0**	15.8	6.2
CAIC	4.8	0	**95.2**	0	0	0	98.4	**1.4**	0.2	0
SACAIC	0	0	**100**	0	0	0	46.4	**53.4**	0.2	0
BIC	1.0	0	**99.0**	0	0	0	92.2	**7.6**	0.2	0
SABIC	0	0	**100**	0	0	0	21.6	**78.2**	0.2	0
ICLBIC	100	0	**0**	0	0	69.2	30.8	**0**	0	0
SAICLBIC	100	0	**0**	0	0	61.0	39.0	**0**	0	0

Note: high and low group separation under a 4-class model with 1500 subjects in each data set and each subject having 1-3 observations. The bold numbers are percentages of data sets that correct number of classes were chosen.

Since SABIC had the best performance in all scenarios, we calculated it in our applications of the Uniform Data Set (UDS) from the National Alzheimer's Coordinating Center (NACC) [40]. We observed the same trend as BIC and 4-class model still seems reasonable. This is expected because there are 1507 subjects in that data set and each subject has at least 4 visits. From the right part of table 3.2, which has the similar sample size as our real data, we can see BIC and SABIC all have perfect correct identification rates.

3.6 Conclusion and Discussion

In this chapter, we performed a simulation study to investigate the performances of most commonly used IC-based fit indices including AIC, BIC, as well as CAIC, sample adjusted CAIC, sample adjusted BIC, ICLBIC, and sample adjusted ICLBIC for selecting the number of latent classes in multivariate finite mixture latent trajectory model. We also investigated the effects of number of subjects, number of observations, and level of separation between classes on their performance. We found SABIC performed uniformly better in all situations and level of class separation has the biggest impact on their performance.

Among the two popular IC-based criteria, AIC outperformed BIC. Although AIC has a slight overestimation problem, it has correct identification rate > 94% in all scenario except low class separation, in which it had the second best performance. BIC did not perform as well as AIC when sample size is small or when class separation is low. SABIC greatly improved the performance of BIC and similar improvement was seen for CAIC and SACAIC over that of AIC. SABIC had the best performance in all simulation

scenarios and should be the preferred model selection method for latent trajectory models. Classification based information criteria had the worst performance in all scenarios and in most cases they only selected the simplest model. Therefore, they shouldn't be used in complex models.

In this study, information criteria were tested under our multivariate latent trajectory model, therefore, generalization of our conclusion to other models should be preceded with caution, especially for less complicated models. In addition, the simulation setting was based on our analysis of NACC data sets and we only tested true model with 4 classes. For data sets that are much simpler or more complicated, the behavior of these indices may be different. Also in this simulation, each class had similar numbers of samples, i.e. they are balanced, however, in real data analysis, some classes may have extreme bigger or smaller sample size than other classes. In the future, we will expand our simulation study to more latent class models with more simulation settings to give a general guidance for model selection in finite mixture modeling.

CHAPTER 4. A MULTIVARIATE FINITE MIXTURE LATENT TRAJECTORY MODEL WITH PARTIALLY LABELLED DATA: SIMULATIONS AND APPLICATION TO DEMENTIA STUDIES

4.1 Summary

In Chapter 2 we developed a multivariate finite mixture latent trajectory model aimed at identifying the classes of subjects sharing the same multivariate longitudinal trajectories. In practice, a subset of subjects may have known class memberships, e.g. dementia patients may have known subtype if they underwent autopsy after death. These data are referred as the labelled data and adding this information can improve the model's ability to classify the remaining data with unknown latent class membership. In this study, we first extended the multivariate finite mixture latent trajectory model by incorporating the partially labelled information. Then we performed simulation studies to investigate the effect of adding labelled information to our model under different considerations. Results showed that the performances were improved in all scenarios, even in situations where there were only 10% of data labelled, and latent classes were not well separated. We also re-analyzed the Uniform Data Set (UDS) from National Alzheimer's Coordinating Center (NACC) by adding pathological information. Compared with our previous analysis, we found that with labelled data the newly defined latent classes can be more phenotypically homogenous.

4.2 Introduction

Dementia is a common disease among the elderly population and is characterized by the impairment of cognitive function. According to different disease etiology, there are several subtypes of dementia, such as Alzheimer's disease (AD), vascular dementia

(VD), Frontotemporal dementia (FTD), and Lewy Body Dementia (LBD), etc [39]. Accurate clinical diagnosis of dementia subtype is critical for therapeutic intervention and for scientific investigations. However, the exact dementia subtype is often defined by pathological finding after patients' death. Previous studies have demonstrated differences in dementia profiles by various dementia subtypes with AD patients having dominant problems in memory [2-4] and FTD patients showing more deficit in executive function [5-9]. Based on those findings, we have developed a multivariate finite mixture latent trajectory model using patients' longitudinal neuropsychological test results. Using data from patients in the Uniform Data Set (UDS) collected by the Alzheimer's Disease Centers (ADCs) across the nation [40], this model identified four distinct cognitive decline patterns.

Although it is difficult to determine the exact dementia subtype for all subjects, there is a subsample of patients in the UDS with known dementia subtype based on the pathological data. These data were obtained through autopsy. Methodological approaches that incorporate these dementia subtypes are attractive because information from these patients could potentially improve the accuracy of inferring patients' unknown dementia subtype. Such data are often called partially labelled data in the latent class analysis literature [19, 33-36]. By combining labelled and unlabeled data, the classification rule established from labelled data is updated and will have better performance [19, 36], or the classification rule will be more accurate than using labelled data or unlabeled alone [19, 33, 34]. In their studies, Hosmer, Hosmer and Dick found that when only 10% of the data were labelled, accuracy of classification was improved [37, 38]. This improved performance was acquired mostly by the co-existence of features that were not captured

by using labelled data or unlabeled data only [33-35]. Beside better classification abilities, the existence of labelled data can make estimation faster, i.e. improve its efficiency [19]. In addition, since most methods used numerical estimations and the values of starting points played an important role in finding the global maxima, labelled data can be analyzed first if sample size is sufficiently large, and the results can be used as starting points for subsequent analyses.

In this chapter, we extend the multivariate finite mixture latent trajectory model, originally proposed to identify subpopulations using multivariate longitudinal data without labelled data, to the situation where partially labelled data are available. We will perform simulation studies to investigate how labelled data can improve the classification performance of the multivariate finite mixture latent trajectory modeling under different conditions. In particular, we will examine whether observing the class membership for a small proportion of the sample improves the classification of the unlabeled cases, especially in situations where the classes are not well separated. Then we will re-analyze the UDS data utilizing the partially available pathological information on dementia subtype. The remainder of this chapter is organized as follows. Section 4.3 describes the model and estimation. Section 4.4 presents the simulation study. In Section 4.5, we apply our model to UDS data. Conclusion and discussion are in section 4.6.

4.3 Multivariate Finite Mixture Latent Trajectory Model With Partially Labelled Data

This model was directly motivated by studies of cognitive decline among dementia patients, in which multiple neuropsychological tests are typically performed to characterize patients' level of cognition in several cognitive domains. Our model allows

more than one latent quantity, each of which can be measured by multiple tests, and identifying subpopulations of patients who exhibit distinct longitudinal patterns in these latent quantities.

Assume that the population consists of G subpopulations represented by G latent classes. For individual $i,\ i = 1, \dots, N$, we define a G-dimentional vector $\boldsymbol{\omega_i}$ denoting the latent class membership, with $\omega_{ig} = 1$ if individual i belongs to class g and 0 otherwise. Suppose there are K neuropsychological tests with continuous outcomes representing cognitive function in D cognitive domains. Let $\boldsymbol{y_i} = \left(\boldsymbol{y_{i1}^T}, \dots, \boldsymbol{y_{ik}^T}, \dots, \boldsymbol{y_{iK}^T}\right)^{\boldsymbol{T}}$ be the vector of all measurements for individual i, where $\boldsymbol{y_{ik}}$ is a vector of length n_{ik}, which denotes the number of longitudinal measurements for individual i and test k $(k = 1, \dots, K)$, hence the length of $\boldsymbol{y_i}$ is $\sum_{k=1}^{K} n_{ik}$. Let $\boldsymbol{X_{1i}(t)}$ and $\boldsymbol{Z_i(t)}$ be the matrices of covariates collected for individual i. $\boldsymbol{Z_i(t)}$ can have partial or all columns of $\boldsymbol{X_{1i}(t)}$ but contains at least one time variable. Then a measurement model if individual i is in latent class g, $g = 1, \dots, G,$ is:

$$\boldsymbol{y_{i|\omega_{ig}=1}} = \boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)} + \boldsymbol{V_i c_i} + \boldsymbol{\varepsilon_i}, \tag{4.1}$$

Where the latent trajectory is defined as:

$$\boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)} = \boldsymbol{X_{1i}(t)\beta_g} + \boldsymbol{Z_i(t)b_{ig}}, \tag{4.2}$$

The length of latent process $\boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)}$ is also $\sum_{k=1}^{K} n_{ik}$. Note that for the tests that are in the same domain, they share the same latent process by having the same values in $\boldsymbol{\Lambda_{i|\omega_{ig}=1}(t)}$. $\boldsymbol{\beta_g}$ is the vector of class-specific fixed effects from all cognitive domains in latent class g. $\boldsymbol{b_{ig}}$ is the class specific random effects for all domains in latent class g. We assume that $\boldsymbol{b_{ig}}$ has a multivariate normal distribution $N\left(\boldsymbol{0}, W_g^2 \boldsymbol{B}\right)$ with $W_1^2 = 1$ and

$\boldsymbol{B}$ is the covariance matrix of first latent class, similarly defined as in Proust et al [16]. $\boldsymbol{c_i}$ in (4.1) is the K-vector of test-specific random intercept. It introduces correlation among scores of the same test from the same individual. Here we assume $\boldsymbol{c_i}$ is distributed as $N(\mathbf{0}, \boldsymbol{\Sigma_c})$, where $\boldsymbol{\Sigma_c}$ is a diagonal matrix with σ_{ck}^2 in its diagonal. $\boldsymbol{\varepsilon_i}$ in (4.1) is an vector of random error with distribution $N(\mathbf{0}, \boldsymbol{\Sigma_\varepsilon})$, where $\boldsymbol{\Sigma_\varepsilon}$ is a block matrix with $\sigma_{\varepsilon k}^2 \boldsymbol{I}_{n_{ik}}$ at diagonal and all other entries are 0s.

For individual i, the probability that this individual belongs to a latent class g is π_{ig}, with $\sum_{g=1}^{G} \pi_{ig} = 1$. This can be modeled through a multinomial logistic regression as:

$$\pi_{ig} = P(\omega_{ig} = 1 | \boldsymbol{X_{2i}^T}) = \frac{\exp(\boldsymbol{X_{2i}^T \gamma_g})}{1+\sum_{h=1}^{G-1} \exp(\boldsymbol{X_{2i}^T \gamma_h})}, \tag{4.3}$$

where $\boldsymbol{\gamma_g}$ is the vector of the class-specific regression coefficients. For identifiability purpose, $\boldsymbol{\gamma_G}$ are set to 0s. Covariates $\boldsymbol{X_{2i}^T}$ used here can be the same or different from $\boldsymbol{X_{1i}(t)}$ in equation (4.2).

Assume for these N individuals belonging to one of the G latent classes, M individuals ($M < N$) have known latent class membership (i.e. labelled data), thus for those M individuals, $\boldsymbol{\omega_i}$ is observed. Then the density function of $(\boldsymbol{\omega_i}, \boldsymbol{y_i})$ is $\pi_{ig} f_{ig}(\boldsymbol{y_i})$ if individual i is known to belong to class g. Therefore, the density function of labelled data can be written as $\prod_{g=1}^{G} \{\pi_{ig} f_{ig}(\boldsymbol{y_i})\}^{\omega_{ig}}$. If individual i does not have known class membership, then the density function of $(\boldsymbol{\omega_i}, \boldsymbol{y_i})$ is a mixture distribution:

$$\sum_{g=1}^{G} \pi_{ig} f_{ig}(\boldsymbol{y_i}).$$

Let

$$\boldsymbol{\Psi} = (\boldsymbol{\beta_1}, \dots, \boldsymbol{\beta_g}, \dots, \boldsymbol{\beta_G}, W_2^2, \dots, W_g^2, \dots, W_G^2, \boldsymbol{B}, \boldsymbol{\Sigma_c}, \boldsymbol{\Sigma_\varepsilon}, \boldsymbol{\gamma_2}, \dots, \boldsymbol{\gamma_g}, \dots, \boldsymbol{\gamma_G})$$

be the parameters to be estimated, then the observed likelihood is:

$$L(\boldsymbol{\Psi}) = \prod_{i=1}^{M} \prod_{g=1}^{G} \{\pi_{ig} f_{ig}(\boldsymbol{y_i})\}^{\omega_{ig}} \prod_{i=m+1}^{N} \sum_{g=1}^{G} \pi_{ig} f_{ig}(\boldsymbol{y_i}) \tag{4.4}$$

Since the latent class memberships for some subjects are unobserved and there are also multiple random effects in our model, the expectation-maximization (EM) algorithm can be used. Augmenting the observed data $\boldsymbol{y_i}$ with unobserved variables $(\boldsymbol{\omega_i}, \boldsymbol{b_{i1}} \dots, \boldsymbol{b_{ig}}, \dots, \boldsymbol{b_{iG}}, \boldsymbol{c_i})$, the complete-data likelihood function is:

$$L^c(\boldsymbol{\Psi}) = \prod_{i=1}^{M} \prod_{g=1}^{G} \{\pi_{ig} f(\boldsymbol{y_i}|\boldsymbol{b_{ig}}, \boldsymbol{c_i}) f(\boldsymbol{b_{ig}}) f(\boldsymbol{c_i})\}^{\omega_{ig}}$$
$$\prod_{i=M+1}^{N} \prod_{g=1}^{G} \{\pi_{ig} f(\boldsymbol{y_i}|\boldsymbol{b_{ig}}, \boldsymbol{c_i}) f(\boldsymbol{b_{ig}}) f(\boldsymbol{c_i})\}^{\omega_{ig}} \tag{4.5}$$

In (4.5), although the first and second parts look exactly same, ω_{ig} in second part is unobserved. The log likelihood is:

$$\log(L^c(\boldsymbol{\Psi})) = \sum_{i=1}^{M} \sum_{g=1}^{G} \omega_{ig} \{\log(\pi_{ig}) + \log\left(f(\boldsymbol{y_i}|\boldsymbol{b_{ig}}, \boldsymbol{c_i})\right) + \log\left(f(\boldsymbol{b_{ig}})\right) + \log(f(\boldsymbol{c_i}))\}$$
$$+ \sum_{i=M+1}^{N} \sum_{g=1}^{G} \omega_{ig} \{\log(\pi_{ig}) + \log\left(f(\boldsymbol{y_i}|\boldsymbol{b_{ig}}, \boldsymbol{c_i})\right) + \log\left(f(\boldsymbol{b_{ig}})\right) + \log(f(\boldsymbol{c_i}))\}$$
$$= \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} \log(\pi_{ig}) - \frac{\sum_{k=1}^{K} n_{ik} + l + K}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} \omega_{ig} \log|\boldsymbol{\Sigma_\varepsilon}|$$

$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G}\omega_{ig}\left(\boldsymbol{y_i}-\boldsymbol{X_{1i}(t)\beta_g}-\boldsymbol{Z_i(t)b_{ig}}-\boldsymbol{V_i c_i}\right)^{T}\boldsymbol{\Sigma_{\varepsilon}^{-1}}\left(\boldsymbol{y_i}-\boldsymbol{X_{1i}(t)\beta_g}-\boldsymbol{Z_i(t)b_{ig}}-\boldsymbol{V_i c_i}\right)$$

$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G}\omega_{ig}\log|\boldsymbol{B}|-\frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G}l*\omega_{ig}\log\left(W_g^2\right)-\frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G}\omega_{ig}\boldsymbol{b_{ig}^{T}}\left(W_g^2\boldsymbol{B}\right)^{-1}\boldsymbol{b_{ig}}$$

$$-\frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G}\omega_{ig}\log|\boldsymbol{\Sigma_c}|-\frac{1}{2}\sum_{i=1}^{N}\sum_{g=1}^{G}\omega_{ig}\boldsymbol{c_i^{T}\Sigma_c^{-1}c_i},\tag{4.6}$$

where l is the dimension of square matrix $\boldsymbol{B}$.

From (4.6), we can see at o^{th} step, we need to calculate:

$$E_{\boldsymbol{\Psi}^{(o)}}\left(\omega_{ig}|\boldsymbol{y_i}\right);\ E_{\boldsymbol{\Psi}^{(o)}}\left(\omega_{ig}\boldsymbol{b_{ig}}|\boldsymbol{y_i}\right);\ E_{\boldsymbol{\Psi}^{(o)}}\left(\omega_{ig}\boldsymbol{b_{ig}b_{ig}^{T}}|\boldsymbol{y_i}\right);$$

$$E_{\boldsymbol{\Psi}^{(o)}}\left(\omega_{ig}\boldsymbol{c_i}|\boldsymbol{y_i}\right);\ E_{\boldsymbol{\Psi}^{(o)}}\left(\omega_{ig}\boldsymbol{c_ic_i^{T}}|\boldsymbol{y_i}\right);\ E_{\boldsymbol{\Psi}^{(o)}}\left(\omega_{ig}\boldsymbol{c_ib_{ig}^{T}}|\boldsymbol{y_i}\right).$$

However, for $E_{\boldsymbol{\Psi}^{(o)}}\left(\omega_{ig}|\boldsymbol{y_i}\right)$, if an individual has known class membership, then it is either 1 or 0 depending on their latent class memberships and no calculation of the expected value is needed. For other individuals without known class membership, it will need to be evaluated; therefore, the maximization steps only involve those $E_{\boldsymbol{\Psi}^{(o)}}\left(\omega_{ig}|\boldsymbol{y_i}\right)$ from unlabeled data. Closed-form solutions to the maximization of the conditional expectation of the complete-data log-likelihood for the majority of the parameters are available except for γ_g in (4.3), which has to be updated numerically. The E-step and M-

step will be repeated until the difference of observed likelihood becomes smaller than a pre-specified threshold.

4.4 Simulation Study

The goals of the simulation study are three-fold. First, we evaluated whether labelled information improves classification accuracy. The classification accuracy is measured by misclassification rate, defined as the percentage of samples that are classified incorrectly. Second, we examined whether labelled data improves the model estimation efficiency with smaller number of EM iterations and hence faster convergence of the algorithm. Lastly, we evaluated how improvements in classification accuracy and model estimation efficiency are influenced by various factors including sample size, number of longitudinal measurements per subject, the number of latent classes, and the level of separation of the latent classes.

4.4.1 Simulation Setup

Data were simulated based on the multivariate finite mixture latent trajectory model described in section 4.3. In this simulation study, we assumed 4 tests from 2 different domains with each domain having 2 tests. Tests within each domain shared the same fixed effect but have different test specific random effects. In addition, there were also domain-specific random effects. Fixed effects in each domain include a linear trajectory over time, as well as a binary covariate and a continuous covariate. Domain-specific random effects include intercept and slope, while only the intercept was included

for the test specific random effect. The latent class membership was associated with one continuous covariate.

Performance of the model was examined in three simulation studies. In simulation I, we evaluated the model performance with varied number of subjects (N=500, 1000, and 1500), assuming 1-3 longitudinal measurements per subject and 4 latent classes. In Simulation II, we examined the performance while varying the number of longitudinal measurements per subject (1-3 observations and 1-4 observations), assuming 500 subjects and 4 latent classes. In Simulation III, we assessed the performance with varied number of latent classes (3, 4, and 5 classes), assuming 500 subjects and 1-3 longitudinal measurements per subject. For each simulation, we generated data with low and high levels of separation of the latent classes. Since each latent class in our model was determined jointly by latent trajectory and multinomial model, we used expected misclassification rate to measure the class separation. High expected misclassification rate indicates the latent classes are close to each other and it is difficult to distinguish them, therefore, it has low class separation. For every scenario, we generated 500 replicates. For each replicate, we fitted 3 models with 0%, 10%, and 20% subjects randomly selected with known class membership. All models were fitted assuming the same number of latent classes as the truth, including the model with no labelled data. After each model fitting, unlabeled subjects were assigned to their most likely latent class according to the posterior probabilities. Misclassification rates were then calculated and summarized across 500 replicates using averages and ranges. In order to evaluate whether labelled data improves the efficiency of the EM algorithm, we also reported the number of iterations it took for the algorithm to converge.

4.4.2 Simulation Results

Table 4.1 is the summary of performance with different numbers of subjects. In all settings, performance improved with more labelled data. However, when the numbers of subjects increased, the improvements were getting smaller. For example, for 1500 subjects, when expected misclassification rate was low, the improvements were almost negligible. In table 4.2, the same trend was observed when the numbers of observations increased. When sample size or number of observations is large, the effective sample sizes of both labelled and unlabeled data increase and performances are good even without labelled data; therefore, the room for improvement is small. The performance under different numbers of latent classes is shown in table 4.3. We noticed that for more classes, it was getting slower to reach the expected misclassification rates and reduction of iterations needed were also smaller. For example, for 5-class model when expected misclassification rate was high, even with 20% labelled data, the misclassification rate was still 10% higher than expected; and there was only 63% of reduction of iterations needed. The reason for this is, for same number of subjects, when adding more classes, there are more parameters and number of subjects in each class are smaller; therefore, there are more errors for estimation.

From tables 4.1 to 4.3, when expected misclassification rates were high, there were big improvements, especially from no labelled data to only 10% labelled; and when expected misclassification rates were low, although the improvements were not dramatic, there were still clear trends of better performances with increasing proportions of labelled data. The small decrease of misclassification rate when expected misclassification was low is due the fact that the misclassification rates were already close to expected even

without labelled data and again there was no much room for improvement. Just as misclassification rates, the numbers of iterations needed also dropped as the proportions of labelled data increased. And when expected misclassification rates were high, the reduction can be as high as 75%. For both misclassification rates and iterations needed, while there were big improvements from no labelled data to 10% labelled, the improvements were much smaller from 10% to 20%.

To further check where these improvements were from, we listed the average parameter estimations and their standard errors for some of parameters in table 4.4. As can be seen, with more labelled data added, while there were just marginal improvements of the average of estimations, the standard errors were several times smaller, therefore classification accuracy was improved.

Table 4.1: Misclassification rates and average iterations used in Simulation I.

	500 Subjects			1000 Subjects			1500 Subjects		
% of Labelled	% of Exp.MR	% of MR (range)	Avg.iter (range)	% of Exp.MR	% of MR (range)	Avg.iter (range)	% of Exp.MR	% of MR (range)	Avg.iter (range)
0	9.13	9.90 (6.00-18.40)	127.37 (38-318)	9.07	9.32 (6.60-12.60)	121.39 (42-266)	9.08	9.17 (6.80-11.73)	125.34 (41-265)
10		9.76 (6.18-17.98)	121.24 (10-290)		9.28 (6.12-12.58)	118.40 (22-259)		9.14 (6.77-11.40)	123.20 (14-246)
20		9.70 (5.78-14.96)	118.68 (7-268)		9.23 (5.97-12.67)	115.84 (12-252)		9.11 (6.60-12.09)	121.78 (16-248)
0	29.60	38.15 (27.60-59.40)	241.02 (76-855)	28.28	33.55 (26.20-53.50)	235.83 (67-945)	28.33	32.00 (26.67-45.33)	229.73 (61-787)
10		32.61 (25.33-49.33)	55.86 (12-417)		29.79 (24.67-36.43)	65.24 (10-544)		29.35 (25.26-34.28)	61.77 (12-505)
20		31.27 (24.81-39.90)	34.36 (9-387)		29.16 (24.65-34.70)	38.42 (9-552)		28.90 (24.15-34.04)	35.97 (7-330)

Note: Exp.MR, expected misclassification rates; MR, misclassification rates; Avg.iter, average numbers of iterations.

Table 4.2: Misclassification rates and average iterations in Simulation II.

	1-3 Observations			1-4 Observation		
% of Labelled	% of Exp.MR	% of MR (range)	Avg.iter (range)	% of Exp.MR	% of MR (range)	Avg.iter (range)
0	9.13	9.90 (6.00-18.40)	127.37 (38-318)	8.12	9.06 (5.00-15.00)	142.72 (55-375)
10		9.76 (6.18-17.98)	121.24 (10-290)		8.94 (4.40-15.56)	138.04 (41-373)
20		9.70 (5.78-14.96)	118.68 (7-268)		8.95 (4.96-16.25)	137.01 (6-368)
0	29.60	38.15 (27.60-59.40)	241.02 (76-855)	27.93	35.35 (24.00-53.80)	248.72 (77-842)
10		32.61 (25.33-49.33)	55.86 (12-417)		31.17 (23.50-47.17)	79.72 (10-561)
20		31.27 (24.81-39.90)	34.36 (9-387)		29.93 (22.66-39.39)	46.98 (7-356)

Note: Exp.MR, expected misclassification rates; MR, misclassification rates; Avg.iter, average numbers of iterations.

Table 4.3: Misclassification rates and average iterations in Simulation III

% of Labelled	3-Group			4-Group			5-Group		
	% of Exp. MR	% of MR (range)	Avg.iter (range)	% of Exp. MR	% of MR (range)	Avg.iter (range)	% of Exp. MR	% of MR (range)	Avg.iter (range)
0	7.66	8.21 (4.80-13.80)	103.79 (34-271)	9.13	9.90 (6.00-18.40)	127.37 (38-318)	12.66	14.53 (8.40-30.20)	165.70 (50-393)
10		8.17 (5.07-12.81)	101.75 (33-249)		9.76 (6.18-17.98)	121.24 (10-290)		14.13 (8.32-20.40)	150.49 (16-330)
20		8.16 (3.93-12.50)	101.97 (36-248)		9.70 (5.78-14.96)	118.68 (7-268)		13.99 (8.44-19.55)	141.31 (7-272)
0	25.87	33.06 (21.80-62.00)	212.77 (53-549)	29.60	38.15 (27.60-59.40)	241.02 (76-855)	30.26	36.55 (28.80-51.60)	262.72 (86-735)
10		27.49 (20.22-38.70)	50.33 (15-314)		32.61 (25.33-49.33)	55.86 (12-417)		34.38 (25.11-44.76)	128.44 (9-478)
20		26.78 (20.20-33.83)	34.80 (10-170)		31.27 (24.81-39.90)	34.36 (9-387)		33.53 (25.43-42.17)	97.64 (8-605)

Note: Exp.MR, expected misclassification rates; MR, misclassification rates; Avg.iter, average numbers of iterations.

Table 4.4: Selected parameter estimations and standard errors

Parameter	True Value	0% Labelled		10% Labelled		20% Labelled	
		Average estimation	SE	Average estimation	SE	Average estimation	SE
γ_{11}	15	16.90	13.58	16.19	3.15	16.15	2.96
γ_{12}	-14	-15.47	9.09	-14.99	2.30	-14.96	2.17
γ_{21}	12	13.80	13.59	13.09	3.13	13.06	2.94
γ_{22}	-9	-10.26	9.08	-9.78	2.20	-9.76	2.06
γ_{31}	8	9.45	13.59	8.80	2.75	8.76	2.61
γ_{32}	-6	-6.95	9.08	-6.52	1.79	-6.49	1.70

Note: Data were simulated under a 4-class model with 500 subjects in each data set and each subject having 1-3 observations, 9.13% expected misclassification rate. γ is as defined in section 4.3.

4.5 Application to the UDS data:

The proposed multivariate finite mixture latent trajectory model with partially labelled data was applied to the UDS data. For the purpose of comparison with our previous result when no labelled information was used, we analyzed the same data set: only Caucasian patients with dementia who had at least four cognitive evaluations were included. Again, we restricted analyses to those with cognitive decline after 60 years in order to exclude patients with early onset dementia. Tests from two domains were used: logical memory immediate recall and delayed recall tests for the memory domain; Animal Fluency Test and the Boston Naming Test for the language domain. Age of onset, gender and education were included in both the latent class membership model and the latent trajectory model as in previous study. Final analysis data set included 30,004

observations from 1517 patients, in which 52.74% are male and average years of education and age of onset are 15.07 and 73.33 respectively. Since these four test scores have different ranges, all outcomes were rescaled to be between 0 and 10 to have a similar magnitude. In addition, education (in number of years) and age of onset (in years) were rescaled to be between 0 and 1. The time variable, age, was measured by decades and centered on the mean age. For previous analysis, our final result was a linear latent trajectory model with 4 latent classes; therefore, the exactly same model was fitted.

Among the 1517 subjects we used, there were 196 subjects who had pathological data available. Most of them were primary AD only or primary AD combined with other subtypes of dementia. Some of the subjects had 3 or more subtypes of dementia. Therefore, there are many different ways to classify these subtypes. Due to the fact that the numbers of primary VD, LBD and FTD only patients were very small and most of VD, LBD and FTD patients also have AD, we divided those 196 patents into the following four classes: a. 46 patients with primary AD only and no any other subtypes of dementia; b. 52 primary or contributing VD patients and no any other subtypes of dementia except AD; c. 35 primary or contributing LBD patients and no any other subtypes of dementia except AD; d, all any other subtypes of dementia except AD, VD and LBD. Here we named classes alphabetically to distinguish them from classes 1 to 4 we identified previously when no labeled information was used.

Assignment of each subject into latent classes was achieved by using the posterior probability. A subject was classified in the latent class for which he or she has the highest posterior probability. To evaluate classification errors of the latent class assignment, we again calculated a $G \times G$ classification table as we did previously for model without

labelled data, with each row representing the average posterior probabilities for each latent class among subjects assigned to a given latent class [44]. Therefore, the diagonal part of this table is the average posterior probabilities of correct classifications. High diagonal values close to 1 and low off-diagonal values close to 0 indicate good classification accuracy. The result is presented in table 4.5. The diagonal elements are close to or large than 0.8 indicating adequate model classification performance.

Table 4.5: Average posterior probabilities of 4 latent classes identified

Classified class	a	b	c	d
A	0.83	0.02	0.11	0.04
B	0.01	0.81	0.07	0.11
C	0.07	0.07	0.82	0.04
D	0.05	0.12	0.04	0.79

To compare the class assignments of models with and without labelled data, we listed the crosstab of subjects classified in each class based on the two models in table 4.6. Table 4.6A is for all 1517 samples. Columns are classes identified with labelled data and rows are classes identified without labelled data. The classes identified by models with and without labelled data were matched by common samples in both classes, for example, for those 300 subjects assigned into latent classes 1 when no labelled data was used, by adding labelled information, 264 samples are in class a and they consist almost 70% of entire sample in class a; therefore, class 1 corresponds to class a. Similarly, classes b, c and d are corresponding to classes 2, 3 and 4 respectively. In table 4.6, if two models agree with each other, then off diagonal part will be close to 0. From table 4.6A, latent

classes 1 and 4 show higher agreement rates with classes a and d. For previously identified latent class 2, labelled information makes more than half of patients being reclassified to latent classes c and d. For previously identified latent class 3, after adding labelled information, although majority patients were in class c, some of them were reclassified to other latent classes, mostly to class a. Table 4.6B is break-up of 196 subjects with pathological information only. By comparing the numbers in tables 4.6A and 4.6B, it can be seen for classes 1 and 4; those disagreements are almost all from labelled subjects only, meaning those subjects didn't have similar trajectories with unlabeled data in classes 1 and 4, and their existences didn't change the latent class assignments, or stated in another way: class 1 is not prime AD only and class 4 is not consisted of all any other subtypes of dementia except AD, VD and LBD. For previously identified latent class 3, those reclassified into latent classes b and d after adding labelled information are almost all from labelled subjects only, however, 58 subjects now in latent class a. For previously identified latent class 2, most of subjects were reclassified into other latent classes.

Table 4.6: Comparison of class assignments of models with and without labelled information.

A.

		Partial labelled				
		a	b	c	d	Sum
No label	1	264	12	16	8	300
	2	38	192	176	104	510
	3	74	19	435	32	560
	4	9	11	3	124	147
	sum	385	234	630	268	1517

B.

		Partial labelled				
		a	b	c	d	sum
No label	1	13	12	12	8	45
	2	10	11	13	17	51
	3	16	19	8	31	74
	4	7	10	2	7	26
	sum	46	52	35	63	196

Note: A, all 1517 subjects; B, 196 subjects with pathological information only.

We presented the characteristics of these 4 latent classes in table 4.7. Not surprisingly, there was not much change for latent classes a and d compared with previous latent classes 1 and 4. However, for latent classes b and c, the percentages of patients with APOE e4 allele and clinically diagnosed AD dropped, especially for latent class b. Therefore, adding labelled information makes these two groups more

homogeneous, or put it in another way, latent class b is more likely to have VD and latent class c is more likely to have LBD. In figure 4.1 we plotted model trajectories of male patients with education and age of onset at the sample means in 4 latent classes using linear model for memory domain and language domain. Again, there were obvious changes for latent classes b and c.

Table 4.7: Characteristics of 4 latent classes identified

Latent class	Male %	Average years of Education (SD)	Average age of onset (SD)	APOE e4 carrier (%)	Probable AD only (%)
a	50.91	15.75(2.93)	71.46(6.79)	70.18	74.09
b	61.11	14.64(3.29)	72.94(7.45)	37.86	32.77
c	51.43	14.34(3.33)	73.41(7.01)	55.91	64.98
d	51.12	16.17(2.60)	76.19(7.71)	47.37	53.53

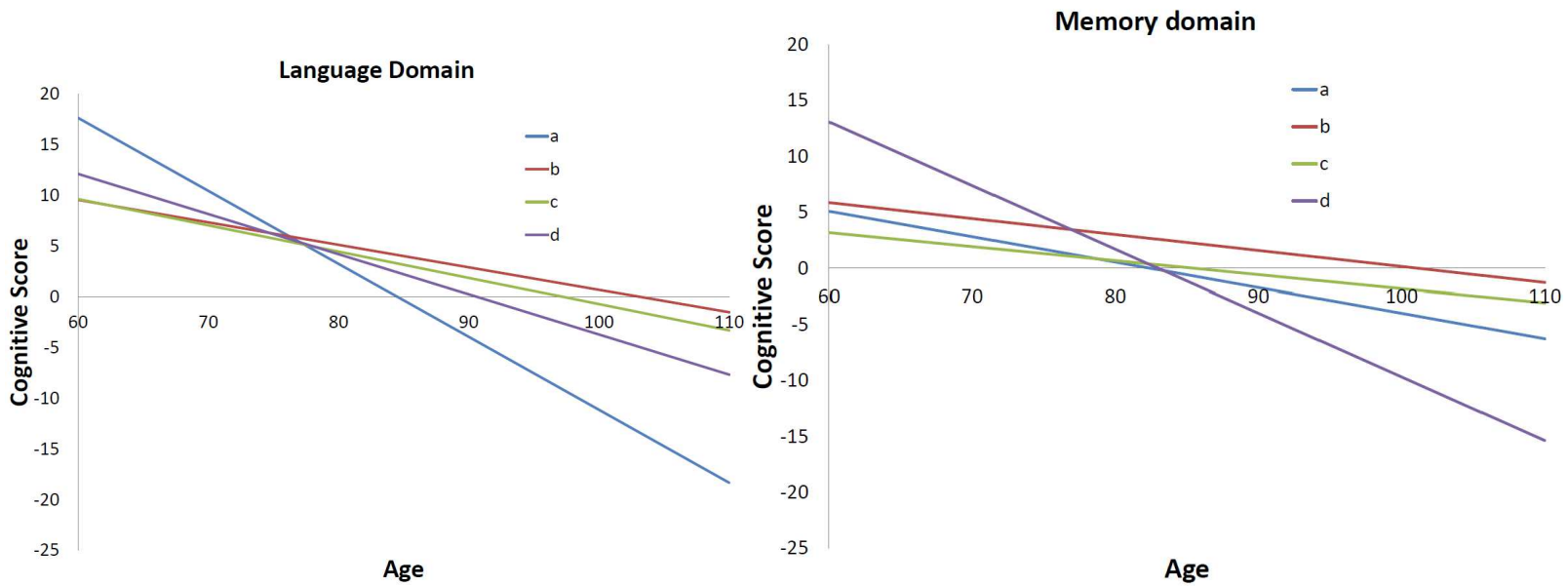

Figure 4.1: Estimated trajectories of language (left) and memory (right) decline for male dementia patients in four latent classes.

4.6 Conclusion and discussion

In this work, we performed simulation studies to survey the effects of adding labelled data to our finite mixture latent trajectory model under different considerations and re-analyzed UDS data by incorporating pathological information. We found that adding as little as 10% labelled data improves the classification accuracy and efficiency, especially when data were not well separated.

In practice, usually we don't have prior information about whether data are well separated or not, thus, labelled information should always be used. However, we found that there was just a little improvement of performance when proportion of labelled data changed from 10% to 20%, or when number of overall subjects was big. Therefore, although it is always ideal to get as more labelled data as possible, if that process is very expensive and/or resource is limited, recruiting more unlabeled data to increase the overall number of subjects can also improve the performance.

In our application of UDS data, we found when labelled data was not representative of underlying classes, the group assignments almost didn't change for those unlabeled subjects (e.g. latent classes a and d). On the other hand, when labelled data was representative of some underlying classes, the newly defined classes became more homogenous (e.g. latent classes b and c). Due to the lack of background information and small sample sizes, the 4 classes we defined from data with pathological information was arbitrary and they were not well matched to the latent classes we previously identified. UDS data collection is still ongoing and more pathological data will be available in the near future. We will re-analyze UDS data when we have better understanding of disease subtypes and their cognitive decline patterns.

Our study is aimed at analyzing UDS data and in order to fit the general purpose, there are several extensions we need to consider. First, we assume those labelled data were chosen randomly. This is not a problem in our dementia study. However, it is not always true, for example, if one disease is much harder to diagnoses than other diseases, we will have disproportionally smaller number of patents with that particular disease. Second, in some cases the proportion of each latent class is already known, for example, in Hosmer's study, the proportion of male fish was already known [37], and that information can also be included to improve the performance. Third, in current studies, we assume that there were no unknown or unobserved latent classes, i.e. unlabeled data has to belong to one of the latent classes observed in labelled data, which is a very strong assumption. Ideally, a series of model, with numbers of latent classes equal to or larger than the number of classes observed should be fitted, then using appropriate criteria to identify correct number of latent classes that are most biologically meaningful.

CHAPTER 5. CONCLUSION AND DISCUSSION

In this dissertation, we first proposed a multivariate finite mixture latent trajectory model motivated by data that are often encountered in dementia studies. This model has the capability to identify latent subpopulations in multiple cognitive domains with each domain measured by multiple neuropsychological tests. Simulation results showed adequate performance of this model. By applying this model to the UDS data set, four distinct cognitive decline patterns were identified. In the second part of this dissertation, through a simulation study, we investigated the performances of several commonly used IC-based fit indices including AIC, BIC, as well as CAIC, sample adjusted CAIC, sample adjusted BIC, ICLBIC, and sample adjusted ICLBIC for selecting the number of latent classes using multivariate finite mixture latent trajectory models. The level of separation between the latent classes had the greatest impact on the performance of these indices. The sample adjusted BIC performed uniformly better in all situations and is therefore the preferred approach for multivariate finite mixture trajectory models. In the third part of this dissertation, the multivariate finite mixture latent trajectory model was further extended to incorporate labeled class information. Our results showed that even small amount of labeled data can improve classification accuracy and estimation efficiency, especially for not well separated data. This model was applied to the same UDS data set and compared to previous analysis based on the unlabeled data. Results showed that subjects classified in the same class based on the partially labelled data have more similar trajectories with each other.

In dementia and aging studies, multiple cognitive measures from several cognitive domains are routinely collected. Neuropsychological tests within the same domain can be

considered as measures of the same underlying latent construct. Joint modeling of test results from multiple domains can improve our ability to identify unique patterns of cognitive decline. In addition, by adding labeled information, the classes identified can be linked directly to the known classes, hence making results more biological meaningful. For complex latent mixture models, the total number of parameters is usually large, thus it is important to apply appropriate amount of penalty when using IC based criteria for model selection. When a new class is added, all parameters associated with that class are also added leading to a large penalty from number of parameters. Our simulation study on comparing information based criteria suggests that SABIC, which uses adjusted sample size, reduced the overall penalty and outperformed commonly used AIC and BIC in all scenarios.

Currently the proposed models assumed the normal distribution due to its tractability and ease of implementation. For further research, non-normal distributions and/or mixed distribution need to be included. For many neuropsychological test scores, there are floor or ceiling effects and approaches to model those data also need to be considered. Additional extension to our current work is when the proportions of each class in the population are known and the models will be extended to consider these additional constrains.

This work was motivated by studies of cognitive decline among dementia patients and the ultimate goal is to find patients that have similar cognitive decline pattern and therefore possibly share the same disease etiology. In our application of UDS data, 4 distinct cognitive decline patterns were found. However, these 4 classes cannot be directly linked to known dementia subtypes, even with the help of pathological

information. One reason for that is, although there are only a few dementia subtypes, many patients have more than one type of dementia. Therefore, the combination of different kinds of dementia creates many classes, some of them with sample sizes too small to model. Since the UDS data collection is still ongoing and more pathological data will be available in the future, a further extension of the current work would be the capability of modeling mixed types of dementia.

BIBLIOGRAPHY

1. Hayden, K.M., et al., *Cognitive decline in the elderly: an analysis of population heterogeneity.* Age Ageing, 2011. **40**(6): p. 684-9.
2. Galton, C.J., et al., *Atypical and typical presentations of Alzheimer's disease: a clinical, neuropsychological, neuroimaging and pathological study of 13 cases.* Brain, 2000. **123 Pt 3**: p. 484-98.
3. Martin, A., et al., *Towards a behavioral typology of Alzheimer's patients.* J Clin Exp Neuropsychol, 1986. **8**(5): p. 594-610.
4. Neary, D., et al., *Neuropsychological syndromes in presenile dementia due to cerebral atrophy.* J Neurol Neurosurg Psychiatry 1986. **49**(2): p. 163-174.
5. Miller, B.L., et al., *Frontal lobe degeneration: clinical, neuropsychological, and SPECT characteristics.* Neurology, 1991. **41**(9): p. 1374-82.
6. Neary, D., J. Snowden, and D. Mann, *Frontotemporal dementia.* Lancet Neurol, 2005. **4**(11): p. 771-80.
7. Neary, D., et al., *Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria.* Neurology, 1998. **51**(6): p. 1546-54.
8. Neary, D., et al., *Dementia of frontal lobe type.* J Neurol Neurosurg Psychiatry, 1988. **51**(3): p. 353-61.
9. Perry, R.J. and J.R. Hodges, *Differentiating frontal and temporal variant frontotemporal dementia from Alzheimer's disease.* Neurology, 2000. **54**(12): p. 2277-84.
10. Nagin, D.S., *Analyzing developmental trajectories: A semiparametric, group-based approach.* Psychol Methods, 1999. **4**: p. 139-57.
11. Nagin, D.S., *Group-based modeling of development*. 2005, Cambridge, Mass: Harvard University Press.
12. Nagin, D.S. and C.L. Odgers, *Group-based trajectory modeling in clinical research.* Annu Rev Clin Psychol, 2010. **6**: p. 109-38.
13. Muthen, B., *Beyond SEM: General latent variable modeling.* Behaviormetrika, 2002. **29**: p. 81–117.
14. Muthen, B., *Latent variable analysis: growth mixture modeling and related techniques for longitudinal data*. The Sage Handbook of Quantitative Methodology for the Social Sciences, ed. D. Kaplan. 2004, Newbury Park, CA. 345–68.
15. Muthen, B. and K. Shedden, *Finite mixture modeling with mixture outcomes using the EM algorithm.* Biometrics, 1999. **55**: p. 463–469.
16. Proust, C., et al., *A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data.* Biometrics, 2006. **62**(4): p. 1014–1024.
17. Proust-Lima, C., L. Letenneur, and H. Jacqmin-Gadda, *A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome.* Statistics in Medicine, 2007. **26**: p. 2229–2245.
18. Kim, S.Y., *Determining the Number of Latent Classes in Single- and Multi-Phase Growth Mixture Models.* Struct Equ Modeling, 2014. **21**(2): p. 263-279.
19. McLachlan, G.J. and D. Peel, *Finite Mixture Models*. 2004, New York: Wiley-Intersci.

20. Nylund, K.L., T. Asparouhov, and B.O. Muthen, *Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study.* Struct. Equ. Model., 2007. **14**: p. 535-69.
21. Tofighi, D. and C.K. Enders, *Identifying the correct number of classes in growth mixture models.* Advances in Latent Variable Mixture Models, 2008: p. 317-341.
22. Yang, C.-C., *Evaluating latent class analysis models in qualitative phenotype identification.* Computational Statistics & Data Analysis, 2006. **50**(4): p. 1090-1104.
23. Lo, Y.T., N.R. Mendell, and D.B. Rubin, *Testing the number of components in a normal mixture.* Biometrika, 2001. **88**(767-78).
24. Akaike, H., *New look at statistical-model identification.* IEEE Trans. Automatic Control, 1974. **19**: p. 716-23.
25. Schwartz, G., *Estimating the dimension of a model.* Ann. Stat., 1978(6): p. 461-64.
26. Baudry, J., et al., *Combining Mixture Components for Clustering.* J Comput Graph Stat., 2010 June 1. **9**(2): p. 332-353.
27. Bozdogan, H., *Model Selection and Akaike Information Criterion (Aic) - the General-Theory and Its Analytical Extensions.* Psychometrika, 1987. **52**(3): p. 345-370.
28. Sclove, S.L., *Application of Model-Selection Criteria to Some Problems in Multivariate-Analysis.* Psychometrika, 1987. **52**(3): p. 333-343.
29. Rissanen, J., *A universal prior for integers and estimation by minimum description length.* The Annals of Statistics, 1983. **11**(2): p. 416-431.
30. Rissanen, J., *Modeling by shortest data description.* Automatica, 1978. **14**: p. 465-471.
31. Rissanen, J., *Minimum-description-length principle*. Encyclopedia of Statistical Sciences. Vol. 5. 1985, New York: John Wiley & Sons.
32. Biernacki, C., G. Celeux, and G. G, *Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood*. Technical Report No. 3521. 1998, Rhone-Alpes: INRIA.
33. Nigam, K., et al., *Text Classification from Labeled and Unlabeled Documents using EM.* Machine Learning, 2000. **39**: p. 103-134.
34. Seeger, M., *Learning with labeled and unlabeled data (Technical Report)* 2001, Institute for Adaptive and Neural Computation, University of Edinburgh: Edinburgh, United Kingdom.
35. Zhu, X. and A.B. Goldberg, *Introduction to semi-supervised learning.* Synthesis lectures on artificial intelligence and machine learning, 2009. **3**(1): p. 1-130.
36. McLachlan, G.J., *Discriminant Analysis and Statistical Pattern Recognition*. 1992, New York: Wiley.
37. Hosmer, D.W., *A comparison of iterative maximum likelihood estimates for the parameters of a mixture of two normal distributions under three different types of sample.* Biometrics, 1973. **29**: p. 761-770.
38. Hosmer, D.W. and N.P. Dick, *Information and mixtures of tow normal distributions.* Journal of Statistical Computation and Simulation, 1977. **6**: p. 137-148.

39. Salmon, D.P. and M.W. Bondi, *Neuropsychological assessment of dementia.* Annu Rev Psychol, 2009. **60**: p. 257-82.
40. *The National Alzheimer's Coordinating Center*. Available from: https://www.alz.washington.edu/.
41. McLachlan, G.J. and T. Krishnan, *The EM Algorithm and Extensions*. 1997, New York: John Wiley & Sons, Inc. .
42. Ng, S.K., et al., *A mixture model with random-effects components for clustering correlated gene-expression profiles.* Bioinformatics, 2006. **22**(14): p. 1745-52.
43. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, Series B, 1977. **39**(1): p. 1-38.
44. Muthen, B., et al., *General growth mixture modeling for randomized preventive interventions.* Biostatistics, 2002. **3**(4): p. 459-75.
45. Morris, J.C., et al., *The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers.* Alzheimer Dis Assoc Disord, 2006. **20**(4): p. 210-6.
46. Weintraub, S., et al., *The Alzheimer's Disease Centers' Uniform Data Set (UDS): the neuropsychologic test battery.* Alzheimer Dis Assoc Disord, 2009. **23**(2): p. 91-101.
47. *Alzheimer's Disease Genetics Consortium*. Available from: https://alois.med.upenn.edu/adgc/about/overview.html.
48. *The AlzGene database* Available from: http://www.alzgene.org/.
49. Corder, E.H., et al., *Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families.* Science, 1993. **261**(5123): p. 921-3.
50. Strittmatter, W.J., et al., *Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease.* Proc Natl Acad Sci U S A, 1993. **90**(5): p. 1977-81.
51. McKhann, G., et al., *Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease.* Neurology, 1984. **34**(7): p. 939-44.
52. McKhann, G.M., et al., *The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease.* Alzheimers Dement, 2011. **7**(3): p. 263-9.
53. Gray, S.M. and R. Brookmeyer, *Estimating a treatment effect from multidimensional longitudinal data.* Biometrics, 1998. **54**(3): p. 976-88.
54. Jacqmin-Gadda, H., C. Proust-Lima, and H. Amieva, *Semi-parametric latent process model for longitudinal ordinal data: Application to cognitive decline.* Statistics in Medicine, 2010. **29**(26): p. 2723–2731.
55. Biernacki, C. and G. Govaert, *Using the Classification Likelihood to Choose the Number of Clusters.* Computing Science and Statistics, 1997. **29**(3): p. 451-457.
56. Celeux, G. and G. Soromenho, *An entropy criterion for assessing the number of clusters in a mixture model.* Journal of Classification, 1996. **13**(2): p. 195-212.

57. Bezdek, J.C., *Pattern recognition with fuzzy objective function algorithms*. Advanced applications in pattern recognition. 1981, New York: Plenum Press. xv, 256 p.
58. Biernacki, C., G. Celeux, and G. Govaert, *An improvement of the NEC criterion for assessing the number of clusters in a mixture model.* Pattern Recognition Letters, 1999. **20**(3): p. 267-272.
59. Windham, M.P. and A. Cutler, *Information Ratios for Validating Mixture Analyses.* Journal of the American Statistical Association, 1992. **87**(420): p. 1188-1192.
60. Kullback, S. and R.A. Leibler, *On information and sufficiency.* Annals of Mathematical Statistics, 1951. **22**: p. 79-96.

CURRICULUM VITAE

Dongbing Lai

Education:

08/09-08/15: Ph. D. in Biostatistics, Department of Biostatistics, IUPUI.
08/07-05/09: M.S. in Applied Statistics, Department of Mathematical Sciences, IUPUI.
08/01-05/03: M.S. in Bioinformatics, School of Informatics, IUPUI.
09/89-07/93: B.S. in Biochemistry, Department of Biology, Nankai University.

Employment:

05/08-Present: Applied statistician II, Department of Medical and Molecular Genetics, IUPUI

05/05/-04/08: Applied statistician, Department of Medical and Molecular Genetics, IUPUI

07/03-04/05: Research data analyst, Department of Medical and Molecular Genetics, IUPUI

09/01-06/03: Research assistant, Department of Medical and Molecular Genetics, IUPUI

09/96-05/99: Research assistant, Department of Biochemistry & Molecular Biology, Peking Union Medical College & Chinese Academy of Medical Science.

09/93-08/96: Research technician, Center of Biochemical Immune Preparation, China Institute for Radiation Protection