Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Kari R. Hart Date

Robust Latent Class Analysis for Longitudinal Data

By

Kari R. Hart

Doctor of Philosophy

Biostatistics

John J. Hanfelt, Ph.D.
Advisor

Felicia Goldstein, Ph.D.
Committee Member

Robert H. Lyles, Ph.D.
Committee Member

Tianwei Yu, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Robust Latent Class Analysis for Longitudinal Data

By

Kari R. Hart
B.S., Lafayette College, 2006
M.S., Emory University, 2010

Advisor: John J. Hanfelt, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Abstract

Robust Latent Class Analysis for Longitudinal Data

By

Kari R. Hart

Latent class analysis is a likelihood-based approach that is designed to elucidate the structure underlying population heterogeneity. More specifically, in latent class analysis, researchers study the patterns of interrelationship among a set of observed feature variables in order to understand and characterize underlying population subtypes or classes. While, typically, these underlying classes cannot be observed directly, they often have meaningful physical interpretations. As such, latent class analysis is useful in many health applications, where it is a powerful statistical tool for detecting disease subtypes and diagnostic subcategories.

Existing latent class methods do not offer a robust and efficient approach applicable to longitudinal data. Most existing methods for latent class analysis apply only to cross-sectional data, while likelihood-based extensions for longitudinal data tend to be computationally intensive and sensitive to modeling assumptions. Thus, we propose a novel robust artificial-likelihood-based approach to longitudinal latent class analysis. In particular, we consider a finite mixture of latent-class-specific generalized estimating equations in which the class mixing proportions can be influenced by a set of covariates. The proposed model is fit under the assumption that the number of latent classes is fixed and known. However, since the number of classes is typically not known a priori, we explore novel model diagnostics for assessing the number of latent classes. The diagnostics rely on longitudinal extensions of information criteria, which account for how well the model fits the data, model complexity, and class membership uncertainty.

A major application of this research is in modeling latent trajectories based on the clinical presentation of diseases. In this research, we applied the proposed methods to a longitudinal data set from the National Alzheimer's Coordinating Center comprised of patients with a baseline consensus diagnosis of mild cognitive impairment (MCI). The proposed methods were used to statistically validate the presence of MCI subtypes and to model the progression of MCI within each subtype over time. Cognitive, functional, and neuropsychiatric assessments were considered as feature variables involved in the conceptualization of MCI subtypes, while an indicator of cerebrovascular disease was incorporated as a risk factor for MCI subtype membership.

Robust Latent Class Analysis for Longitudinal Data

By

Kari R. Hart
B.S., Lafayette College, 2006
M.S., Emory University, 2010

Advisor: John J. Hanfelt, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Acknowledgements

Thank you to the faculty, staff, and students in the Department of Biostatistics and Bioinformatics for providing me with me a caring and welcoming environment during my time at Emory. In particular, I am extremely grateful to my advisor, John Hanfelt, for his consistent patience, guidance, and support. He has been an invaluable mentor and teacher throughout the completion of my degree, and has generously offered his advice in areas beyond the realm of this research. I truly could not have asked for a better advisor. In addition, I would like to acknowledge my committee members, Robert Lyles, Tianwei Yu, and Felicia Goldstein, for their thoughtful comments and constructive feedback. Many thanks also to Lance Waller for his course and career advice throughout my graduate studies and to Tracy Wachholz for always being there to listen and encourage me.

I would like to take this opportunity to thank all of the family and friends who have helped me along the way. In particular, thank you to Laura Ward and Julia Cleveland for being wonderful friends and for helping me to clear my head whenever my research took an unexpected twist. Thank you to Sameera Wijayawardana for his friendship, advice, and belief in me. Thank you to my parents, Daniel and Dahni Barkley, and my brother, Jeremiah Barkley, for their love and support throughout all of the highs and lows of my graduate career. Thank you to my in-laws, John and Cherie Hart, for their constant encouragement. Last, but not least, thank you to my husband, Rory Hart, for offering me unfailing love and support and for helping me to keep everything in perspective.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Latent class analysis is a statistical method used to identify population subtypes and to classify related subjects into their most likely subtype. As such, latent class analysis is useful in many health applications, where it is a powerful statistical tool for detecting disease subtypes or diagnostic subcategories. More precisely, latent class analysis is a likelihood-based approach designed to elucidate the structure underlying the heterogeneity exhibited by individuals in a certain population of interest. In latent class analysis, researchers observe a set of clinically-relevant feature variables, which they believe are associated with a set of underlying "classes" in the population of interest. These classes represent mutually exclusive and exhaustive subpopulations. The idea behind latent class analysis is then to study the patterns of interrelationship among the observed feature variables in order to better understand and characterize the underlying population subtypes. Note that, within a given latent class, the observed variables are assumed to be independent. This is known as the "local independence" assumption.

1.2 Motivating Example

Many phenomena in the biological, social, and physical sciences cannot be plainly viewed. Rather, only symptoms or indicators of the phenomena can be observed. For example, consider mild cognitive impairment (MCI). MCI refers to the clinical state in which a subject is cognitively impaired, usually in the memory domain, but is not suffering from dementia [41]. Although neuropsychological testing is often used to differentiate elderly individuals with MCI from those who experience normal aging, MCI is not a neuropsychological diagnosis, and no specific test or battery of tests currently exist to confirm a diagnosis of MCI. Determining that a patient has MCI is further complicated because not all patients present with an identical set of symptoms. Indeed, research has suggested tremendous heterogeneity in the clinical presentation of MCI. As a result, MCI is frequently classified into four subtypes: Amnestic MCI, Multidomain MCI-Amnestic, Multidomain MCI-Non-Amnestic, or Single Nonmemory MCI [14]. These subtypes were determined based on clinical observation rather than on a rigorous clustering approach. Further, MCI patients are typically classified based on a single clinical assessment, which ignores potential variation in the progression of symptoms over time. Thus, the motivation for this dissertation research is to empirically validate the presence of MCI subtypes, to incorporate longitudinal data into subtype classification, and to model the progression of MCI within each subtype over time.

In order to conceptualize MCI subtypes, longitudinal data on a large sample of patients with a consensus diagnosis of MCI at baseline was obtained from the National Alzheimer's Coordinating Center (NACC). The data included cognitive, functional, and behavioral assessments from the NACC Uniform Data Set (UDS)[2, 25]. Functioning was assessed using the Functional Assessment Questionnaire (FAQ)[23] as reported by an informant. Behavioral disturbances were assessed using the Geriatric Depression Scale (GDS)[73] and select items from the Neuropsychiatric Questionnaire

(NPI-Q)[15]. Cognitive performance was assessed using the following ten neuropsychological items: mini-mental state exam (MMSE)[32], Trail-Making Test[83], Boston Naming Test[40], Category Fluency[75], Digit Span subtest, Digit Symbol subtest[84], Logical Memory, and Story A[85]. In addition to these assessments, the Rosen Modification of the Hachinski Ischemic Score (RMHIS) [27] was used as an indicator of cerebrovascular disease (CVD). Although the RMHIS score was not thought to contribute to the conceptualization of MCI subtypes, it was considered as a potential risk factor for belonging to a particular MCI subtype.

1.3 Limitations of Existing Methodology

Longitudinal studies involve the repeated measurement of subjects over time. They are considered in contrast to cross-sectional studies in which each subject is observed only once. One of the main advantages of a longitudinal study is its ability to assess change. In particular, longitudinal studies can distinguish changes within individuals over time from differences between subjects at their baseline or initial starting values. Analysis of longitudinal data requires specific methodology because repeated measurements on the same subject tend to be correlated. As a result, the assumption of independent observations, which underlies most standard cross-sectional methodology, is no longer satisfied.

Both discrete and continuous longitudinal data can be modeled using extensions of generalized linear models (GLMs), which accommodate correlated observations. Specifically, three distinct model formulations are typically considered: marginal, random effects, and transition models [12]. Estimation for both random effects and transition models is based on maximum likelihood methods. In contrast, marginal models require specification of only the first two moments and can be estimated using generalized estimating equations (GEEs) [46, 87, 63].

Presently, latent class methods primarily deal with cross-sectional data; however, extending latent class methodology to longitudinal data and latent class trajectories may expound the underlying subpopulation structure. Both mixed models and transition models for longitudinal latent class analysis have been explored and implemented in the literature (see for example [58, 65, 16, 7, 6]). Although such models aid in understanding population heterogeneity, these fully-parametric approaches tend to be computationally complex and often require strict modeling assumptions, particularly when one or more of the feature variables are discrete [17]. As an alternative, Reboussin et al.(2002) [70] presented a latent transition approach for analyzing multiple longitudinal binary health outcomes with multiple-cause non-response when the data is missing at random and non-likelihood-based analysis is performed. While Reboussin et al.'s transition model overcomes some of the computational complexities associated with a full-likelihood-based approach, it still has several limitations. In particular, parameter estimation is based on unconditional moments and may be inefficient. Additionally, the proposed approach uses a first-order transition model for which every latent class must be present at every time point. In the context of slowly progressing illnesses- such as neurodegenerative diseases- this restriction may not be realistic because the conceptualization of the latent classes may vary with time. Semi-parametric approaches to longitudinal latent class analysis, which have the potential to overcome many of these limitations, have not yet been adequately examined in the literature.

1.4 Outline and Objectives

Current latent class methods tend to rely on fully parametric modeling approaches. The primary objective of this research is to develop a robust artificial-likelihood-based approach to latent class analysis for high-dimensional longitudinal data. Chapter 2

will provide a brief literature review of cross-sectional finite mixture models, artificial likelihood, and model selection diagnostics for longitudinal data. Chapter 3 focuses on establishing generalized estimating equation (GEE) methodology for modeling mixtures of longitudinal data under the assumption that the number of latent classes is fixed and known. The performance of the proposed methods will be explored via simulation studies. Then, since it is often not realistic to assume that the number of latent classes is known, Chapter 4 discusses model selection diagnostics for determining the appropriate number of latent trajectories in a heterogenous population. The proposed measures of model fit extend cross-sectional information criteria to finite mixtures of generalized estimating equations. Simulation studies will be performed to assess and compare the effectiveness of these model selection diagnostics in correctly identifying the number of latent classes. In Chapter 5, the proposed methodology will be applied to the longitudinal data from the NACC- UDS in order to identify and model subtypes of mild cognitive impairment. Finally, Chapter 6 will discuss potential areas of future research and extensions of the proposed methodology.

Chapter 2

Literature Review

2.1 Cross-Sectional Finite Mixture Models

2.1.1 Overview

Finite mixture models offer a way to model heterogeneity in a cluster analysis context and to accommodate situations in which a single parametric family is unable to satisfactorily model local variations in observed data. As described in McLachlan and Peel(2000) [54], let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ denote a random sample of size n where $\mathbf{Y}_i$ is a J-dimensional random vector with probability density function $f(\mathbf{y}_i)$ on $\Re^J$. For cross-sectional applications, $\mathbf{Y}_i$ is a vector of random variables corresponding to J measurements taken on the i^{th} subject. Note that, although this notation assumes that there are J measurements on each subject, the methodology described below can be naturally extended to accommodate unbalanced data. Then, let $\mathbf{Y} = \left(\mathbf{Y}_1^T, \ldots, \mathbf{Y}_n^T\right)^T$ represent the entire sample. It follows that the probability density for $\mathbf{Y}_i$ under a C-component mixture model is

$$f\left(\mathbf{y}_i; \boldsymbol{\psi}\right) = \sum_{g=1}^{C} \pi_g f_g(\mathbf{y}_i; \boldsymbol{\theta}_g),$$

where $f_g(\mathbf{y}_i; \boldsymbol{\theta_g})$ are component densities and π_g are mixing proportions or weights satisfying

$$0 \leq \pi_g \leq 1$$
$$\sum_{g=1}^{C} \pi_g = 1$$

for g = 1, . . . , C. The vector $\boldsymbol{\psi} = (\pi_1, \ldots, \pi_{C-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C)$ is the vector containing all unknown parameters. Assume that $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C$ are distinct and let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C)$.

In this framework, a C-component mixture model can be viewed as arising when $\mathbf{Y}_i$ is drawn from a population consisting of C subgroups, $G_1, \ldots, G_C$, in proportions $\pi_1, \ldots, \pi_C$. To clarify this interpretation of finite mixture models, let $\mathbf{Z}_i$ be a C-dimensional component-label vector, where the g^{th} element of $\mathbf{Z}_i$ is defined to be one if the origin of $\mathbf{Y}_i$ is the g^{th} mixture component and zero otherwise. In this framework, $\mathbf{Z}_i$ is distributed according to a multinomial distribution consisting of one draw on C categories with probabilities $\pi_1, \ldots, \pi_C$, i.e. $\mathbf{Z}_i \sim Mult_C(1, \boldsymbol{\pi})$. Further, by Bayes' Rule, the posterior probability that subject i belongs to the g^{th} component of the mixture model given observation $\mathbf{y}_i$ can be expressed as

$$\begin{aligned} \tau_{ig} &= \tau_{ig}(\mathbf{y}_i; \boldsymbol{\psi}) \\ &= Pr(Z_{ig} = 1 | \mathbf{y}_i) \\ &= \frac{\pi_g f_g(\mathbf{y}_i; \boldsymbol{\theta_g})}{\sum_{d=1}^{C} \pi_d f_d(\mathbf{y}_i; \boldsymbol{\theta_d})}, \end{aligned}$$

where $g = 1, \ldots, C$ and $i = 1, \ldots, n$

Note that the mixing proportions may also be modeled as functions of a $P \times 1$ vector of covariates, $\mathbf{x}_i$[19]. In this situation, the mixing proportions are subject-specific and can be modeled using a polytomous logistic regression model. Specifically,

for the i^{th} subject with observed feature variables $\mathbf{y}_i$ and covariates $\mathbf{x}_i$,

$$\begin{aligned}\pi_{ig} &= \pi_g\left(\mathbf{x}_i;\boldsymbol{\alpha}\right) \\ &= \frac{exp(\boldsymbol{\alpha}_g^T\mathbf{x}_i)}{1+\sum_{h=1}^{C-1}exp(\boldsymbol{\alpha}_h^T\mathbf{x}_i)} \quad g=1,\ldots,C,\end{aligned}$$

where $\boldsymbol{\alpha}_C = 0$ and $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^T,\ldots,\boldsymbol{\alpha}_{C-1}^T\right)^T$. Further, when the mixing proportions depend on covariates, the parameter vector of interest becomes $\boldsymbol{\psi} = \left(\boldsymbol{\alpha}^T,\boldsymbol{\theta}^T\right)^T$.

The log-likelihood with respect to $\boldsymbol{\psi} = (\boldsymbol{\alpha},\boldsymbol{\theta})$ can be expressed as

$$l(\boldsymbol{\psi}) = \sum_{i=1}^{n} log\left\{\sum_{g=1}^{C}\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})f_g\left(\mathbf{y}_i;\boldsymbol{\theta}_g\right)\right\}.$$

It follows that the score equation with respect to $\boldsymbol{\alpha}$ is given by

$$\begin{aligned}S(\boldsymbol{\alpha}) &= \frac{\partial l(\boldsymbol{\alpha},\boldsymbol{\theta})}{\partial\boldsymbol{\alpha}} \\ &= \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\alpha}}\left[log\left\{\sum_{g=1}^{C}\pi_g\left(\mathbf{x}_i;\boldsymbol{\alpha}\right)f_g(\mathbf{y}_i;\boldsymbol{\theta}_g)\right\}\right] \\ &= \sum_{i=1}^{n}\left[\frac{\frac{\partial}{\partial\boldsymbol{\alpha}}\left\{\sum_{g=1}^{C}\pi_g\left(\mathbf{x}_i;\boldsymbol{\alpha}\right)f_g(\mathbf{y}_i;\boldsymbol{\theta}_g)\right\}}{\sum_{d=1}^{C}\pi_d\left(\mathbf{x}_i;\boldsymbol{\alpha}\right)f_d(\mathbf{y}_i;\boldsymbol{\theta}_d)}\right] \\ &= \sum_{i=1}^{n}\left[\frac{\sum_{g=1}^{C}\frac{\partial}{\partial\boldsymbol{\alpha}}\left\{\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})\right\}f_g(\mathbf{y}_i;\boldsymbol{\theta}_g)}{\sum_{d=1}^{C}\pi_d\left(\mathbf{x}_i;\boldsymbol{\alpha}\right)f_d(\mathbf{y}_i;\boldsymbol{\theta}_d)}\right] \\ &= \sum_{i=1}^{n}\left[\frac{\sum_{g=1}^{C}\left\{\frac{\partial\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}}{\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})}\right\}\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})f_g(\mathbf{y}_i;\boldsymbol{\theta}_g)}{\sum_{d=1}^{C}\pi_d(\boldsymbol{x}_i;\boldsymbol{\alpha})f_d(\mathbf{y}_i;\boldsymbol{\theta}_d)}\right] \\ &= \sum_{i=1}^{n}\sum_{g=1}^{C}\tau_{ig}(\mathbf{y}_i;\boldsymbol{\psi})\frac{\partial log\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}.\end{aligned}$$

where, in the last equality, the posterior class membership probability is defined as

$$\begin{aligned}\tau_{ig}(\mathbf{y}_i;\boldsymbol{\psi}) &= Pr(z_{ig}=1|\mathbf{y}_i,\mathbf{x}_i)\\ &= \frac{\pi_g(\mathbf{x}_i,\boldsymbol{\alpha})f_g(\mathbf{y}_i;\boldsymbol{\theta}_g)}{\sum_{d=1}^{C}\pi_d(\mathbf{x}_i,\boldsymbol{\alpha})f_d(\mathbf{y}_i;\boldsymbol{\theta}_d)}.\end{aligned}$$

Bandeen-Roche(1997)[19] notes that the class-specific score equation for $\boldsymbol{\alpha}$ can be re-expressed in the form

$$\frac{\partial l}{\partial \alpha_{pg}} = \sum_{i=1}^{n} x_{ip}\left\{\tau_{ig}\left(\mathbf{y}_i;\boldsymbol{\psi}\right) - \pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})\right\}, g = 1,\ldots,C-1,$$

where x_{ip} refers to the p^{th} covariate for $p = 1,\ldots,P$ and $\pi_C\left(\mathbf{x}_i;\boldsymbol{\alpha}\right) = 1-\sum_{h=1}^{C-1}\pi_h\left(\mathbf{x}_i;\boldsymbol{\alpha}\right)$. Analogously, the score equation with respect to $\boldsymbol{\theta}$ can be expressed as

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{g=1}^{C}\tau_{ig}(\mathbf{y}_i;\boldsymbol{\psi})\frac{\partial log f_g(\mathbf{y}_i;\boldsymbol{\theta}_g)}{\partial\boldsymbol{\theta}}.$$

2.1.2 Maximum Likelihood Estimation

Estimation for finite mixture models is typically done via the Expectation-Maximization (EM) algorithm due to its easy implementation and stable convergence [54, 53]. In the EM framework, estimation of $\boldsymbol{\psi}$ for a finite mixture model can be approached as an incomplete-data problem. More specifically, the observed data vector, $\mathbf{y} = \left(\mathbf{y}_1^T,\ldots,\mathbf{y}_n^T\right)^T$, can be viewed as incomplete because the component-label vectors, $\mathbf{z} = \left(\mathbf{z}_1^T,\ldots,\mathbf{z}_n^T\right)^T$, are unknown. Thus, the complete-data vector is given by $\mathbf{y}_c = \left(\mathbf{y}^T,\mathbf{z}^T\right)^T$. It follows that the complete-data log-likelihood is given by:

$$l_C\left(\boldsymbol{\alpha},\boldsymbol{\theta}\right) = \sum_{i=1}^{n}\sum_{g=1}^{C} z_{ig}\left\{log\left[\pi_g\left(\mathbf{x}_i;\boldsymbol{\alpha}\right)\right] + log\left[f_g\left(\mathbf{y}_i;\boldsymbol{\theta}_g\right)\right]\right\}.$$

Treating the component-label vectors as missing, it is then possible to proceed iteratively between the expectation and maximization steps of the EM algorithm. In the expectation step, the conditional expectation of the complete log-likelihood given the observed data vector $\mathbf{y}$ is computed. Then, in the M step, the conditional expectation of the complete log-likelihood is maximized with respect to $\boldsymbol{\psi}$.

McLachlan and Peel[54] also outline a more direct and computationally appealing approach to the EM algorithm, which they refer to as the direct approach for applying the EM algorithm to finite mixtures models. The direct approach obtains an estimate of $\boldsymbol{\psi}$, denoted $\hat{\boldsymbol{\psi}}$, by iterating between

$$\tau_{ig}(\mathbf{y}_i;\boldsymbol{\psi}) = \frac{\pi_g\left(\mathbf{x}_i;\boldsymbol{\alpha}\right) f_g(\mathbf{y}_i;\boldsymbol{\theta}_g)}{\sum_{d=1}^{C}\pi_d\left(\mathbf{x}_i;\boldsymbol{\alpha}\right) f_d(\mathbf{y}_i;\boldsymbol{\theta}_d)}$$

and solving the score equations with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$, i.e.

$$\begin{aligned}\textstyle\sum_{g=1}^{C}\sum_{i=1}^{n}\tau_{ig}(\mathbf{y}_i;\boldsymbol{\psi})\frac{\partial log\pi_g(\mathbf{x}_i;\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} &= \mathbf{0}\\ \textstyle\sum_{g=1}^{C}\sum_{i=1}^{n}\tau_{ig}(\mathbf{y}_i;\boldsymbol{\psi})\frac{\partial log f_g(\mathbf{y}_i;\boldsymbol{\theta}_g)}{\partial\boldsymbol{\theta}} &= \mathbf{0}\end{aligned} \quad ,$$

until a pre-specified convergence criteria is met. Ideally, this iterative algorithm will converge to the global solution; however, it is possible that the algorithm may converge to a local solution instead. Local solutions can often lead to substantially different interpretations from those suggested by the global solution. Thus, steps need to be taken in order to help avoid local solutions. The most common approach to dealing with local solutions is to generate multiple random starting points for the iterative algorithm. If the starting values lead to multiple solutions, then the solution that maximizes the likelihood function is selected.

2.1.3 Model Identifiability and Boundary Solutions

When estimating $\boldsymbol{\psi}$ for a mixture distribution, model identifiability needs to be considered. In general, a parametric family of densities $f(\mathbf{y}_i;\boldsymbol{\psi})$ is identifiable if distinct values of the parameter $\boldsymbol{\psi}$ determine distinct members of the family of densities $\{f(\mathbf{y}_i;\boldsymbol{\psi}) : \boldsymbol{\psi} \in \boldsymbol{\Omega}\}$, where $\boldsymbol{\Omega}$ is the specified parameter space[54]. Based on this definition of identifiability, a mixture distribution would not be identifiable because $f(\mathbf{y}_i;\boldsymbol{\psi})$ is invariant under the g! permutations of the component labels in $\boldsymbol{\psi}$. Thus, the definition of identifiability is slightly modified in the context of mixture distributions. Specifically, finite mixture models are said to be identifiable for $\boldsymbol{\psi} \in \boldsymbol{\Omega}$ if $f(\mathbf{y}_i;\boldsymbol{\psi}) = f(\mathbf{y}_i;\boldsymbol{\psi}^*)$ if and only if $\boldsymbol{\psi} = \boldsymbol{\psi}^*$ up to a permutation of the component labels. In practice, a constraint is sometimes imposed on $\boldsymbol{\psi}$ that uniquely determines the component labels after estimation.

In the context of finite mixture models, nonidentifiability due to overfitting also needs to be considered [33]. To illustrate nonidentifiability due to overfitting, consider a finite mixture model where the true number of components is $C = 2$. This mixture model can also be written with $C = 3$ components if the third component has a weight of zero, i.e.

$$f(\mathbf{y}_i;\boldsymbol{\psi}) = \pi_1 f_1(\mathbf{y}_i;\boldsymbol{\theta}_1) + \pi_2 f_2(\mathbf{y}_i;\boldsymbol{\theta}_2) + 0 \times f_3(\mathbf{y}_i;\boldsymbol{\theta}_3),$$

or if two of the components are the same, i.e.

$$f(\mathbf{y}_i;\boldsymbol{\psi}) = \pi_1 f_1(\mathbf{y}_i;\boldsymbol{\theta}_1) + (\pi_2 - \pi_3) f_2(\mathbf{y}_i;\boldsymbol{\theta}_2) + \pi_3 f_3(\mathbf{y}_i;\boldsymbol{\theta}_2).$$

More generally, Crawford(1994) [8] notes that any mixture with $C-1$ components defines a nonidentifiable subset in the larger parameter space corresponding to mixtures with C components. Although nonidentifiability due to label switching can be easily

addressed, nonidentifiability due to overfitting can be more problematic. Specifically, overfitting can result in numerical difficulties because the matrix of second derivatives will be close to singular. As such, Crawford suggests that, as a practical matter, it is usually preferable to reduce the number of components in the mixture rather than to work in the full-dimensional space.

2.1.4 Bayesian Estimation of Finite Mixture Models

An alternate way to conceptualize and estimate a finite mixture model is to use a Bayesian approach. Assume that there exists a prior distribution for all unknown parameters in the mixture model, $p(\boldsymbol{\psi})$. Then, define the posterior density as

$$p(\boldsymbol{\psi}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\psi})\, p(\boldsymbol{\psi}).$$

Although the Bayesian approach tends to reduce the risk of obtaining spurious modes in cases where the EM algorithm leads to degenerate solutions[33], there are no natural conjugate priors available for the mixture likelihood function. As such, the posterior density $p(\boldsymbol{\psi}|\mathbf{y})$ does not belong to any standard distributional family. Thus, until the development of Markov Chain Monte Carlo (MCMC) methods, Bayesian estimation approaches for finite mixture models were infeasible.

As computational resources have increased, MCMC methods for finite mixture models have become more common; however, the Bayesian approach does present some unique challenges. For example, as described above, the likelihood function associated with a finite mixture model is invariant under a permutation of the component labels. Although maximum likelihood estimation via the EM algorithm is not affected by potential switching of component labels during different iterations, label switching can be problematic for Bayesian estimation, which relies on the simulation of realizations of $\boldsymbol{\psi}$ from posterior distributions. In addition, priors on the mixing

proportions have been used to draw the mixing proportions away from the boundary of the parameter space and to avoid the numerical issues that sometimes arise due to boundary solutions. This should be done with care, however, since it can eliminate the possibility of reducing the number of components when the model is actually overfit and informative priors tend to force too many distinct components[33]. A more comprehensive overview of the Bayesian approach to finite mixture models, the challenges it presents, and methods for overcoming some of these challenges can be found in Fruhwirth-Schnatter(2006)[33].

It should also be noted, that recently Bayesian approaches to finite mixture models for repeated measurements have begun to appear in the literature (see, for example, [22]). These approaches are often referred to as Bayesian growth mixture models. In growth mixture model approaches, the latent class variable is not directly identified by the feature variables. Instead, the latent class variable captures heterogeneity in the growth model parameters. In this context, it is important to note that the distribution assumed for the growth model parameters, i.e. the heterogeneity distribution, is influential and that misspecification of the distribution can lead to substantial changes in the parameter estimates [33].

2.1.5 Assessing the Number of Components: Information Criteria

When fitting a finite mixture model, the number of components, C, is typically assumed to be fixed and known. Unfortunately, in many applications, a priori information regarding the number of components, C, is not available. In latent class analysis, selecting the appropriate number of components relies on the fundamental assumption of local independence. The axiom of local independence for latent class models states that observed features are statistically independent within a given latent class. Thus, selecting the appropriate number of latent classes ideally leads to a model in

which the underlying classes fully account for the population heterogeneity.

Even in the cross-sectional context, there is currently no consensus regarding the best approach for selecting the number of components in a finite mixture model. With that said, there are two general approaches to address this issue. The first approach is based on information criteria, while the second involves hypothesis testing.

Model selection based on information criteria is motivated by the Kullback-Leibler (KL) information [43]. Intuitively, the KL information is a measure of the difference between the proposed statistical model and the true distribution of the observed data. Using the notation of McLachlan and Peel [54], assume that the true density of the observed data is $f(\boldsymbol{\omega})$ and denote the estimated model being considered by $f\left(\boldsymbol{\omega};\hat{\boldsymbol{\psi}}\right)$. The KL information of $f(\boldsymbol{\omega})$ with respect to $f\left(\boldsymbol{\omega};\hat{\boldsymbol{\psi}}\right)$ is then:

$$\begin{aligned} I\left\{f(\boldsymbol{\omega}); f\left(\boldsymbol{\omega};\hat{\boldsymbol{\psi}}\right)\right\} &= \int f(\boldsymbol{\omega}) log\left\{\frac{f(\boldsymbol{\omega})}{f\left(\boldsymbol{\omega};\hat{\boldsymbol{\psi}}\right)}\right\} d\boldsymbol{\omega} \\ &= \int f(\boldsymbol{\omega}) log f(\boldsymbol{\omega})\, d\boldsymbol{\omega} - \int f(\boldsymbol{\omega}) log f\left(\boldsymbol{\omega};\hat{\boldsymbol{\psi}}\right) d\boldsymbol{\omega} \geq 0. \end{aligned}$$

Since the first term above does not depend on the fitted model, estimation of the Kullback-Leibler information is based solely on the second term. Now,

$$\begin{aligned} \boldsymbol{\eta}(\mathbf{y}; F) &= \int f(\boldsymbol{\omega}) log f\left(\boldsymbol{\omega};\hat{\boldsymbol{\psi}}\right) d\boldsymbol{\omega} \\ &= \int log f\left(\boldsymbol{\omega};\hat{\boldsymbol{\psi}}\right) dF(\boldsymbol{\omega}), \end{aligned}$$

where F denotes the true cumulative distribution and $\mathbf{y} = \left(\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T\right)^T$ is the observed data. Replacing the distribution function, F, with the empirical distribution function, $\hat{F}_n$, yields the following estimate of $\boldsymbol{\eta}(\mathbf{y}; F)$:

$$\boldsymbol{\eta}\left(\mathbf{y};\hat{F}_n\right) = \frac{1}{n}\sum_{i=1}^{n} log f\left(\mathbf{y}_i;\hat{\boldsymbol{\psi}}\right) = \frac{1}{n} log L\left(\hat{\boldsymbol{\psi}}\right).$$

Since the empirical distribution function $\hat{F}_n$ is generally closer to the fitted distribution function $F_{\hat{\boldsymbol{\psi}}}$ than the true distribution F, this estimator typically overestimates the model fit. In order to account for this bias, an information criterion can be defined as

$$logL\left(\hat{\boldsymbol{\psi}}\right) - b(F),$$

where $b(F)$ denotes an appropriate estimate for the bias of $\boldsymbol{\eta}\left(\mathbf{y}; \hat{F}_n\right)$ as an estimator of the expected log density. Information criteria based on the KL information are more commonly expressed as twice the negative value of this difference. Namely, information criteria based on KL information are of the general form

$$Information\ Criterion = -2logL\left(\boldsymbol{\psi}\right) + 2p_n\left(\boldsymbol{\psi}\right),$$

where the first term measures the lack of fit for the proposed model and $p_n\left(\boldsymbol{\psi}\right)$ is a penalty term that measures model complexity. Model selection is then done by minimizing the information criterion or, equivalently, minimizing the KL information.

Estimation of the Kullback-Leibler information gives rise to commonly known information criteria such as Akaike's Information Criteria (AIC)[1] and Bayesian Information Criteria (BIC)[72]. In the context of finite mixture models, AIC selects the model that minimizes

$$-2logL(\hat{\boldsymbol{\psi}}) + 2d$$

and BIC selects the model that minimizes

$$-2logL(\hat{\boldsymbol{\psi}}) + dlogn,$$

where d is the total number of parameters in the mixture model and n represents the number of subjects. Unfortunately, previous work has suggested that both AIC and BIC tend to overestimate the correct number of components for a finite mixture model. For this reason, alternative information criteria have been proposed. Specifically, in brief simulation studies performed by McLachlan and Peel[54], the integrated classification likelihood (ICL), the large cluster size approximation ICL referred to as ICL-BIC, and the Laplace-Empirical criterion (LEC) most often selected the true number of components. Of these three criteria, ICL-BIC is the easiest criterion to apply and will be described in detail.

Consider the fuzzy classification matrix whose elements are given by τ_{ig}, where $i = 1, \ldots, n$ and $g = 1, \ldots, C$. The entropy of this matrix is defined to be

$$EN(\boldsymbol{\tau}) = -\sum_{g=1}^{C}\sum_{i=1}^{n} \tau_{ig} log\left(\tau_{ig}\right).$$

If the components of the mixture are well separated then $EN(\hat{\boldsymbol{\tau}})$ will be close to its minimum value of 0. In contrast, if the components are poorly separated then $EN(\hat{\boldsymbol{\tau}})$ will have a large value. Thus, the degree of separation between the fitted components determines the severity of the penalty term, with more severe penalties imposed for situations in which class membership is more ambiguous. For large cluster sizes, the ICL-BIC criterion selects the number of components for a finite mixture model by minimizing

$$-2logL(\hat{\boldsymbol{\psi}}) + 2EN(\hat{\boldsymbol{\tau}}) + dlogn,$$

where d denotes the number of unknown parameters in $\boldsymbol{\psi}$.

2.1.6 Assessing the Number of Components: Hypothesis Testing

The most common hypothesis test for assessing the number of components in a mixture model relies on the likelihood ratio test statistic [54]. Consider a hypothesis test of $H_0 : C = C_0$ versus $H_A : C = C_1$, where $C_1 > C_0$ and C_0 represents the true-order of a C-component mixture model. Further, let $\hat{\boldsymbol{\psi}}_0$ and $\hat{\boldsymbol{\psi}}_1$ represent the maximum likelihood estimates of $\boldsymbol{\psi}$ under the null and alternative hypothesis, respectively. By definition, the likelihood ratio test statistic is

$$T_{LR} = 2\left\{logL(\hat{\boldsymbol{\psi}}_1) - logL(\hat{\boldsymbol{\psi}}_0)\right\}.$$

Unfortunately, for a mixture model, the likelihood ratio test statistic does not follow its traditional chi-square asymptotic distribution under the null hypothesis. When the null hypothesis holds, the parameter vector is on the boundary of the parameter space and in a non-identifiable subspace. As a result, the regularity conditions required for T_{LR} to have a chi-square asymptotic distribution break down. Although some theoretical results exist for determining the distribution of the likelihood ratio test statistic under the null hypothesis for specific mixture models, McLachlan(1987)[52] proposed a more general bootstrapping approach for the likelihood ratio test.

A less popular alternative for hypothesis testing proposed by Liang and Rathouz(1999)[45] is the score test. In their paper, Liang and Rathouz consider testing a two-component mixture model against a one-component mixture model. They let $y_1, \ldots, y_n$ be independent observations with f_i^* denoted as the probability density function for the i^{th} observation of the form

$$f_i^*\left(y_i; \alpha, \theta\right) = \alpha f_i\left(y_i; \theta\right) + \left(1 - \alpha\right) f_i\left(y_i; \theta_0\right).$$

Here, f_i is the pdf, θ_0 is known to investigators, $\alpha \in [0,1]$, and the hypothesis of interest tests $H_0 : \alpha = 0$. Note that, under the null hypothesis, θ is meaningless and, thus, the standard asymptotic results required for likelihood ratio hypothesis testing are not applicable. Now, the score function for α evaluated at $\alpha = 0$ is

$$S(\theta) = \sum_{i=1}^{n} S_i(\theta) = \sum_{i=1}^{n} \left\{ \frac{f_i(y_i;\theta)}{f_i(y_i;\theta_0)} - 1 \right\}.$$

Based on this score equation, $E\left[S(\theta)\right] \geq 0$ with equality occurring only under the null hypothesis. As with the likelihood ratio test, the score test does not behave well under a naive estimate of the parameter vector; however, unlike the likelihood ratio test, the score test is able to handle this situation by using an estimate of the parameter vector, which is not based on maximum likelihood. In other words, in order to make $S(\theta)$ computable, θ must be replaced by an estimator $\widetilde{\theta}$, which is well-behaved under the null hypothesis. Then, let $\hat{\theta}_\lambda$ be the value of λ that maximizes $L(\lambda, \theta)$ with fixed θ and define

$$T_\lambda = S(\hat{\theta}_\lambda),$$

where $0 < \lambda < 1$ is the user-specified value plugged in for α. Under some regularity conditions on the f_i's, the statistic $T^*_\lambda = \lambda T_\lambda$ asymptotically follows either a chi-squared distribution under H_0 if θ_0 is an interior point of the parameter space or a mixture of chi-square distributions if θ_0 is on the boundary of the parameter space. Note that, although Liang and Rathouz[45] consider the score test specifically for a two-component mixture model, it can likely be extended to handle tests involving more than two components.

2.2 Artificial Likelihood

2.2.1 Overview

Longitudinal studies are characterized by repeated measurements on individuals over time. The advantage of longitudinal studies over cross-sectional studies is that longitudinal studies can distinguish changes over time within individuals from differences among people in their baseline levels [12]. Since observations on a single subject over time tend to be correlated with one another, special statistical methods are required to analyze longitudinal data. In the context of generalized linear models (GLMs), there are three possible extensions for longitudinal data: marginal, random effects, and transition models. Estimation for both the random effects and transitional extensions of GLMs are based on traditional maximum likelihood methods. In contrast, the marginal model specifies only the first two moments and, as such, it does not generally have a well-defined likelihood function. In the absence of a fully-specified distribution for the observations, a reasonable approach to estimation is to use generalized estimating equations(GEEs), which are essentially a multivariate analogue of quasi-likelihood [46, 87, 63]. Quasi-likelihood, extended quasi-likelihood, and GEEs are each briefly described in this section.

As Hanfelt and Liang(1995) [37] point out, estimating functions sometimes have limited utility due to multiple roots for the estimating function, a poorly behaved Wald test, or lack of a goodness-of-fit test. In order to address these limitations, several artificial likelihood approaches for approximating the likelihood ratio in the absence of a well-defined likelihood function have been proposed[44, 68, 80]. Here, focus will be placed on the projection-based approach of Li(1993) [44] and the empirical likelihood based approach of Qin and Lawless(1994) [68]. Quadratic inference functions, as described in Qu(2000) [69], will also be discussed.

2.2.2 Quasi-likelihood and Extended Quasi-likelihood

Briefly consider the situation where there are n independent observations $(y_i, x_i; i = 1, \ldots n)$, where y_i is the i^{th} response variable with mean μ_i and variance $V(\mu_i)$, and x_i is an associated vector of covariates. Wedderburn(1974) [86] defined the quasi-likelihood for a single observation y_i, $Q(y_i; \mu)$, by the relation

$$\frac{\partial Q(y_i; \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)}$$

or, equivalently,

$$Q(y_i; \mu_i) = \int^{\mu_i} \frac{y_i - \mu}{V(\mu)} du$$

plus a function of y_i only. The quasi-likelihood for the sample, $Q(\mathbf{y}; \boldsymbol{\mu})$ is then defined to be the sum of the individual quasi-likelihoods. As shown in Table 2.1, the quasi-likelihood function often takes a simple closed-form for independent observations.

Table 2.1: Quasi-likelihood for a single observation y_i associated with some simple variance functions

Distribution	Variance Function	Quasi-likelihood $Q(y_i; \mu_i)$
Normal	1	$-\frac{(y_i-\mu_i)^2}{2}$
Poisson	μ_i	$y_i log(\mu_i) - \mu_i$
Binary	μ_i^2	$y_i log\left(\frac{\mu_i}{1-\mu_i}\right) + log(1-\mu_i)$

Note that, while defining a likelihood function requires that the full form of the distribution of the observations be specified, a quasi-likelihood function requires only that the relation between the mean and the variance of the observations be specified. Further, Wedderburn[86] and McCullagh(1983)[51] showed that the quasi-likelihood has many properties that are analogous to those of log-likelihood functions. In particular, the maximum quasi-likelihood estimate, $\hat{\boldsymbol{\beta}}_{QL}$, follows an asymptotic normal

distribution with mean $\boldsymbol{\beta}_{QL}$ and asymptotic covariances that can be computed in the standard fashion from the second derivative matrix of $Q(\mathbf{y};\boldsymbol{\mu})$.

In the formulation of the quasi-likelihood above, Wedderburn [86] relaxes the assumption of a known variance function of $\mathbf{y}$ by permitting a constant of proportionality or dispersion parameter, ϕ. When the response variable is binary, ϕ is assumed to be fixed at 1; however, for other response types, the dispersion parameter is unknown and must be estimated. Wedderburn [86] recommends estimating the scale parameter using

$$\tilde{\phi} = \frac{\chi^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

Estimation of the dispersion parameter based on the 'bias corrected' mean χ^2 statistic above is implemented in most standard statistical software packages. Other approaches for estimating the dispersion parameter are summarized in Wang and Hin(2010) [82].

In order to compare different variance functions on the same data, Nelder and Pregibon(1987) [59] proposed the extended quasi-likelihood function, which includes a covariance penalty. The extended quasi-likelihood function for a single observation y_i with mean μ_i and variance $\phi V(\mu_i)$ is defined to be

$$Q^{+}(y_i;\mu_i) = -\frac{1}{2}log\left\{2\pi\phi V(y_i)\right\} - \frac{\frac{1}{2}D(y_i;\mu_i)}{\phi},$$

where $D(y_i;\mu_i)$ denotes the deviance as defined by

$$D(y_i;\mu_i) = -2\left\{Q(y_i;\mu_i) - Q(y_i;y_i)\right\}.$$

As was the case with quasi-likelihood, the extended quasi-likelihood and deviance of the sample are simply the sum of the individual extended quasi-likelihoods and

deviances, respectively. It follows that the extended quasi-likelihood of the sample is

$$Q^{+}\left(\mathbf{y};\boldsymbol{\mu}\right)=Q\left(\mathbf{y};\boldsymbol{\mu}\right)-\frac{1}{2}\sum_{i=1}^{n}log\left\{2\pi\phi V(\mathbf{y})\right\}.$$

Note that the extended quasi-likelihood does not require a full distributional assumption. Rather, like the quasi-likelihood, it requires only that the first two moments be specified. Further, the extended quasi-likelihood is the unnormalized saddle point approximation for exponential families.

2.2.3 Generalized Estimating Equations (GEEs)

Consider the longitudinal observations $(y_{ij},\mathbf{x}_{ij})$ for times $t_{ij}, j=1,\ldots,m_i$ and subjects $i=1,\ldots,n$. Here, y_{ij} denotes the outcome variable and $\mathbf{x}_{ij}$ denotes a $p\times 1$ vector of covariates. Let $\mathbf{y}_i=(y_{i1},\ldots,y_{im_i})'$ be an $m_i\times 1$ vector of outcomes for subject i and $\mathbf{x}_i=(\mathbf{x}_{i1},\ldots,\mathbf{x}_{im_i})'$. Further, let $\boldsymbol{\beta}$ be a $p\times 1$ vector of regression parameters and ϕ denote a scale parameter. In this framework, the GEE approach models the mean and covariance matrix of $\mathbf{Y}_i, i=1,\ldots,n$ as

$$E\left(\mathbf{Y}_i\right)=\boldsymbol{\mu}_i\left(\boldsymbol{\beta}\right)$$

and

$$Var\left(\mathbf{Y}_i\right)=\mathbf{V}_i\left(\boldsymbol{\mu}_i,\boldsymbol{\alpha},\phi\right)=\phi\mathbf{A}_i^{1/2}\left(\boldsymbol{\mu}_i\right)\mathbf{R}_i\left(\boldsymbol{\alpha}\right)\mathbf{A}_i^{1/2}\left(\boldsymbol{\mu}_i\right),$$

where $\mathbf{A}_i$ is a diagonal matrix with $var\left(Y_{ij}\right)=v\left(\boldsymbol{\mu}_{ij}\right)$ as the j^{th} diagonal element and $\mathbf{R}_i\left(\boldsymbol{\alpha}\right)$ is the $n_i\times n_i$ working correlation matrix for each $\mathbf{Y}_i$. It is assumed that $\mathbf{R}_i\left(\boldsymbol{\alpha}\right)$ is completely specified by an $s\times 1$ vector of unknown parameters, $\boldsymbol{\alpha}$, which is the same for all subjects and can be estimated using the method of moments or

another set of estimating equations. It follows that the GEE model is specified by

$$g\left(\boldsymbol{\beta};\phi,\boldsymbol{\mu}_i\right)=\sum_{i=1}^{n}\frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}\mathbf{V}_i^{-1}\left(\boldsymbol{\mu}_i,\boldsymbol{\alpha},\phi\right)\left(\mathbf{Y}_i-\boldsymbol{\mu}_i\left(\boldsymbol{\beta}\right)\right).$$

The above model is sometimes referred to as a GEE1 model to distinguish it from the more recently developed GEE2 approach[47, 64, 88]. GEE1 models focus on estimation of the regression parameters and provide consistent, but not necessarily fully efficient, estimators regardless of whether the working correlation matrix is correctly specified [46]. Additionally, in contrast to GEE2 models, GEE1 models assume orthogonality of the estimating equations for the regression and association parameters [38].

2.2.4 Projection-Based Approach

The projection-based method for approximating the likelihood ratio described by Li [44] depends solely on the first two moments of the data. Thus, this approach does not require any additional knowledge about the moment structure of the feature vector than would be required to construct the estimating functions. In past work, McLeish and Small(1992)[55] worked directly with the likelihood ratio and projected it onto the subspace of L^2 spanned by $\prod_{i=1}^{n} X_i$, where the $X_i, i = 1, \ldots, n$ represent independent observations. In contrast, Li considers projecting the log likelihood ratio onto a subspace linear in the observations $X_1, \ldots, X_n$. Unlike in the work of McLeish and Small, Li's approach applies even when the $X_i, i = 1, \ldots, n$ are dependent observations; thus, it can be used for longitudinal studies. Unfortunately, the log likelihood ratio is not amenable to projection when only the first two data moments are known. Thus, rather than projecting the log likelihood ratio directly, Li obtains a linear approximation of the log likelihood ratio via a Taylor series expansion before conducting the projection.

Briefly, assume that $\mathbf{Y} = (Y_1, \ldots, Y_b)^T$ is a $b \times 1$ vector of possibly dependent observations with distribution p_ω. Further, suppose only the mean and covariance matrix of $\mathbf{Y}$ are known. Specifically, let $\boldsymbol{\mu}_\omega$ be the mean vector of $\mathbf{Y}$ and $\mathbf{V}_\omega$ be the variance-covariance matrix of $\mathbf{Y}$ under ω. Now, for simplicity of notation, let $a = p_\omega(\mathbf{Y})$ and $b = p_\nu(\mathbf{Y})$ and note that

$$\begin{aligned} 2log\left[\frac{a}{b}\right] &= log\left[\frac{a/b}{b/a}\right] \\ &= log\left[\frac{a}{b}\right] - log\left[\frac{b}{a}\right] \\ &= log\left[\frac{b+a-b}{b}\right] - log\left[\frac{a+b-a}{a}\right] \\ &= log\left[1+\frac{a-b}{b}\right] - log\left[1+\frac{b-a}{a}\right]. \end{aligned}$$

Then, a Taylor series expansion yields the following linear approximation of the log likelihood ratio

$$\begin{aligned} 2log\left[\frac{a}{b}\right] &\approx \frac{a-b}{b} + \frac{a-b}{a} + \frac{(b-a)^3(b+a)}{2a^2b^2} \\ &\approx \frac{a-b}{b} + \frac{a-b}{a}. \end{aligned}$$

Thus, by substituting back in $a = p_\omega(\mathbf{Y})$ and $b = p_\nu(\mathbf{Y})$, the following approximation of $log\frac{p_\omega(\mathbf{Y})}{p_\nu(\mathbf{Y})}$ can be obtained using Li's approach:

$$log\left[\frac{p_\omega(\mathbf{Y})}{p_\nu(\mathbf{Y})}\right] \approx \frac{p_\omega(\mathbf{Y}) - p_\nu(\mathbf{Y})}{2p_\nu(\mathbf{Y})} + \frac{p_\omega(\mathbf{Y}) - p_\nu(\mathbf{Y})}{2p_\omega(\mathbf{Y})}.$$

Next, the approximation of the log likelihood ratio is projected onto a suitable Hilbert subspace. Let $R_1 = \frac{p_\omega(\mathbf{Y}) - p_\nu(\mathbf{Y})}{2p_\nu(\mathbf{Y})} = \frac{1}{2}\left\{\frac{p_\omega(\mathbf{Y})}{p_\nu(\mathbf{Y})} - 1\right\}$. Consider the Hilbert space L^2 and the closed subspace of L^2 defined as $L_\nu = span\left\{Y_1 - \mu_{1\nu}, \ldots, Y_b - \mu_{b\nu}\right\}$ with inner product $\langle g_1, g_2 \rangle_\nu = E_\nu\left(g_1 g_2\right)$. Since R_1 is a member of L^2 it can be projected onto L_ν. Also, all elements of L_ν are linear in $\mathbf{Y}$ and, thus, take the form $\mathbf{a}^T\left(\mathbf{Y} - \boldsymbol{\mu}_\nu\right)$

for some vector $\mathbf{a}$. Denote the projection of R_1 onto L_ν by $\hat{R}_1$. It can be shown that $\hat{R}_1 = \frac{1}{2}\left\{(\boldsymbol{\mu}_\omega - \boldsymbol{\mu}_\nu)^T \mathbf{V}_\nu^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\nu)\right\}$ is the unique projection of R_1 onto L_ν by showing that the residual $R_1 - \hat{R}_1$ is orthogonal to all $h \in \mathbf{a}^T(\mathbf{Y} - \boldsymbol{\mu}_\nu)$. In other words, the proposed form of $\hat{R}_1$ is the correct form of the projection if $\left\langle R_1 - \hat{R}_1, h\right\rangle_\nu = 0$. To see that the proposed form of $\hat{R}_1$ does indeed satisfy this property, note that

$$\begin{aligned}
\left\langle R_1 - \hat{R}_1, h\right\rangle_\nu &= \left\langle \frac{1}{2}\left\{\frac{p_\omega(\mathbf{Y})}{p_\nu(\mathbf{Y})} - 1 - (\boldsymbol{\mu}_\omega - \boldsymbol{\mu}_\nu)^T \mathbf{V}_\nu^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\nu)\right\}, h\right\rangle_\nu \\
&= \frac{1}{2}\left\{\left\langle \frac{p_\omega(\mathbf{Y})}{p_\nu(\mathbf{Y})}, h\right\rangle_\nu - \langle 1, h\rangle_\nu - (\boldsymbol{\mu}_\omega - \boldsymbol{\mu}_\nu)^T \mathbf{V}_\nu^{-1}\langle \mathbf{Y} - \boldsymbol{\mu}_\nu, h\rangle_\nu\right\} \\
&= \frac{1}{2}\left\{E_\nu\left(\frac{p_\omega(\mathbf{Y})}{p_\nu(\mathbf{Y})}h\right) - E_\nu(h) - (\boldsymbol{\mu}_\omega - \boldsymbol{\mu}_\nu)^T \mathbf{V}_\nu^{-1} E_\nu\left[(\mathbf{Y} - \boldsymbol{\mu}_\nu)h\right]\right\} \\
&= \frac{1}{2}\left\{\int \frac{p_\omega(\mathbf{Y})}{p_\nu(\mathbf{Y})} h p_\nu - 0 - (\boldsymbol{\mu}_\omega - \boldsymbol{\mu}_\nu)^T \mathbf{V}_\nu^{-1} E_\theta\left[(\mathbf{Y} - \boldsymbol{\mu}_\nu)\mathbf{a}^T(\mathbf{Y} - \boldsymbol{\mu}_\nu)\right]\right\} \\
&= \frac{1}{2}\left\{\int p_\omega(\mathbf{Y}) h - (\boldsymbol{\mu}_\omega - \boldsymbol{\mu}_\nu)^T \mathbf{V}_\nu^{-1}\mathbf{V}_\nu \mathbf{a}\right\} \\
&= \frac{1}{2}\left\{E_\omega(h) - (\boldsymbol{\mu}_\omega - \boldsymbol{\mu}_\nu)^T \mathbf{a}\right\} \\
&= \frac{1}{2}(0) \\
&= 0.
\end{aligned}$$

Now, let $R_2 = \frac{p_\omega(\mathbf{Y}) - p_\nu(\mathbf{Y})}{2p_\omega(\mathbf{Y})}$. Here, consider the closed subspace of L^2 given by $L_\omega = span\{Y_1 - \mu_{1\omega}, \ldots, Y_b - \mu_{b\omega}\}$ with inner product $\langle g_1, g_2\rangle_\omega = E_\omega(g_1 g_2)$. Denote the projection of R_2 onto L_ω by $\hat{R}_2$. By the same reasoning that was used to show that $\hat{R}_1$ was the unique projection of R_1 onto L_ν, it can be shown that $\hat{R}_2 = \frac{1}{2}\left\{(\boldsymbol{\mu}_\omega - \boldsymbol{\mu}_\nu)^T \mathbf{V}_\omega^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\omega)\right\}$ is the unique projection of R_2 onto L_ω. The sum of two projections is itself a projection. Thus, the projection of the log likelihood approximation onto $L_\nu \oplus L_\omega$ is given by the linear deviance

$$\hat{R}(\omega, \nu, \mathbf{Y}) = \hat{R}_1 + \hat{R}_2 = \frac{1}{2}\left\{(\boldsymbol{\mu}_\omega - \boldsymbol{\mu}_\nu)^T \mathbf{V}_\nu^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\nu) + (\boldsymbol{\mu}_\omega - \boldsymbol{\mu}_\nu)^T \mathbf{V}_\omega^{-1}(\mathbf{Y} - \boldsymbol{\mu}_\omega)\right\}.$$

The linear deviance approximates the log likelihood ratio. Thus, an approximation of the likelihood ratio can be obtained via an exponential transformation.

The projection approach proposed by Li[44] has several useful properties. First, Li's approach holds very generally. Unlike quasi-likelihood, which requires continuous parameter spaces, Li's approach holds even for discrete parameter spaces. In addition, as noted previously, Li's approach applies to dependent observations and can be used with longitudinal data. Further, $\hat{R}$ is antisymmetric and linear in observations. Since $\hat{R}$ is linear in $\mathbf{Y}$, the deviance is defined for all $\mathbf{Y}$ in the sample space and its behavior can be observed locally at an alternative parameter by using noncentral moments. Finally, $\hat{R}$ is invariant under affine transformations of $\mathbf{Y}$ and under a change of coordinate system in the parameter space.

2.2.5 Empirical Likelihood Approach

Qin & Lawless [68] propose an alternative approach for approximating the likelihood ratio, which is based on empirical likelihood. Briefly, let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent and identically distributed (iid) observations from a d-variate distribution F with an associated $p \times 1$ parameter $\boldsymbol{\theta}$. Although Qin & Lawless require that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be iid observations, this work has been generalized to handle the situation where observations are independent but not identically distributed. This extension requires one to assume that the estimating functions are iid from a super population of estimating functions. Thus, for the remainder of this section, we take $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to be independent observations from F with an associated parameter $\boldsymbol{\theta}$. By definition, the empirical likelihood function is

$$L(F) = \prod_{i=1}^{n} dF(\mathbf{x}_i) = \prod_{i=1}^{n} p_i,$$

where $p_i = dF(\mathbf{x}_i) = Pr(\mathbf{X} = \mathbf{x}_i)$. It can be shown that the empirical likelihood function is maximized by the empirical distribution function $F_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{x}_i < \mathbf{x})$. The empirical likelihood ratio is then

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i.$$

Next, assume that the information available about $\boldsymbol{\theta}$ and F is in the form of $r \geq p$ functionally independent unbiased estimating functions, $g_j(x;\boldsymbol{\theta}); j = 1, 2, \ldots, r$, where $E_F\{g_j(\mathbf{x};\boldsymbol{\theta}) = 0\}$. In vector form, Qin and Lawless consider

$$g(\mathbf{x};\boldsymbol{\theta}) = (g_1(\mathbf{x};\boldsymbol{\theta}), \ldots, g_r(\mathbf{x};\boldsymbol{\theta}))^T$$

where

$$E_F\{g(\mathbf{x};\boldsymbol{\theta})\} = 0.$$

In this framework, the empirical likelihood approach to approximating the likelihood ratio is essentially a constrained maximization problem. Qin and Lawless suggest that the empirical likelihood function, $L(F) = \prod_{i=1}^{n} p_i$, be maximized subject to the following restrictions: $p_i \geq 0$, $\sum_i p_i = 1$, and $\sum_i p_i g(\mathbf{x}_i;\boldsymbol{\theta}) = 0$. For a given $\boldsymbol{\theta}$, a unique maximum exists provided that 0 is inside the convex hull of the points $g(\mathbf{x}_1;\boldsymbol{\theta}), \ldots, g(\mathbf{x}_n;\boldsymbol{\theta})$. The maximum can be found using Lagrange multipliers. The objective function of interest is

$$H = \sum_i log p_i + \lambda\left(1 - \sum_i p_i\right) - n\mathbf{t}^T \sum_i p_i g(\mathbf{x}_i;\boldsymbol{\theta}),$$

where λ and $\mathbf{t} = (t_1, t_2, \ldots, t_r)^T$ are Lagrange multipliers. It follows that, by taking

derivatives with respect to p_i,

$$\frac{\partial H}{\partial p_i} = \frac{1}{p_i} - \lambda - n\mathbf{t}^T g(\mathbf{x}_i;\boldsymbol{\theta}) = 0.$$

Multiplying both sides by p_i and then summing over all possible values of i yields

$$\begin{aligned} 0 &= \frac{p_i}{p_i} - \lambda p_i - n\mathbf{t}^T p_i g(\mathbf{x}_i;\boldsymbol{\theta}) \\ 0 &= 1 - \lambda p_i - n\mathbf{t}^T p_i g(\mathbf{x}_i;\boldsymbol{\theta}) \\ 0 &= \sum_{i=1}^{n} 1 - \lambda \sum_{i=1}^{n} p_i - n\mathbf{t}^T \sum_{i=1}^{n} p_i g(\mathbf{x}_i;\boldsymbol{\theta}) \\ 0 &= n - \lambda. \end{aligned}$$

Thus, $n = \lambda$. Further, an estimate of p_i is given by

$$\begin{aligned} \frac{1}{p_i} - \lambda - n\mathbf{t}^T g(\mathbf{x}_i;\boldsymbol{\theta}) &= 0 \\ \frac{1}{p_i} = n + n\mathbf{t}^T g(\mathbf{x}_i;\boldsymbol{\theta}) \quad &since\ \lambda = n \\ \frac{1}{p_i} &= n\left(1 + \mathbf{t}^T g(\mathbf{x}_i;\boldsymbol{\theta})\right) \\ p_i &= \left(\frac{1}{n}\right)\frac{1}{1 + \mathbf{t}^T g(\mathbf{x}_i;\boldsymbol{\theta})} \end{aligned}$$

with the restriction that $\frac{1}{n}\sum_i \frac{1}{1+\mathbf{t}^T g(\mathbf{x}_i;\boldsymbol{\theta})} g(\mathbf{x}_i;\boldsymbol{\theta}) = 0$ since $\sum_i p_i g(\mathbf{x}_i;\boldsymbol{\theta}) = 0$. Thus, $\mathbf{t}$ can be determined in terms of $\boldsymbol{\theta}$. Note that, since $0 \le p_i \le 1$, $\mathbf{t}$ and $\boldsymbol{\theta}$ must satisfy $1 + \mathbf{t}^T g(\mathbf{x}_i;\boldsymbol{\theta}) \ge \frac{1}{n}$. For a fixed $\boldsymbol{\theta}$, define $D_{\boldsymbol{\theta}}$ to be the set of all $\mathbf{t}$ that satisfy this condition, i.e. $D_{\boldsymbol{\theta}} = \left\{\mathbf{t} : 1 + \mathbf{t}^T g(\mathbf{x}_i;\boldsymbol{\theta}) \ge \frac{1}{n}\right\}$. It can be shown that $D_{\boldsymbol{\theta}}$ is convex, closed, and bounded if 0 is inside the convex hull of the $g(\mathbf{x}_i,\boldsymbol{\theta})$'s. In addition,

$$\frac{\partial}{\partial \mathbf{t}}\left\{\frac{1}{n}\sum_i \frac{1}{1 + t^T g(\mathbf{x}_i;\boldsymbol{\theta})} g(\mathbf{x}_i;\boldsymbol{\theta})\right\} = -\frac{1}{n}\sum_i \frac{g(\mathbf{x}_i;\boldsymbol{\theta}) g^T(\mathbf{x}_i;\boldsymbol{\theta})}{\left\{1 + \mathbf{t}^T g(\mathbf{x}_i;\boldsymbol{\theta})\right\}^2}$$

is negative definite for $\mathbf{t}$ in $D_{\boldsymbol{\theta}}$ provided that $\sum_i g(\mathbf{x}_i;\boldsymbol{\theta}) g^T(\mathbf{x}_i;\boldsymbol{\theta})$ is positive definite.

Using this information, Qin and Lawless conclude that the empirical likelihood function for θ is given by

$$L_E(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \left(\frac{1}{n}\right) \frac{1}{1 + \mathbf{t}^T g(\mathbf{x}_i; \boldsymbol{\theta})} \right\}$$

and the empirical log-likelihood ratio comparing the proposed model to the unconstrained model is

$$l_E(\boldsymbol{\theta}) = \sum_{i=1}^{n} log \left[1 + \mathbf{t}^T(\boldsymbol{\theta}) g(\mathbf{x}_i; \boldsymbol{\theta})\right].$$

The values of $\mathbf{t}$ and $\boldsymbol{\theta}$ must be solved iteratively. An outer loop can be used to maximize $l_E(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, while an inner loop will solve $\mathbf{t}$ for each value of $\boldsymbol{\theta}$ and be updated each time a new $\boldsymbol{\theta}$ is chosen or computed. Due to the inner and outer loops involved in obtaining an estimator of the log likelihood ratio, the empirical likelihood approach of Qin and Lawless can be computationally intensive.

2.2.6 Quadratic Inference Function

Although GEEs consistently estimate regression parameters even when the correlation structure is misspecified, the estimator of the regression parameter can be inefficient under such misspecification. Qu et al. [69] introduced a quadratic inference function (QIF) method, which is essentially a generalized method of moments approach based on an extended GEE. The QIF method provides an alternate estimation approach for longitudinal data, which does not require direct estimation of the correlation parameter and remains optimal even under misspecification of the correlation structure. Additionally, Qu et al. proposed using the QIF approach as a measure of goodness-of-fit.

Let y_{ij} be an outcome variable and $\mathbf{x}_{ij}$ denote a $p \times 1$ vector of covariates observed at the j^{th} measurement occasion for subject i, where $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$.

Further, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})'$ be an $m_i \times 1$ vector of outcomes for subject i and $\boldsymbol{\beta}$ be a $p \times 1$ vector of regression parameters. Assume that the $E(y_{it}) = \mu(x'_{it}\beta)$ and let $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})$. Finally, let V_i be the covariance matrix of the vector $\mathbf{y}_i$. As seen in the GEE approach to modeling longitudinal data, the optimal estimating equation is given by the quasi-likelihood equation

$$g_{opt} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} V_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i).$$

When using the GEE approach, V_i is often unknown and a working variance model $V_i = \phi \mathbf{A}_i^{1/2}(\boldsymbol{\mu}_i) \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}(\boldsymbol{\mu}_i)$ with working correlation matrix $R(\boldsymbol{\alpha})$ is used during estimation. The QIF approach assumes that the inverse of the working correlation matrix can be written as

$$\mathbf{R}^{-1} = \sum_{i=1}^{m} a_i \mathbf{M}_i,$$

where $\mathbf{M}_1, \ldots, \mathbf{M}_m$ are known matrices and $a_1, \ldots, a_m$ are unknown constants. This class of matrices accommodates most of the commonly used correlation structures including the exchangeable, and autoregressive correlation structures. In the QIF approach, the components of the above linear expression for the inverse of the working correlation matrix are used to form an extended score equation based on each of the $M_1, \ldots, M_m$. Specifically, the extended score g_N takes the form

$$g_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} g_i(\boldsymbol{\beta}) = \frac{1}{N} \begin{pmatrix} \sum_{i=1}^{N} (\dot{\boldsymbol{\mu}}_i) \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_1 \mathbf{A}_i^{-\frac{1}{2}} (\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \sum_{i=1}^{N} (\dot{\boldsymbol{\mu}}_i) \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_m \mathbf{A}_i^{-\frac{1}{2}} (\mathbf{y}_i - \boldsymbol{\mu}_i) \end{pmatrix},$$

where $\boldsymbol{\beta}$ represents the regression parameters of interest and $\mathbf{A}_i$ is the diagonal marginal covariance matrix for the i^{th} cluster.

The vector g_N contains more estimating equations then parameters; however,

using the generalized method of moments,the estimating equations can be combined optimally. Based on the extended score, g_N, an objective function is constructed. This objective function is called the QIF and is defined as

$$Q_N(\boldsymbol{\beta}) = g_N' C_N^{-1} g_N,$$

where $C_N = \left(\frac{1}{N^2}\right) \sum_{i=1}^{N} g_i(\boldsymbol{\beta}) g_i'(\boldsymbol{\beta})$. The QIF can be used in two manners. First, one can optimally estimate $\boldsymbol{\beta}$ directly by selecting the value of $\boldsymbol{\beta}$ that minimizes the QIF. Additionally, QIF provides a chi-squared inference function for testing nested GEE models and a chi-squared regression misspecification test. In this context, it is important to note that that the QIF mimics the properties of the log-likelihood function [48]. In particular, $Q_N(\boldsymbol{\beta}_0) - Q_N\left(\hat{\boldsymbol{\beta}}\right)$ is asymptotically chi-squared with degrees of freedom equal to the dimension of $\boldsymbol{\beta}$ and $Q_N\left(\hat{\boldsymbol{\beta}}\right)$ is asymptotically chi-squared as a test statistic for testing whether the semi-parametric model is true.

2.3 Model Selection Diagnostics for Longitudinal Data

2.3.1 Overview

One of the major difficulties of extending latent class analysis to accommodate longitudinal data via GEEs is that a likelihood function is no longer available. Thus, the aforementioned likelihood-based approaches for selecting the order of a finite mixture model are no longer directly applicable. For hypothesis testing, one might consider using an estimate of the likelihood ratio based on a projection or based on the empirical likelihood function. Li[44] notes that the linear deviance function that approximates the log likelihood-based deviance via projection follows a chi-square distribution asymptotically when regularity conditions hold; for example, when the

parameter space is an open set. Similarly, Qin and Lawless[68] note that the empirical likelihood ratio test asymptotically follows a chi-square distribution. Unfortunately, the regularity conditions required for such asymptotic properties to hold may break down in the mixture model context or under transformation.

To avoid the distributional requirements of hypothesis testing, focus is placed primarily on extensions of information criteria for longitudinal data. Direct extensions of information criteria might involve replacing the likelihood portion of the criteria with an alternate measure of model fit such as the empirical likelihood function [68], quasi-likelihood function, or quadratic inference function (QIF) [69]. A literature review of model selection procedures for estimating equations did reveal several extensions of information-based approaches. A brief summary of some of these approaches, as well as other methods available for model selection follows. Note that when choosing a model selection criteria for use with finite mixture models, the approach must be invariant to the scale of the covariates and avoid component labeling issues. Thus, approaches involving direct comparisons of the estimated parameter vectors from two models, such as Wang(2007) [79], are omitted due to the class labeling issues for finite mixture models.

2.3.2 Quasi-Likelihood Under the Independence Model Criterion (QIC)

Pan(2001)[60] proposed an information criterion that he refers to as the quasi-likelihood under the independence model criterion, denoted QIC(R). QIC(R) is an extension of Akaike Information Criterion (AIC) for generalized estimating equations. Consider a random sample of n individuals with longitudinal measurements $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})$ with corresponding covariates $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i})'$. Assume that $\mathbf{Y}_i$ and $\mathbf{Y}_i'$ are independent for $i \neq i'$, but that the components of a given $\mathbf{Y}_i\,(i = 1, \ldots, n)$ may be correlated. The set of all available data will be denoted by $D = \{(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_n, \mathbf{X}_n)\}$.

The relationship between the response and covariates can be then be modeled via a GEE with regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$

Now, recall that AIC is derived from the the Kullback-Leibler (KL) information. It assumes a fully-specified likelihood and takes the form:

$$AIC = -2l\left(\hat{\boldsymbol{\beta}}; D\right) + 2p,$$

where $l\left(\hat{\boldsymbol{\beta}}; D\right)$ denotes the log-likelihood function and p denotes the dimension of $\boldsymbol{\beta}$. In the GEE context, we do not have a likelihood function; however, a quasi-likelihood function may be available. QIC(R) is developed by replacing the likelihood in the KL information with the quasi-likelihood under the independence model, $Q\left(\boldsymbol{\beta}; \mathbf{I}, D\right)$. More specifically, for any working correlation matrix R, QIC(R) is defined as

$$QIC(R) = -2Q\left(\hat{\boldsymbol{\beta}}(R); \mathbf{I}, D\right) + 2trace\left(\hat{\sigma}_I \hat{V}_r\right),$$

where $\hat{V}_r$ is the sandwich covariance estimator of $cov(\hat{\boldsymbol{\beta}})$ and $\hat{\sigma}_I = \frac{-\partial^2 Q(\boldsymbol{\beta}; \mathbf{I}, D)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$.

The authors recommend using QIC(I), i.e. QIC with an independent working correlation structure, whenever possible, due to its superior performance in model selection simulations. Additionally, they note that for correlated data QIC(R) can be approximated by

$$QIC_u(R) = -2Q\left(\hat{\boldsymbol{\beta}}(R); \mathbf{I}, D\right) + 2p.$$

It is important to recognize that while QIC(R) can be used for variable selection, it is not designed to select the working correlation matrix.

The QIC approach described was developed under the assumption that the dispersion parameter, ϕ was known and, hence, could be ignored in the quasi-likelihood function. In practice, the dispersion parameter is typically unknown. Thus, an es-

timate of the dispersion parameter based on the largest model available, say $\hat{\phi}$, is sometimes plugged in. As an alternative, the authors note that using the extended quasi-likelihood [59] would provide a more difficult, but general approach. Wang and Hin [82] present an extension of QIC based on the extended quasi-likelihood, which they refer to EQIC.

2.3.3 Empirical Information Criterion (EIC)

Kolaczyk(1995) [42] proposed an alternate extension of Akaike's information criterion (AIC) for longitudinal data that is based on empirical likelihood. Using the same notation as in the previous discussion of empirical likelihood, let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent and identically distributed d-variate random variables with a common distribution function F. Recall that, by definition, the empirical likelihood function is

$$L(F) = \prod_{i=1}^{n} dF(\mathbf{x}_i) = \prod_{i=1}^{n} p_i,$$

where $p_i = dF(\mathbf{x}_i) = Pr(\mathbf{X} = \mathbf{x}_i)$. Further, the empirical log-likelihood ratio can be expressed as

$$l_{EL}(\theta) = \sum_{i=1}^{n} log\left[1 + \mathbf{t}^T(\boldsymbol{\theta}) g(\mathbf{x}_i; \boldsymbol{\theta})\right].$$

Kolaczyk [42] notes that, by construction, the set $\{p_i(\theta); i = 1 \ldots, n\}$ is a proper probability distribution on the sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Hence, an empirical likelihood analogue of the Kullback-Leibler (KL) information can be obtained by replacing the log-likelihood ratio statistic with the empirical likelihood alternative

$$\sum_{i=1}^{n} log\left\{\frac{p_i(\boldsymbol{\theta}^*)}{p_i(\boldsymbol{\theta})}\right\} p_i(\boldsymbol{\theta}^*),$$

where $\boldsymbol{\theta}^*$ denotes the true value of the parameter vector of interest. In words, this means that the loss entailed in modeling $\boldsymbol{\theta}^*$ by θ is defined to be the Kullback-Leibler distance between two discrete empirical distributions corresponding to these two parameter values. It follows that, under the regularity conditions outlined in Kolaczyk[42], the empirical information criterion can be expressed as

$$EIC(k) = -2l_{EL}\left(\tilde{\boldsymbol{\theta}}_k\right) + d,$$

where $\tilde{\boldsymbol{\theta}}_k$ denotes the estimated parameter of the fitted submodel.

The EIC(k) statistic has many appealing properties. First, it does not depend on any parametric assumptions regarding the distribution of the data and requires only that the estimating function being considered is unbiased. Further, EIC(k) is an asymptotically unbiased estimate of the risk of modeling $\boldsymbol{\theta}^*$ by the fitted submodel $\tilde{\boldsymbol{\theta}}_k$. Unfortunately, the fact that EIC(k) is an asymptotically unbiased estimate of risk does not ensure that it will be a successful criterion for selecting the optimal model from among a set of candidate models and the effectiveness of EIC for model selection has not yet been investigated.

2.3.4 Bayesian Information Quadratic Inference Function (BIQIF)

Wang and Qu(2009) [81] proposed the Bayesian Information Quadratic Inference Function (BIQIF). The BIQIF approach modifies the Bayesian information criterion (BIC) by replacing the negative of two times the log-likelihood with the quadratic inference function (QIF) [69]. This substitution is motivated by the observation that the role of QIF in the semi-parametric setting is similar to the role of the negative of two times the log-likelihood in the parametric setting.

Consider selecting an appropriate marginal regression model from among a class of candidate models that correspond to select different subsets of covariates. Let $\mathbf{M}$ be

the class of candidate models being considered. Each member of $\mathbf{M}$ can be identified with a unique set m, where m is a subset of $\{1, \ldots, q\}$ and contains the indices of the covariates that are included in the candidate model. Then, let $\boldsymbol{\beta}(m)$ denote the $(q+1) \times 1$ vector which sets the corresponding components of $\boldsymbol{\beta}$ to zero if they are not selected by this model. Further, let $\mathbf{B}(m)$ be the corresponding parameter space. Let $Q_N(\beta)$ be the QIF as defined in Qu [69]. The QIF-based BIC then selects the model in $\mathbf{M}$ which minimizes

$$BIQIF(m) = Q_N\left\{\hat{\boldsymbol{\beta}}(m)\right\} + |\boldsymbol{\beta}(m)|log(N),$$

where $Q_N\left\{\hat{\boldsymbol{\beta}}(m)\right\} = inf_{\boldsymbol{\beta}\in\mathbf{B}(m)}\{Q_N(\boldsymbol{\beta})\}$ and $|\boldsymbol{\beta}(m)|$ denotes the number of non-zero elements in $\boldsymbol{\beta}(m)$. As in BIC, BIQIF contains a term to penalize the lack of fit of the model and a term to penalize the complexity of the model.

Although BIQIF approach to model selection has several appealing properties, the quadratic inference function is not on the same scale as the negative of two times the log-likelihood. More specifically, the asymptotic distribution of $Q_N\left\{\hat{\boldsymbol{\beta}}\right\}$ is χ^2 with $r-q$ degrees of freedom, where r is the dimension of the extended score equation, g_N, and q is the dimension of $\boldsymbol{\beta}$, where $q < r$. Thus, while the quadratic inference function may be useful for a likelihood-ratio-type test, it is not generally applicable as a measure of model fit for information criteria.

2.3.5 Expected Predictive Bias (EBP)

Pan(2001)[61] also proposed a more flexible and data driven model selection approach for estimating equations. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an iid sample of size n from an unknown distribution F and $\boldsymbol{\beta}$ be the $p\times 1$ regression coefficient vector. For a system of estimating equations $S(.|\boldsymbol{\beta})$ satisfying $E_{X_1}(S(X_1|\boldsymbol{\beta})) = 0$, an estimate of $\boldsymbol{\beta}$, denoted

by $\hat{\boldsymbol{\beta}}(X)$ can be obtained by solving the p equations

$$S(\mathbf{X}|\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} S(X_i|\boldsymbol{\beta}) = 0.$$

Define $\hat{\beta}(\mathbf{X})$ to be the resulting estimate of $\boldsymbol{\beta}$.

To proceed with the proposed method, let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be another iid sample from F that is independent from $\mathbf{X}$. The model selection criterion is then constructed to minimize the expected predicted bias (EPB) of the estimating equations, where

$$EPB = E_X E_Y |S\left(\mathbf{Y}|\hat{\boldsymbol{\beta}}(\mathbf{X})\right)|.$$

In practice, only one sample $\mathbf{X}$ is typically available. Thus, a resampling approach is used to estimate the EPB. Specifically, Pan considers an estimate of EPB developed using a bootstrap smoothed cross-validation (BCV) estimate. The BCV approach for estimating the EPB was selected because cross-validation typically gives an almost unbiased estimate but results in large variance. The bootstrap then smooths unstable estimates and reduces variability. The BCV estimate is given by

$$\hat{EPB}_{BCV} = E_{X_*}\left|S\left(\mathbf{X}^{*-}|\hat{\boldsymbol{\beta}}(\mathbf{X}^*)\right)\right|,$$

where $\mathbf{X}^*$ denotes a bootstrap sample taken from $\mathbf{X}$ and $\mathbf{X}^{*-} = \mathbf{X} - \mathbf{X}^*$ contains the observations in $\mathbf{X}$ but not in $\mathbf{X}^*$.

Finally, Monte Carlo simulations are usually used to approximate the $\hat{EPB}_{BCV}$ since a closed-form solution is often not available. In general, a weighted sum of the components of $\hat{EPB}_{BCV}$ can be used as a summary statistic/criterion. This summary statistic is denoted $\hat{EPB}_{BCV_a}$ and the weights are usually chosen to be inversely proportional to the variances of the components of $\hat{EPB}_{BCV}$. As mentioned, this approach is appealing because it is motivated by the observed data; however, it can

be computationally burdensome.

2.4 Generating Correlated Discrete Data

In order to validate the methodology proposed in this dissertation, it will be necessary to perform a series of simulation studies. These simulation studies will require normal, Poisson, and binary data to be generated under an AR(1) correlation structure. A multivariate normal distribution with an appropriate correlation matrix can easily be used to generate appropriately correlated normal data; however, generating correlated discrete data is more challenging.

2.4.1 Correlated Count Data

Madsen and Dalthorp(2007) [10] describe two approaches for simulating count-valued random vectors with a specified mean and correlation structure. The first approach uses a lognormal-Poisson hierarchy (L-P method). The idea behind this approach is to generate a vector of correlated normal variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$ by multiplying a vector of i.i.d. standard normals by the Cholesky decomposition of the appropriate covariance matrix. From there, $\mathbf{Z}$ is transformed to a vector of lognormals, $\mathbf{X}$, by exponentiation. Then, $\mathbf{Y}$ is generated as conditionally independent Poissons with means X_i. Although the L-P method is relatively fast and easy to implement, it can sometimes be tedious to determine the appropriate correlation among the normally distributed random variables that will induce the target correlation among the Poissons. In addition, the L-P method cannot accommodate underdispersed random variables or pairs of strongly correlated random variables with similar means and variances.

An alternative approach for simulating count-valued random vectors is known as the overlapping sums (OS) method. The OS algorithm is based on Holgate's(1964)

observation that if $Y_1 = X + X_1$ and $Y_2 = X + X_2$, then Y_1 and Y_2 are correlated even if X, X_1, and X_2 are independent because they share the common component X. Holgate's observation was generalized to n-vectors of correlated Y_i's by Park and Shin(1998) [62] who took each Y_i to be a sum of independent X's. Park and Shin's algorithm parses the target covariance matrix into a long vector of variances of independent X_j's that, when multiplied by the appropriate matrix T of zeros and ones, sum to the vector of Y_i's with the desired correlations[9]. Although the OS method is not as fast as the L-P method, it can simulate strongly correlated random variables provided that the random variables have similar means and variance. In addition, it accommodates under-dispersed random variables or combinations of under- and over-dispersed variables.

2.4.2 Correlated Binary Data

A variety of approaches have been proposed for generating longitudinally correlated binary data (see [29] for a summary). One such approach was proposed by Qaqish(2003) [67], who introduced a multivariate binary distribution to easily and efficiently simulate correlated binary variables with a given marginal mean vector and correlation matrix. Let $\mathbf{Y} = (Y_1, \ldots, Y_T)^T$ denote a sequence of T binary responses. Further, let the marginal mean of Y_t be $E(Y_t) = Pr(Y_t = 1) = \mu_t$. The proposed approach is then implemented by generating a binary sequence using the conditional distribution for Y_t given $(Y_1, \ldots, Y_{t-1})$, where $t = 2, \ldots, n$. Specifically, Qaqish defines

$$\lambda_t = P(Y_t = 1 | Y_1, \ldots, Y_{t-1}) = \mu_t + \sum_{j=1}^{t-1} b_{jt} (Y_j - \mu_j),$$

where b_{jt} reflects the generic correlation structure. Qaqish's approach is computationally efficient for large T, while offering the flexibility to accommodate non-stationary data and unpatterned correlation. In addition, when compared to other methods

for common stationary processes, the proposed approach allows for a wider range of correlation parameters.

Chapter 3

A Latent Trajectory Model for Longitudinal Data

3.1 Overview

The cross-sectional finite mixture model methodology described in Chapter 2 relied on a well-defined likelihood function. Likelihood-based extensions of finite mixture models that accommodate correlated data tend to be both computationally intensive and sensitive to modeling assumptions. As an alternative to these fully-parametric longitudinal approaches, we propose a computationally simpler and more robust latent trajectory model based on generalized estimating equations (GEEs). In this chapter, the proposed approach is outlined and the results of simulation studies to assess its effectiveness are described.

3.2 The Proposed Latent Trajectory Model

Assume that the observed longitudinal data is of the form $(y_{ijk}, t_{ijk})\,; i = 1, \ldots, n, j = 1, \ldots, J, k = 1, \ldots, m_{ij}$. Here, y_{ijk} denotes the k^{th} measurement of the j^{th} feature variable for the i^{th} subject and t_{ijk} denotes the time of observation y_{ijk}. Note that

this data framework is very flexible and allows for unbalanced data.

Since the feature vector for each subject is observed at multiple time points, a natural way to proceed is to model the latent class specific longitudinal trajectories using a generalized estimating equation (GEE) approach, which accommodates the correlation between repeated observations. For the remainder of this chapter, assume that the number of latent classes, C, is fixed and known. As it may not always be realistic to assume that the number of classes is known a priori, methods for assessing the number of components for a mixture of GEEs will be proposed and discussed in Chapter 4. Additionally, assume that the aforementioned longitudinal trajectories are fully characterized by the parameter vector $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \boldsymbol{\phi}^T, \boldsymbol{\gamma}^T\right)^T$, where $\boldsymbol{\beta}$ contains the slope and intercept parameters of the latent class-specific trajectories, $\boldsymbol{\phi}$ contains the latent class-specific dispersion parameters, and $\boldsymbol{\gamma}$ parameterizes the temporal correlation structure of the longitudinal trajectories. Now, let $\mathbf{y}_i = [y_{ijk}]$ and $[\mu_{ijkg}(\boldsymbol{\beta})] = [E(y_{ijk}|(\mathbf{Z}_i)_g = 1)]$, where $(\mathbf{Z}_i)_g = z_{ig} = 1$ is an indicator of whether subject i belongs to class g for $i = 1, \ldots, n$ and $g = 1, \ldots, C$. Note that, although time is the only covariate that will be considered in the simulations and applications that follow, this model formulation can accommodate other covariates when modeling the class-specific mean structure. Further, denote the squared residual by $s_{ijk} = (y_{ijk} - \mu_{ijkg}(\boldsymbol{\beta}))^2$ and the cross-product by $r_{ijklg} = (y_{ijk} - \mu_{ijkg}(\boldsymbol{\beta}))(y_{ijl} - \mu_{ijlg}(\boldsymbol{\beta}))$ for $k < l$. Then, let $\mathbf{s}_{ig} = [s_{ijk}]$ and $\boldsymbol{\sigma}^2_{ig}(\boldsymbol{\phi}, \boldsymbol{\beta}) = E[\mathbf{s}_{ig}]$, where $\sigma^2_{ijkg} = \phi_j v(\mu_{ijkg})$. Finally, let $\mathbf{r}_{ig} = [r_{ijklg}]$ and $\boldsymbol{\eta}_{ig}(\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\beta}) = E[\mathbf{r}_{ig}]$, where $\eta_{ijklg} = \rho_{ijklg}(\boldsymbol{\gamma})\boldsymbol{\phi}_j\sqrt{v(\mu_{ijkg})v(\mu_{ijlg})}$. The latent-class specific second order GEE1 is then given by

$$h_g(\mathbf{y}_i; \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_{ig}(\boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}} \mathbf{V}^{-1}_{ig}(\boldsymbol{\theta}) \left\{\mathbf{y}_i - \boldsymbol{\mu}_{ig}(\boldsymbol{\beta})\right\} \\ \frac{\partial \boldsymbol{\eta}_{ig}(\boldsymbol{\phi}, \boldsymbol{\beta})^T}{\partial \boldsymbol{\phi}} \mathbf{I} \left\{\mathbf{s}_{ig} - \boldsymbol{\sigma}^2_{ig}(\boldsymbol{\phi}, \boldsymbol{\beta})\right\} \\ \frac{\partial \boldsymbol{\psi}_{ig}(\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\beta})^T}{\partial \boldsymbol{\gamma}} \mathbf{I} \left\{\mathbf{r}_{ig} - \boldsymbol{\eta}_{ig}(\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\beta})\right\} \end{pmatrix},$$

where $V_{ig}(\boldsymbol{\theta})$ is a block diagonal, and thus easily invertible, subject-specific variance-

covariance matrix fully-specified by $\boldsymbol{\theta}$. A second order model was used because, although primary interest is in modeling the mean, the correlation and scale parameters are still of secondary interest. The estimating functions are orthogonal to ensure that inferences on the mean trajectories remain valid even when the variance-covariance structure is misspecified.

While the first component of the proposed class-specific GEE model is optimally weighted, the remaining components are not. The suboptimal weighting is used because optimal weighting of the remaining components would require additional knowledge of the third and fourth moments. Additionally, note that each component of the GEE above is permitted to have its own unique link function. The mean link has been well studied and, when multiple feature variables are present, each feature variable may have a different link function. Here, the identity link was used for both the scale and correlation components of the estimating equation to allow for estimation via standard software packages. Finally, a natural choice for the working correlation structure is an autoregressive (AR1) correlation model. The AR1 correlation structure accounts for temporal correlations resulting from repeated measurements while avoiding overly complicated computations. Note, however, that the AR1 correlation structure implicitly assumes that repeated observations occur at evenly spaced time intervals. More complicated correlation structures may be considered for data with unevenly spaced measurements provided that estimation remains feasible.

The proposed GEE model assumes that missing data is missing completely at random (MCAR) within a given latent class. That is, given a participant's observed feature variables and latent class, it is assumed that the missingness mechanism does not depend on past or future values for that feature. The MCAR assumption seems justified for the motivating example on mild cognitive impairment because patients with mild cognitive impairment are not so cognitively impaired as to prevent them from completing the neuropsychological evaluations. If a violation of the MCAR

assumption is suspected, the above approach can be naturally extended to handle the situation in which missing data is missing at random (MAR) by adding suitable weights to the GEE above [71]. The GEE model with suitable weights should yield an unbiased estimating function and consistent parameter estimates.

Now, let $\mathbf{x}_i$ be a $P \times 1$ vector of covariates not involved in modeling the class-specific longitudinal trajectories but potentially influential in determining the class membership probabilities. Denote the class membership probabilities by $\pi(\mathbf{x}_i; \boldsymbol{\alpha}_g)$. Then, recall that estimation of cross-sectional finite mixture models relies on the score equations:

$$\begin{aligned} \textstyle\sum_{g=1}^{C} \sum_{i=1}^{n} \tau_{ig}(\mathbf{y}_i; \boldsymbol{\psi}) \frac{\partial log \pi_g(\mathbf{x}_i; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} &= \mathbf{0} \\ \textstyle\sum_{g=1}^{C} \sum_{i=1}^{n} \tau_{ig}(\mathbf{y}_i; \boldsymbol{\psi}) \frac{\partial log f_g(\mathbf{y}_i; \boldsymbol{\theta}_g)}{\partial \boldsymbol{\theta}} &= \mathbf{0} \quad . \end{aligned}$$

Assuming that the covariates used to model π_{ig} are not time-dependent, the score equation with respect to $\boldsymbol{\alpha}$ remains unaltered in the proposed GEE extension. The likelihood-based score equation with respect to $\boldsymbol{\theta}$ can be generalized to handle correlated longitudinal data by replacing $\frac{\partial f_g(\mathbf{y}_i; \boldsymbol{\theta}_g)}{\partial \boldsymbol{\theta}_g}$ with the analogous class-specific GEE model $h_g(\mathbf{y}_i; \boldsymbol{\theta})$. The resultant quasi-score equation with respect to $\boldsymbol{\theta}$ will be denoted by $Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{g=1}^{C} \tau_{ig}(\mathbf{y}_i; \boldsymbol{\psi}) h_g(\mathbf{y}_i; \boldsymbol{\theta})$. This quasi-score equation remains unbiased. To see this, note that

$$\begin{aligned} E\{Q(\boldsymbol{\theta})\} &= \sum_{i=1}^{n} E\left[\sum_{g=1}^{C} E(z_{ig}|\mathbf{y}_i, \mathbf{x}_i) h_g(\mathbf{y}_i; \boldsymbol{\theta})\right] \\ &= \sum_{i=1}^{n} E\left[\sum_{g=1}^{C} E\{z_{ig} h_g(\mathbf{y}_i; \boldsymbol{\theta})|\mathbf{y}_i, \mathbf{x}_i)\}\right] \\ &= \sum_{i=1}^{n} E\left[\sum_{g=1}^{C} E\{z_{ig} h_g(\mathbf{y}_i; \boldsymbol{\theta})|\mathbf{z}_i)\}\right] \\ &= \mathbf{0}. \end{aligned}$$

Next, consider the posterior class membership probabilities, τ_{ig}. As shown below, τ_{ig} can be expressed in terms of likelihood ratios comparing the evidence that a given subject is in component g versus component 1 of the population:

$$\begin{aligned}
\tau_{ig} &= \frac{\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})f_g(\mathbf{y}_i;\boldsymbol{\theta}_g)}{\sum_{d=1}^{C}\pi_d(\boldsymbol{\alpha})f_d(\mathbf{y}_i;\boldsymbol{\theta}_d)} \\
&= \frac{\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha}))\frac{f_g(\mathbf{y}_i;\boldsymbol{\theta}_g)}{f_1(\mathbf{y}_i;\boldsymbol{\theta}_1)}}{\sum_{d=1}^{C}\pi_d(\boldsymbol{x}_i;\boldsymbol{\alpha})\frac{f_d(\mathbf{y}_i;\boldsymbol{\theta}_d)}{f_1(\mathbf{y}_i;\boldsymbol{\theta}_1)}} \\
&= \frac{\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})LR_{ig}(\boldsymbol{\theta})}{\sum_{d=1}^{C}\pi_d(\boldsymbol{x}_i;\boldsymbol{\alpha})LR_{id}(\boldsymbol{\theta})}.
\end{aligned}$$

Thus, an estimate of τ_{ig} can be obtained by approximating the subject-specific likelihood ratios.

We propose approximating these likelihood ratios via the projection-based approach of Li(1993) [44]. When applying this approach, let

$$\mathbf{Y}_i = (Y_{i11},\ldots,Y_{i1m_{i1}},Y_{i21},\ldots,Y_{i2m_{i2}},\ldots,Y_{iJ1},\ldots,Y_{iJm_{iJ}})^T,$$

where $i = 1,\ldots,n$ indicates the subject, $j = 1,\ldots,J$ indicates the feature variable, and $k = 1,\ldots,m_{ij}$ denotes the measurement occasion of the j^{th} feature for subject i. Further, for the latent class application being considered, let the parameter ω represent the latent class. More rigorously, $\omega = \{\mathbf{Z}_i = (0,0,\ldots,0,1,0,\ldots,0)\}$ is a $C \times 1$ vector of class indicators for subject i. To indicate that subject i is in class g for $g = 1,\ldots,C$, the g^{th} indicator variable, $(\mathbf{Z}_i)_g$ assumes the value of 1 and $\sum_{g=1}^{C}(\mathbf{Z}_i)_g = 1$. Similarly, subject i belongs to class 1 under $\nu = \{\mathbf{Z}_i = (1,0,\ldots,0)\}$. The subject-specific likelihood ratios can then be approximated by $e^{\hat{R}(\omega,\nu,\mathbf{Y})}$, where $\hat{R}(\omega,\nu,\mathbf{Y})$ is the linear deviance function as defined in Li.

As an alternative, the empirical likelihood approach of Qin and Lawless(1994) [68] can be used to approximate the likelihood ratio. In the context being considered, one must assume that the estimating functions are independent and identically dis-

tributed from a super population of estimating functions in order to accommodate dependent observations. In other words, assume that the estimating functions represent a random sample drawn from a latent class specific super-population. Then, consider $\mathbf{Y}_i = (Y_{i11}, \ldots, Y_{i1m_{i1}}, \ldots, Y_{iJ1}, \ldots, Y_{iJm_{iJ}})^T$ where $i = 1, \ldots, n$ indicates the subject, $j = 1, \ldots, J$ indicates the feature variable, and $k = 1, \ldots, m_{ij}$ denotes the measurement occasion of the j^{th} feature for subject i. It follows that the latent-class specific empirical likelihood within each latent class is

$$L_E^g(\boldsymbol{\theta}) := sup \left\{ \prod_{i=1}^{n} p_i^g : p_i^g \geq 0, \sum_i p_i^g = 1, \sum_i p_i^g h_g(\mathbf{y}_i; \boldsymbol{\theta}) = 0 \right\}, g = 1, \ldots, C.$$

Traditionally, the product of $\{p_i^g; i = 1, \ldots, n\}$ is used to make inferences on $\boldsymbol{\theta}$ within class $g = 1, \ldots, C$. By contrast, we propose a somewhat unorthodox use of empirical likelihood in order to compute a subject-specific empirical approximation of p_i in class g, say $\hat{p}_i^g$. Let $\boldsymbol{\theta}_0$ be defined as the null hypothesis that response trajectories are the same across latent classes, so that $\hat{p}_i^g(\boldsymbol{\theta}_0) = \hat{p}_i^1(\boldsymbol{\theta}_0)$ for all $g = 1, \ldots, C$. The empirical approximation of the relevant subject-specific likelihood ratio is given by:

$$LR_{ig}(\boldsymbol{\theta}) \approx \frac{\hat{p}_i^g(\boldsymbol{\theta}) / \hat{p}_i^g(\boldsymbol{\theta}_0)}{\hat{p}_i^1(\boldsymbol{\theta}) / \hat{p}_i^1(\boldsymbol{\theta}_0)} = \frac{\hat{p}_i^g(\boldsymbol{\theta})}{\hat{p}_i^1(\boldsymbol{\theta})}.$$

While the above approach is non-traditional, Kolaczyk(1995) [42] provides validity to the proposed approach by noting that, by construction, the set $\{p_i^g; i = 1, \ldots, n\}$ is a proper probability distribution on the sample $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ for $g = 1, \ldots, C$. The empirical likelihood approach also has several appealing properties. In particular, the approach does not require any extra information about the moment structure of the feature vector than is required to construct the estimating functions. Further, the empirical likelihood approach does not rely on the strong linearity assumption required for Li's projection-based approach. With that said, an empirical likelihood approach would have an increased computational burden. Implementation of the

empirical likelihood-based approach is reserved for future research.

For the proposed model, parameter estimation is done by using a modification of the direct approach to the EM algorithm for GEEs[54]. Specifically, an estimate of $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$ is obtained by iterating between the approximation of

$$\tau_{ig}(\mathbf{y}_i; \boldsymbol{\psi}) = \frac{\pi_g(\boldsymbol{x}_i; \boldsymbol{\alpha}) LR_{ig}(\boldsymbol{\theta})}{\sum_{d=1}^{C} \pi_d(\boldsymbol{x}_i; \boldsymbol{\alpha}) LR_{id}(\boldsymbol{\theta})}$$

and solving

$$\begin{aligned} S(\boldsymbol{\alpha}) &= \sum_{g=1}^{C} \sum_{i=1}^{n} \tau_{ig}(\mathbf{y}_i; \boldsymbol{\psi}) \tfrac{\partial log \pi_g(\boldsymbol{x}_i; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \mathbf{0} \\ Q(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{g=1}^{C} \tau_{ig}(\mathbf{y}_i; \boldsymbol{\psi}) h_g(\mathbf{y}_i; \boldsymbol{\theta}) = \mathbf{0} \quad , \end{aligned}$$

until a pre-specified convergence criteria is met. Recall that a weakness of the EM algorithm is that it can sometimes converge to a spurious local solution. In order to help avoid local solutions, the algorithm is initialized using multiple random starting values for $\tau_{ig}(\mathbf{y}_i; \boldsymbol{\psi})$. Approaches for distinguishing between multiple roots and assessing the number of components in a finite mixture model are discussed in Chapter 4.

3.3 Asymptotic Standard Error

Estimation of the proposed latent trajectory models was done via the EM algorithm. One limitation of the EM algorithm is that it does not automatically provide standard errors. A variety of approaches have been proposed to assess the standard errors of the parameter estimates obtained via the EM algorithm[49, 56, 53]; however, many of these approaches rely heavily on likelihood theory and are not applicable in the current context. Thus, we propose two options for approximating the standard errors of the parameter estimates that can be implemented within the proposed framework. The

first option is to use a bootstrap approach to standard error approximation, where the number of latent classes is held fixed and the bootstrap samples are drawn in the unit of individuals to accommodate repeated measures over time[13]. Since a bootstrap approach to estimation can be computationally intensive even for relatively simple problems, bootstrapping the standard errors for the parameter estimates involved in a finite mixture of estimating functions can be computationally burdensome. As an alternative, we propose an analytical approach based on a sandwich variance estimate. Let

$$\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right)=\left[\begin{array}{c}\frac{\partial log\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\\ h_g\left(\mathbf{y}_i;\boldsymbol{\theta}\right)\end{array}\right].$$

Note that the complete-data estimating function is given by

$$g\left(\mathbf{y},\mathbf{z};\boldsymbol{\psi}\right)=\sum_{i=1}^{n}\sum_{g=1}^{C}z_{ig}\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right),$$

where $\mathbf{z}_i=(z_{i1},\ldots,z_{iC})\sim Multinomial\left(1;\pi_{i1},\ldots,\pi_{iC}\right)$ and $\boldsymbol{\psi}=(\boldsymbol{\alpha},\boldsymbol{\theta})$ as in earlier work. Further, the observed data-estimating function is given by

$$g^*\left(\mathbf{y};\boldsymbol{\psi}\right)=E\left\{g\left(\mathbf{y},\mathbf{z};\boldsymbol{\psi}\right)|\mathbf{y}\right\}=g\left\{\mathbf{y},E\left(\mathbf{z}|\mathbf{y}\right);\boldsymbol{\psi}\right\}=\sum_{i=1}^{n}g^*\left(\mathbf{y}_i;\boldsymbol{\psi}\right)=\sum_{i=1}^{n}\sum_{g=1}^{C}\tau_{ig}\left(\mathbf{y};\boldsymbol{\psi}\right)\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right),$$

where $\tau_{ig}\left(\mathbf{y};\boldsymbol{\psi}\right)=E\left(z_{ig}|\mathbf{y};\boldsymbol{\psi}\right),g=1,\ldots,C$. Then, let $\hat{\boldsymbol{\psi}}$ be a solution to $g^*\left(\mathbf{y};\boldsymbol{\psi}\right)=0$. It follows that $avar\left(\hat{\boldsymbol{\psi}}\right)=A^{-1}BA^{-T}$, where

$$A=E\left\{-\frac{\partial g^*\left(\mathbf{y};\boldsymbol{\psi}\right)}{\partial\boldsymbol{\psi}}\right\}\doteq\sum_{i=1}^{n}\sum_{g=1}^{C}-\left\{\frac{\partial\tau_{ig}\left(\mathbf{y};\boldsymbol{\psi}\right)}{\partial\boldsymbol{\psi}}\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right)+\tau_{ig}\left(\mathbf{y}_i;\boldsymbol{\psi}\right)\frac{\partial\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right)}{\partial\boldsymbol{\psi}}\right\},$$

and

$$
\begin{aligned}
B &= var\left\{g^*\left(\mathbf{y};\boldsymbol{\psi}\right)\right\} = var\left[E\left\{g\left(\mathbf{y},\mathbf{z};\boldsymbol{\psi}\right)\right\}\right] \\
&= var\left\{g\left(\mathbf{y},\mathbf{z};\boldsymbol{\psi}\right)\right\} - E\left[var\left\{g\left(\mathbf{y},\mathbf{z};\boldsymbol{\psi}\right)|\mathbf{y}\right\}\right] \\
&= E\left[var\left\{g\left(\mathbf{y},\mathbf{z};\boldsymbol{\psi}\right)|\mathbf{z}\right\}\right] + var\left[E\left\{g\left(\mathbf{y},\mathbf{z};\boldsymbol{\psi}\right)|\mathbf{z}\right\}\right] - E\left[var\left\{g\left(\mathbf{y},\mathbf{z};\boldsymbol{\psi}\right)|\mathbf{y}\right\}\right] \\
&= E\left[\sum_{i=1}^{n}\sum_{g=1}^{C} z_{ig} var\left\{\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right)|z_{ig}=1\right\}\right] + 0 - E\left[\sum_{i=1}^{n} var\left\{\sum_{g=1}^{C} z_{ig}\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right)|\mathbf{y}\right\}\right] \\
&\doteq \sum_{i=1}^{n}\sum_{g=1}^{C} \tau_{ig}\left(\mathbf{y};\boldsymbol{\psi}\right)\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right)\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right)^T \\
&\quad - \sum_{i=1}^{n}\sum_{g=1}^{C} \tau_{ig}\left(\mathbf{y};\boldsymbol{\psi}\right)\left\{1-\tau_{ig}\left(\mathbf{y};\boldsymbol{\psi}\right)\right\}\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right)\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right)^T \\
&\quad + 2\sum_{i=1}^{n}\sum_{g<d} \tau_{ig}\left(\mathbf{y};\boldsymbol{\psi}\right)\tau_{id}\left(\mathbf{y};\boldsymbol{\psi}\right)\Gamma_g\left(\mathbf{y}_i;\boldsymbol{\psi}\right))\Gamma_d\left(\mathbf{y}_i;\boldsymbol{\psi}\right)^T \\
&= \sum_{i=1}^{n} g^*\left(\mathbf{y}_i;\boldsymbol{\psi}\right) g^*\left(\mathbf{y}_i;\boldsymbol{\psi}\right)^T.
\end{aligned}
$$

Implementation of these approaches is reserved for future research.

3.4 A Simulation Study to Assess the Performance of the Proposed Latent Trajectory Model

Simulation studies were performed to determine the effectiveness of the proposed methodology for identifying latent trajectories when the number of latent classes is fixed and known. Both normal and discrete feature variables were considered. The following subsections describe each simulation's design and results. All simulations were conducted in SAS v9.2 (SAS Institute, Cary, NC). Discrete Poisson data was generated using the discsim 2.1 package in Matlab [9]. Binary data was generated using the CLFsim SAS macro by Qaqish(2003) [67, 66].

3.4.1 Identifying the Mean Structure of Normally Distributed Feature Variables

The first set of simulations focused on whether the proposed approach was able to correctly identify the mean structure of five normally distributed feature variables. Two designs were considered to evaluate the effectiveness of the proposed approach under different class membership probabilities. It was assumed that there were two latent classes. In the first design, subjects were evenly divided between the two classes. In the second, 80% of subjects were assigned to one class and the remaining 20% of subjects were assigned to the other class. For each simulation, 10 distinct realizations of longitudinal data were generated for five hundred individuals. It was assumed that each subject had six evenly spaced measurement occasions ($time = 0, 0.5, 1, 1.5, 2, 2.5$) and no missing data. All feature variables were generated with a standard deviation of 5 and repeated measurements were generated under an autoregressive (AR1) correlation structure with a correlation coefficient of 0.3. Within each latent class, the features were assumed to be independent of one another. This is consistent with the fundamental assumption of local independence underlying latent class analysis. Table 3.1 provides a summary of the intercept/mean of each feature variable. All slopes were taken to be 0.

Table 3.1: Class-specific intercepts of five normally distributed feature variables simulated under an AR(1) correlation structure with a slope of 0, a correlation coefficient of 0.3, and a standard deviation of 5.

Feature	Class 1	Class 2
Feature 1	5	15
Feature 2	25	30
Feature 3	20	20
Feature 4	30	25
Feature 5	15	5

In addition to the feature variables of interest, each subject was randomly assigned a value for a binary covariate. Although this covariate was not involved in the conceptualization of the latent class-specific longitudinal trajectories, it may be a risk factor for belonging to a particular latent class. As such, it was included as a covariate in the polytomous logistic regression models used to determine class membership probabilities.

Estimation was done using the direct extension of the EM algorithm [54] initialized with 100 random starting values for the posterior probabilities of class membership. An AR(1) working correlation structure was used to fit the component-specific GEEs. The stopping criteria for the algorithm was 100 iterations or an absolute difference in parameter estimates between the current and previous iteration of at most 1% for any parameter. Note that all two-class models considered converged to a valid solution in fewer than 100 iterations.

In the first design, 500 subjects were evenly divided between the two latent classes. For each of the 10 data realizations or runs, the proposed latent trajectory model was fit under the assumption of two latent classes. The results do not suggest any problems with multiple roots since, up to class labeling, each of the 100 random starting values converge to the same numeric solution for both the polytomous logistic regression model and the GEEs. A summary of the bias and empirical standard error of the parameter estimates for both simulation designs is shown in Table 3.2, while Table 3.3 shows the bias of the average parameter estimates for each run under Design 1. As shown, the parameter estimates generated under the proposed latent trajectory model are consistently very close to the true parameter values used to generate the data. More specifically, the proposed approach seemed to correctly identify the class-specific intercept and zero slope for the GEEs associated with each of the five normally distributed feature variables. In addition, the intercept and slope for the polytomous logistic regression model reflect an estimated probability of latent class membership

close to 0.50 for each of the two classes. The estimated slope of the polytomous logistic regression model associated with the binary covariate is also close to 0, which is consistent with the fact that each subject was randomly assigned a value for that covariate. Finally, a Kappa coefficient was used to summarize the level of agreement between true class membership and model-based class assignment. Note that model-based class assignment was determined from the estimated posterior probabilities of class membership. Specifically, a subject was classified into the latent class for which they had the largest posterior probability of membership. Here, all subjects were classified into the correct latent class in 7 of the 10 runs. In each of the remaining 3 runs, only a single subject was misclassified ($\kappa = 0.996$). This suggests that the model consistently placed subjects into the correct latent class.

In the second design, 400 subjects were in class one and the remaining 100 subjects were in class 2. Note that class labeling is defined as in Table 3.1. For each of the 10 runs, the proposed latent trajectory model was fit under the assumption of two latent classes. Again, there did not appear to be any problems with multiple roots since, up to class labeling, each of the 100 random starting values converged to the same numeric solution for both the polytomous logistic regression model and the GEEs. Table 3.4 shows the bias of the average parameter estimates for each run, while Table 3.2 provides the average bias and empirical standard error across all runs. Again, the parameter estimates generated under the proposed latent trajectory model are very close to the true parameter values used to generate the data. The approach consistently identified the correct intercept and slope for the five normally distributed feature variables. In addition, when the binary covariate assumed the value 0, the estimated probability of belonging to class 1 ranged from approximately 78% to approximately 83% across the 10 runs. Similarly, when the binary covariate assumed the value 1, the estimated probability of belonging to class 1 ranged from approximately 76% to approximately 82% across the 10 runs. These results are

consistent with the fact that the data was generated under an 80-20 split between two classes and that the true slope associated with the binary covariate in the polytomous logistic regression model is 0. Finally, all subjects were classified into the correct latent class in 50% of the runs. Across the remaining 5 runs, the Kappa coefficient ranged from 0.988 to 0.994. This suggests a high level of agreement between true class membership and model-based class assignment.

Table 3.2: Summary of simulation results for parameter estimates generated for five normally distributed feature variables with equal probabilities of class membership and with unequal probabilities of class membership

			Design 1		Design 2	
			Class Prevalences: 50%, 50%		Class Prevalences: 80%, 20%	
Class	Feature	Parameter	Bias	Standard Error[a]	Bias	Standard Error[a]
1	1	Intercept	0.026	0.263	-0.088	0.209
		Slope	-0.133	0.206	0.052	0.124
2	1	Intercept	0.106	0.225	0.109	0.509
		Slope	-0.106	0.159	0.039	0.213
1	2	Intercept	-0.024	0.316	0.093	0.256
		Slope	0.007	0.204	-0.099	0.170
2	2	Intercept	0.110	0.270	0.037	0.597
		Slope	-0.051	0.140	0.020	0.299
1	3	Intercept	0.052	0.338	-0.008	0.217
		Slope	0.005	0.197	0.008	0.129
2	3	Intercept	-0.061	0.223	-0.039	0.465
		Slope	0.032	0.111	0.090	0.335
1	4	Intercept	-0.012	0.228	0.005	0.170
		Slope	-0.002	0.187	-0.011	0.085
2	4	Intercept	-0.016	0.130	-0.068	0.384
		Slope	0.019	0.166	0.054	0.295
1	5	Intercept	-0.099	0.265	0.014	0.226
		Slope	0.082	0.157	-0.023	0.177
2	5	Intercept	-0.044	0.193	0.037	0.413
		Slope	0.083	0.202	-0.064	0.236
Polytomous		Intercept	0.037	0.109	0.047	0.117
Logistic		Slope	-0.075	0.222	-0.083	0.227

[a] The standard error is determined empirically based on 10 point estimates for the parameter.

Table 3.3: Detailed simulation results for parameter estimates generated for five normally distributed feature variables with equal probabilities of class membership between two latent classes

			Bias of Parameter Estimate									
Class	Feature	Parameter	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
1	1	Intercept	0.045	0.107	0.183	-0.208	0.028	0.597	0.147	-0.203	-0.131	-0.305
		Slope	-0.009	0.054	-0.316	0.157	-0.039	-0.360	-0.137	0.212	-0.062	0.235
2	1	Intercept	0.069	0.011	0.311	0.089	-0.220	-0.142	0.226	0.268	0.389	0.062
		Slope	0.004	-0.189	-0.201	-0.205	0.024	-0.104	-0.006	-0.212	-0.224	0.055
1	2	Intercept	0.052	0.425	0.373	0.236	-0.133	-0.394	-0.575	-0.119	-0.055	-0.046
		Slope	0.057	-0.290	-0.386	-0.077	0.020	0.203	0.191	0.165	0.027	0.162
2	2	Intercept	0.230	-0.238	0.240	0.542	0.302	-0.036	-0.210	0.188	0.144	-0.059
		Slope	-0.024	0.197	-0.059	-0.268	-0.202	-0.072	0.054	-0.075	-0.080	0.018
1	3	Intercept	0.208	-0.196	0.301	-0.049	0.719	-0.097	-0.549	0.114	-0.035	0.101
		Slope	-0.025	0.037	-0.158	0.060	-0.352	0.130	0.394	-0.084	0.109	-0.057
2	3	Intercept	0.255	-0.325	-0.149	-0.066	-0.315	-0.027	-0.236	0.258	-0.119	0.111
		Slope	-0.141	0.218	-0.084	-0.114	0.072	0.065	0.046	-0.120	0.229	0.146
1	4	Intercept	0.043	-0.389	0.276	-0.242	0.385	-0.244	-0.064	-0.022	0.156	-0.022
		Slope	0.014	0.223	-0.130	0.134	-0.453	0.028	-0.004	-0.020	0.043	0.143
2	4	Intercept	0.116	-0.054	-0.030	0.103	-0.199	-0.270	0.093	0.023	0.027	0.035
		Slope	-0.325	0.036	0.234	-0.057	0.189	0.058	-0.077	0.065	-0.094	0.160
1	5	Intercept	-0.115	-0.187	0.034	-0.400	-0.194	0.246	-0.097	0.262	-0.497	-0.037
		Slope	0.170	0.034	-0.024	0.209	0.146	-0.025	0.016	-0.114	0.323	0.089
2	5	Intercept	-0.010	0.044	-0.122	0.140	-0.293	-0.109	-0.069	0.270	0.040	-0.332
		Slope	0.394	-0.061	0.111	-0.196	0.257	0.195	0.045	-0.138	0.052	0.172
Polytomous		Intercept	0.030	0.047	0.038	-0.098	0.045	0.063	0.074	-0.156	0.206	0.117
Logistic		Slope	-0.064	-0.094	-0.085	0.193	-0.071	-0.126	-0.163	0.323	-0.419	-0.243

Table 3.4: Detailed simulation results for parameter estimates generated for five normally distributed feature variables with unequal probabilities of class membership between two latent classes

			Bias of Parameter Estimate									
Class	Feature	Parameter	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
1	1	Intercept	-0.155	-0.297	0.097	0.147	0.108	-0.035	-0.375	0.062	-0.210	-0.220
		Slope	0.042	0.152	-0.056	0.003	-0.062	0.020	0.243	-0.065	0.041	0.205
2	1	Intercept	-0.382	0.430	0.638	-0.083	0.845	-0.100	-0.418	0.180	0.559	-0.583
		Slope	0.115	-0.134	-0.129	0.201	-0.229	0.340	0.254	0.152	-0.246	0.067
1	2	Intercept	0.223	0.056	-0.112	0.374	-0.012	0.461	-0.275	-0.160	0.197	0.180
		Slope	-0.136	0.050	0.063	-0.171	-0.102	-0.289	-0.086	0.104	-0.265	-0.158
2	2	Intercept	-0.175	0.775	0.639	-0.955	0.846	-0.110	-0.412	0.173	0.143	-0.558
		Slope	-0.195	-0.411	-0.129	0.535	-0.229	0.059	0.457	0.125	-0.084	0.074
1	3	Intercept	0.249	-0.089	0.393	-0.220	-0.123	0.193	-0.126	0.031	-0.134	-0.252
		Slope	-0.055	-0.044	-0.134	0.191	0.023	-0.209	0.157	0.136	0.046	-0.029
2	3	Intercept	-0.097	0.822	-0.605	-0.434	0.216	-0.532	-0.412	0.165	0.421	0.066
		Slope	-0.224	-0.429	0.444	0.450	-0.134	0.318	0.458	0.130	-0.170	0.061
1	4	Intercept	0.235	-0.251	-0.056	0.251	0.159	-0.191	0.026	-0.017	-0.114	0.012
		Slope	-0.053	0.100	0.077	-0.098	-0.141	0.061	0.060	-0.016	-0.005	-0.096
2	4	Intercept	-0.039	0.167	-0.404	-0.434	0.023	-0.274	-0.705	0.415	0.421	0.155
		Slope	-0.202	-0.104	0.404	0.449	-0.059	0.228	0.287	-0.427	-0.170	0.130
1	5	Intercept	-0.227	0.111	-0.180	0.450	0.017	-0.198	-0.134	-0.117	0.249	0.167
		Slope	0.152	-0.294	0.243	-0.257	-0.021	0.101	-0.122	0.072	-0.129	0.028
2	5	Intercept	-0.271	0.103	0.614	-0.262	0.287	-0.253	-0.349	-0.470	0.331	0.636
		Slope	-0.065	-0.138	-0.162	-0.061	-0.247	0.223	0.159	0.165	0.018	-0.533
Polytomous		Intercept	-0.073	0.196	0.010	0.198	0.113	0.044	0.087	-0.031	0.042	-0.120
Logistic		Slope	0.148	-0.417	-0.009	-0.357	-0.196	-0.071	-0.171	0.057	-0.060	0.243

3.4.2 Identifying the Intercept and Slope of Normal and Discrete Feature Variables

The results of the first simulation study suggest that the proposed latent trajectory model can effectively detect the presence of the two underlying classes and correctly identify the mean structure based on a set of normally distributed feature variables. In the second set of simulations, we considered a slightly more general simulation design that incorporates non-zero slopes and non-normal feature variables. More specifically, the second simulation contains 6 feature variables. Of these 6 feature variables, 2 are normally distributed, 2 are binary, and 2 follow a Poisson distribution. In addition, the second simulation allows for non-zero slope with respect to time for a subset of the feature variables.

Again, the simulation was performed under the assumption that there were two latent classes and two designs were considered with varying class membership probabilities. Again, in the first design, subjects were evenly divided between two classes. In the second, 80% of subjects belonged to one latent class and the remaining 20% belonged to the other. For each of five hundred individuals, ten distinct realizations of longitudinal data with six evenly spaced measurement occasions and no missing data were generated. Within each latent class, the features were assumed to be independent and data for each feature was generated separately based on an autoregressive (AR1) correlation structure with a correlation coefficient of 0.3. Let β_{0jg} and β_{1jg} for $j = 1, \ldots, 6$ and $g = 1, 2$ be class- and feature-specific intercepts and slopes, respectively. The Poisson feature variables were generated using the overlapping sums (OS) approach of Madsen and Dalthorp(2007) [10] with a mean of $e^{\beta_{0jg}+\beta_{1jg}\times time}(j = 1, 2; g = 1, 2)$. The binary feature variables were generated using the approach of Qaqish(2003)[67] with a mean of $\frac{e^{\beta_{0jg}+\beta_{1jg}\times time}}{1+e^{\beta_{0jg}+\beta_{1jg}\times time}}(j = 3, 4; g = 1, 2)$. The normal feature variables were generated with a standard deviation of 5 and a mean of $e^{\beta_{0jg}+\beta_{1jg}\times time}(j = 5, 6; g = 1, 2)$. As in the first set of simulations, each sub-

ject was also randomly assigned a value for a binary covariate, which may be involved in determining class membership probabilities. Table 3.5 provides a summary of the class-specific intercept and slope used to generate each of the feature variables.

Table 3.5: Class-specific intercepts and slopes of six feature variables simulated under an AR(1) correlation structure with a correlation coefficient of 0.3.

		Class 1		Class 2	
Feature	Distribution	Intercept	Slope	Intercept	Slope
Feature 1	Poisson	0.7	1.0	0.7	0.0
Feature 2	Poisson	3.0	0.0	3.0	-1.0
Feature 3	Binary	-0.5	1.0	-0.5	0.0
Feature 4	Binary	0.5	0.0	0.5	-1.0
Feature 5	Normal	20.0	0.0	20.0	0.0
Feature 6	Normal	5.0	0.0	5.0	5.0

In the first design, 500 subjects were evenly divided between the two classes. For each of the 10 runs, the proposed latent trajectory model was fit under the assumption of two latent classes. There did not appear to be any problems with multiple roots since, up to class labeling, each of the 100 random starting values converged to the same numeric solution for both the polytomous logistic regression model and the GEEs. A summary of the average bias across the 10 runs and empirical standard error of the parameter estimates for the two simulation designs is shown in Table 3.6, while Table 3.7 shows the bias for each individual run under Design 1. As shown, across all 10 runs, the parameter estimates generated under the proposed latent trajectory model are consistently very close to the true parameter values used to generate the data. More specifically, the proposed approach seemed to correctly identify both the class-specific intercept and the class-specific slope for the GEEs associated with each of the six feature variables. In addition, the intercept and slope for the polytomous logistic regression model reflect an estimated probability of latent class membership close to 0.50 for each of the two classes and the estimated slope of the polytomous

logistic regression model associated with the binary covariate is close to its true value of 0. Finally, the Kappa coefficient, which assesses the agreement between true class membership and model-based class assignment, ranged from 0.916 to 1.000 across the 10 runs. This implies that the model usually classified subjects into the correct latent class.

Next, consider the scenario where 400 subjects were in Class 1 and the remaining 100 subjects were in Class 2. The classes are defined as in Table 3.5 . Again, when the number of latent classes was taken to be 2, the 100 random starting values for a particular data realization converged to the same root. This was true for all 10 realizations of the longitudinal data. Table 3.8 shows the bias for each individual run under Design 2. As shown, up to class labeling, the parameter estimates of both the polytomous logistic regression model and the GEEs appear to be the same. Again, the parameter estimates under the proposed latent trajectory model are very close to to the true parameter values used to generate the data. The approach consistently identified the correct intercept and slope for the six feature variables. Finally, when the binary covariate assumed the value 0, the estimated probability of belonging to class 1 ranged from approximately 79% to approximately 84% across the 10 runs. Similarly, when the binary covariate assumed the value 1, the estimated probability of belonging to class 1 ranged from approximately 77% to approximately 82% across the 10 runs. These results are consistent with the fact that the data was generated under an 80-20 split between two classes and that the true slope associated with the binary covariate in the polytomous logistic regression model is 0. Finally, the Kappa coefficient ranged from 0.981 to 1.000 across the 10 runs, which suggests that the model typically classified subjects into the correct latent class.

Table 3.6: Summary of simulation results for parameter estimates generated for six feature variables with equal probabilities of class membership and with unequal probabilities of class membership

			Design 1 Class Prevalences: 50%, 50%		Design 2 Class Prevalences: 80%, 20%	
Class	Feature	Parameter	Bias	Standard Error[a]	Bias	Standard Error[a]
1	1	Intercept	0.011	0.028	-0.002	0.027
		Slope	-0.011	0.019	< 0.001	0.011
2	1	Intercept	-0.012	0.044	-0.007	0.041
		Slope	<0.001	0.015	-0.015	0.041
1	2	Intercept	0.002	0.011	-0.002	0.008
		Slope	-0.005	0.013	-0.002	0.005
2	2	Intercept	-0.005	0.013	0.009	0.030
		Slope	-0.002	0.013	0.006	0.024
1	3	Intercept	-0.034	0.120	-0.024	0.084
		Slope	0.009	0.105	-0.005	0.088
2	3	Intercept	0.007	0.143	0.014	0.118
		Slope	0.001	0.091	-0.009	0.071
1	4	Intercept	-0.030	0.092	-0.034	0.097
		Slope	0.011	0.049	0.019	0.057
2	4	Intercept	-0.010	0.114	0.008	0.138
		Slope	-0.011	0.075	-0.010	0.103
1	5	Intercept	-0.006	0.211	-0.079	0.230
		Slope	0.049	0.118	0.039	0.155
2	5	Intercept	0.187	0.342	0.087	0.410
		Slope	-0.093	0.182	0.045	0.281
1	6	Intercept	0.063	0.159	0.092	0.185
		Slope	0.029	0.146	-0.039	0.139
2	6	Intercept	0.047	0.184	-0.039	0.543
		Slope	0.086	0.162	0.119	0.269
Polytomous Logistic		Intercept	0.068	0.125	0.087	0.144
		Slope	-0.065	0.184	-0.126	0.242

[a] The standard error is determined empirically based on 10 point estimates for the parameter.

Table 3.7: Detailed simulation results for parameter estimates generated for six feature variables with equal probabilities of class membership between two latent classes

			Bias of Parameter Estimate									
Class	Feature	Parameter	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
1	1	Intercept	-0.042	0.004	0.004	0.009	0.011	0.031	0.046	-0.006	0.043	0.007
		Slope	0.019	-0.006	-0.005	-0.007	-0.003	-0.026	-0.027	-0.025	-0.024	-0.002
2	1	Intercept	-0.012	-0.036	0.033	-0.031	-0.039	-0.098	0.013	0.003	0.052	-0.009
		Slope	-0.008	-0.002	-0.007	0.013	0.006	0.033	-0.019	-0.014	-0.008	0.009
1	2	Intercept	0.015	-0.019	-0.005	-0.003	0.010	-0.002	-0.004	0.005	0.017	0.006
		Slope	-0.009	0.009	0.007	-0.002	-0.005	-0.010	0.005	-0.033	-0.007	-0.007
2	2	Intercept	-0.007	-0.015	-0.021	0.010	0.006	0.009	0.006	-0.012	-0.017	-0.010
		Slope	0.011	-0.013	0.021	-0.009	-0.007	-0.017	-0.009	-0.012	-0.003	0.015
1	3	Intercept	-0.172	-0.189	0.038	0.135	0.016	0.069	-0.178	-0.008	-0.006	-0.084
		Slope	0.143	0.083	-0.096	-0.137	-0.061	-0.045	0.057	-0.088	0.090	0.140
2	3	Intercept	0.191	0.148	0.037	-0.102	0.049	0.074	-0.299	0.089	-0.076	-0.046
		Slope	-0.086	-0.135	-0.029	-0.030	-0.016	0.023	0.187	-0.025	0.099	0.025
1	4	Intercept	-0.098	-0.006	0.017	-0.098	-0.088	0.021	-0.176	0.117	0.045	-0.036
		Slope	0.053	-0.059	-0.001	0.024	0.086	-0.010	0.049	-0.066	0.022	0.007
2	4	Intercept	0.052	-0.093	0.112	-0.049	0.105	-0.075	0.039	0.046	0.029	-0.264
		Slope	0.010	0.039	-0.063	0.042	-0.145	-0.080	-0.026	0.039	-0.036	0.110
1	5	Intercept	-0.076	0.044	-0.212	-0.358	-0.011	0.031	-0.182	0.287	0.302	0.115
		Slope	0.179	-0.112	-0.029	0.206	0.140	<0.001	0.087	0.015	0.071	-0.067
2	5	Intercept	0.041	-0.171	0.438	0.234	0.371	0.056	0.449	0.633	-0.172	-0.006
		Slope	0.077	0.124	-0.308	-0.178	-0.232	-0.052	-0.115	-0.246	0.098	-0.101
1	6	Intercept	0.121	0.197	-0.113	0.120	0.119	0.344	-0.108	-0.014	-0.034	-0.005
		Slope	-0.241	-0.052	-0.049	0.079	-0.108	0.115	0.144	0.251	0.116	0.034
2	6	Intercept	0.095	-0.249	-0.011	-0.100	-0.150	-0.022	0.066	-0.358	0.263	0.001
		Slope	-0.029	0.232	0.023	0.026	0.088	0.081	0.110	0.364	-0.116	0.083
Polytomous		Intercept	0.150	-0.012	0.096	-0.033	-0.130	0.104	0.078	0.234	0.097	0.099
Logistic		Slope	-0.273	0.098	-0.177	0.120	0.272	-0.088	-0.127	-0.126	-0.160	-0.184

Table 3.8: Detailed simulation results for parameter estimates generated for six feature variables with unequal probabilities of class membership between two latent classes

			Bias of Parameter Estimate									
Class	Feature	Parameter	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
1	1	Intercept	<0.001	0.016	0.046	-0.029	-0.015	0.014	-0.051	-0.017	0.005	0.006
		Slope	0.001	-0.005	-0.022	0.011	0.006	-0.004	0.020	0.007	0.002	-0.007
2	1	Intercept	0.007	0.025	-0.051	0.033	-0.064	-0.046	-0.006	-0.025	-0.006	0.064
		Slope	-0.034	-0.049	0.042	-0.020	-0.038	0.033	-0.007	0.033	-0.057	-0.053
1	2	Intercept	-0.004	-0.012	-0.007	0.007	-0.007	0.012	-0.011	0.002	0.006	-0.002
		Slope	-0.001	0.002	<0.001	-0.002	-0.005	-0.008	0.007	-0.005	-0.009	<0.001
2	2	Intercept	0.018	0.028	-0.012	0.013	0.011	-0.003	-0.058	-0.028	0.042	-0.021
		Slope	-0.005	-0.016	0.020	0.001	-0.002	-0.015	0.042	0.049	-0.014	-0.003
1	3	Intercept	-0.111	0.037	-0.143	0.042	0.076	-0.048	0.009	-0.054	-0.109	0.063
		Slope	0.073	-0.041	0.070	-0.163	-0.060	-0.043	-0.016	0.103	0.095	-0.072
2	3	Intercept	0.196	-0.043	-0.028	0.090	-0.157	0.159	-0.076	0.098	-0.108	0.012
		Slope	-0.121	0.088	0.043	-0.067	0.041	-0.066	-0.072	-0.026	0.066	0.025
1	4	Intercept	-0.041	-0.020	-0.115	-0.132	0.118	-0.115	0.042	0.090	-0.073	-0.095
		Slope	0.022	0.017	0.028	0.084	-0.059	0.087	-0.034	-0.051	0.018	0.073
2	4	Intercept	-0.029	0.050	-0.014	-0.255	0.202	0.181	0.113	0.009	-0.098	-0.076
		Slope	-0.036	0.026	-0.091	0.219	-0.068	-0.150	-0.006	-0.064	-0.006	0.079
1	5	Intercept	0.160	0.120	0.018	-0.594	0.020	-0.175	-0.172	-0.102	-0.107	0.038
		Slope	-0.018	-0.079	-0.116	0.343	-0.038	0.162	0.016	0.092	0.155	-0.131
2	5	Intercept	0.715	0.603	0.531	-0.157	-0.530	-0.054	0.061	-0.031	-0.072	-0.195
		Slope	-0.562	-0.306	0.001	0.126	0.285	0.332	0.127	0.101	0.120	0.230
1	6	Intercept	0.033	0.035	-0.066	-0.113	0.331	-0.002	0.202	-0.016	0.212	0.303
		Slope	0.035	-0.024	0.220	-0.003	-0.173	0.041	-0.282	-0.066	-0.053	-0.081
2	6	Intercept	0.252	-0.840	1.084	0.052	-0.260	-0.054	0.061	-0.031	0.131	-0.782
		Slope	0.004	0.484	-0.287	-0.213	0.223	0.332	0.127	0.101	0.126	0.292
Polytomous		Intercept	-0.071	-0.030	0.194	0.112	0.286	0.055	0.176	0.099	0.081	-0.033
Logistic		Slope	0.168	0.060	-0.296	-0.168	-0.438	-0.086	-0.332	-0.163	-0.151	0.146

3.5 Discussion

The latent trajectory model based on generalized estimating equations proposed in this chapter is flexible enough to accommodate both discrete and continuous feature variables and overcomes many of the limitations associated with fully-likelihood based extensions of finite mixture models for correlated data. Specifically, the proposed approach does not require the strict modeling assumptions associated with parametric extensions of finite mixture models for longitudinal data and reduces the computational burden associated with likelihood-based models for discrete feature variables. Simulation studies suggest that, when the number of latent classes is known, the proposed approach can correctly detect the presence of underlying classes based on a set of observed feature variables. This is true regardless of whether the class membership probabilities are equal or unbalanced. In addition, for a sufficiently large data set, the proposed approach obtains accurate estimates of the class-specific intercepts and slopes associated with the generalized estimating equations for each feature variable. Finally, the proposed approach accurately estimates the slope and intercept associated with the polytomous logistic regression model used to determine the probabilities of latent class membership.

Chapter 4

Diagnostics for Latent Trajectory Models

4.1 Overview

When developing the latent trajectory model proposed in the previous chapter, it was assumed that the number of latent classes was fixed and known. Since it is often not realistic to assume that the number of latent classes is known a priori, model diagnostics for determining the correct number of latent classes are needed. In addition, recall that a weakness of the EM algorithm is that it can sometimes converge to a spurious local solution. In order to help avoid local solutions, the algorithm is initialized using multiple random starting values for the subject- and class-specific posterior probabilities of class membership, $\tau_{ig}(g = 1, \ldots, C; i = 1, \ldots, n)$. When a fully-specified likelihood function is available, multiple roots can be compared and the root which maximizes the likelihood function is selected; however, in the context of a finite mixture of GEEs, the full distribution of the observations is not known. Thus, in the absence of a likelihood function, an analogous scalar objective function is needed to distinguish between multiple roots and determine the appropriate number

of mixture components. In this chapter, an objective function for distinguishing between multiple roots is proposed. Several information criteria for assessing the number of components are also compared.

4.2 Assessing the Number of Components in a Finite Mixture of Generalized Estimating Equations

4.2.1 Cross-sectional Background

In the cross-sectional context, the likelihood function is used to compare multiple roots. Further, several information criteria based on penalized version of the likelihood function have been proposed for assessing the number of components in a mixture model. A subset of these information criteria relies on what is known as the classification likelihood, $L_C(\boldsymbol{\psi})$. For example, ICL-BIC [3], which was found to perform well in cross-sectional simulation studies [54], is based on the conditional expectation of the classification likelihood.

In the EM framework used for estimation of the cross-sectional finite mixture model, the classification likelihood is often referred to as the complete-data likelihood. Recall that the classification or complete log-likelihood for $\boldsymbol{\psi}$ is given by

$$logL_C(\boldsymbol{\psi}) = \sum_{g=1}^{C}\sum_{i=1}^{n} z_{ig}\left\{log\pi_g(\mathbf{x}_i;\boldsymbol{\alpha}) + logf_g(\mathbf{y}_i;\boldsymbol{\theta}_g)\right\},$$

where z_{ig} is the unobserved component indicator vector for subject i. As noted by Hathaway(1986) [39], the mixture log likelihood, $logL(\boldsymbol{\psi})$, can be expressed as

$$logL(\boldsymbol{\psi}) = logL_C(\boldsymbol{\psi}) - \sum_{g=1}^{C}\sum_{i=1}^{n} z_{ig}log(\tau_{ig}),$$

where τ_{ig} is the posterior probability that subject i belongs to class g given the observed feature variables. The conditional mean of $\sum_{g=1}^{C}\sum_{i=1}^{n} z_{ig} log\left(\tau_{ig}\right)$ given the observed data is then equal to the negative of the entropy. As in the cross-sectional context, the entropy is defined as

$$EN(\boldsymbol{\tau}) = -\sum_{g=1}^{C}\sum_{i=1}^{n} \tau_{ig} log\left(\tau_{ig}\right),$$

where $\boldsymbol{\tau} = \left(\boldsymbol{\tau}_1^T, \ldots, \boldsymbol{\tau}_n^T\right)^T$ and $\boldsymbol{\tau}_i = \left(\tau_{i1}\left(\mathbf{y}_i; \boldsymbol{\psi}\right), \ldots, \tau_{iC}\left(\mathbf{y}_i; \boldsymbol{\psi}\right)\right)$. Following the notation of McLachlan and Peel(2000) [54], denote the complete-data likelihood as $L_C\left(\boldsymbol{\psi}; \mathbf{z}\right)$ to indicate that it is formed on the basis of $\mathbf{z} = \left(\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T\right)^T$ in addition to the observed data. Then, substituting the MLE of $\boldsymbol{\tau}$ for $\mathbf{z}$ in $L_C\left(\boldsymbol{\psi}; \mathbf{z}\right)$ yields

$$logL_C\left(\hat{\boldsymbol{\psi}}; \hat{\boldsymbol{\tau}}\right) = logL\left(\hat{\boldsymbol{\psi}}\right) - EN\left(\hat{\boldsymbol{\tau}}\right).$$

Information criteria based on the classification likelihood make use of this relationship by minimizing

$$Q\left(\boldsymbol{\psi}\right) = -2logL\left(\hat{\boldsymbol{\psi}}\right) + 2EN\left(\hat{\boldsymbol{\tau}}\right).$$

When $\mathbf{y}_i$ includes several feature variables, estimation via the EM algorithm involves fitting a separate class-specific model for each feature variable, $Y_j (j = 1, \ldots, n)$. By the local independence assumption, feature variables are independent within a given latent class. Thus, the sum of the class-specific probability density functions for each individual feature may be taken as a class-specific summary of model fit. This implies that the conditional expectation of the classification log-likelihood when multiple feature variables are present can be expressed as

$$\sum_{g=1}^{C}\sum_{j=1}^{J}\sum_{i=1}^{n} \tau_{ig}(\mathbf{y}_i; \boldsymbol{\psi}) log f_{jg}(\mathbf{y}_{ij}; \boldsymbol{\theta}_g) - EN\left(\boldsymbol{\tau}\right),$$

where $f_{jg}(\mathbf{y}_{ij};\boldsymbol{\theta}_g)$ refers to the feature-specific pdf for subject i in class g.

In the cross-sectional context, $Q(\boldsymbol{\psi})$ is referred to as the classification likelihood criterion (CLC) [5]. In CLC, the term based on the estimated entropy is used to penalize a model for class membership uncertainty, and the finite mixture model is selected by choosing the number of classes that minimizes $Q(\boldsymbol{\psi})$. Recall, that if the components of the mixture are well separated, then the entropy will be close to 0. In contrast, if the mixture components are poorly separated, then the entropy will have a large value. Thus, the degree of separation between the fitted components determines the severity of the penalty term, with more severe penalties imposed for situations in which class membership is more ambiguous. When dealing with cross-sectional finite mixture models, this criterion works well when the mixing proportions are restricted to being equal, but tends to overestimate the number of classes when the mixing proportions are unequal [4]. In order to overcome these limitations, ICL-BIC [3] was proposed. ICL-BIC incorporates an additional penalty of the form $dlogn$, where d refers to the number of unknown parameters in $\boldsymbol{\psi}$ and n refers to the sample size.

4.2.2 Mixture Classification Quasi-Likelihood Approach

In the absence of a likelihood function, we consider replacing the feature-, class- and subject-specific log-likelihood function, $logf_{jg}(\mathbf{y}_{ij};\boldsymbol{\theta}_g)$, with an artificial-likelihood based objective function for assessing the model fit of the class-specific GEE $h_g(\mathbf{y}_i;\boldsymbol{\theta})$. In particular, we draw from Pan's(2001)[60] quasi-likelihood under the independence model criteria (QIC) and consider replacing $logf_{jg}(\mathbf{y}_{ij};\boldsymbol{\theta}_g)$ with a feature-, class-, and subject-specific quasi-likelihood under the independence model. The feature- and class-specific quasi-likelihood under the working independence assumption is defined as

$$Q_{jg}\left(\hat{\boldsymbol{\beta}}_{jg}(\mathbf{R}),\phi_{jg}\right)=\sum_{i=1}^{n}\sum_{k=1}^{m_{ij}}Q_{jg}\left(\hat{\boldsymbol{\beta}}_{jg}(\mathbf{R}),\phi_{jg};(Y_{ik},\mathbf{X}_{ik})\right),$$

where ϕ_{jg} denotes the dispersion parameter, $\mathbf{R}$ denotes the working correlation of interest, and the quasi-likelihood contribution of the k^{th} observation on the i^{th} subject evaluated at the regression parameters $\boldsymbol{\beta}$ is defined as

$$Q_{jg}\left(\hat{\boldsymbol{\beta}}_{jg}\left(\mathbf{R}\right),\phi_{jg};\left(Y_{ik},\mathbf{X}_{ik}\right)\right)=Q_{gijk}/\phi_{jg}.$$

Under the independence model, Q_{gijk} takes a closed-form for many standard distributions. In particular, note that for

- Normal: $Q_{gijk}=-\frac{1}{2}\tau_{ig}(\mathbf{y}_i;\boldsymbol{\psi})\left(y_{ijk}-\mu_{ijk}\right)^2$
- Poisson: $Q_{gijk}=\tau_{ig}(\mathbf{y}_i;\boldsymbol{\psi})\left(y_{ijk}log\left(\mu_{ijk}\right)-\mu_{ijk}\right)$
- Binary: $Q_{gijk}=\tau_{ig}(\mathbf{y}_i;\boldsymbol{\psi})\left[y_{ijk}log(p_{ijk}+\left(1-y_{ijk}\right)log\left(1-p_{ijk}\right)\right]$.

In the above equations, μ_{ijk} refers to the mean or predicted value of y_{ijk} and p_{ijk} refers to the probability that a binary feature variable y_{ijk} attains the value 1.

When Pan[60] developed QIC, he initially assumed that the dispersion parameter ϕ was known; however, in practice ϕ is typically unknown and estimated using the Pearson chi-square residuals. In such situations, one may consider using information criteria based on the extended quasi-likelihood [82]. In the current context, we propose using the following objective function based on the extended quasi-likelihood function [59] as a measure of model fit:

$$Q_{jg}^{+}=Q_{jg}^{+}\left(\hat{\boldsymbol{\beta}}_{jg}\left(\mathbf{R}\right),\phi_{jg}\right)=\sum_{i=1}^{n}\sum_{k=1}^{m_{ij}}Q_{jg}\left(\hat{\boldsymbol{\beta}}_{jg}\left(\mathbf{R}\right),\phi_{jg};\left(Y_{ik},\mathbf{X}_{ik}\right)\right)-\frac{1}{2}log\left(\hat{\phi}_{jg}\right),$$

where

$$\hat{\phi}_{jg}=\frac{\sum_{i=1}^{n}\sum_{k=1}^{m_{ij}}\left(y_{ijk}-\mu_{ijk}\right)^2}{n-p},$$

where p denotes the number of parameters in the model. Note that, by using an objective function based on the extended quasi-likelihood function, models are given an increased penalty for larger dispersion parameters.

Here, we propose using the mixture classification extended quasi-likelihood to distinguish between multiple roots. More specifically, for a given number of latent classes, the root that maximizes the mixture classification extended quasi-likelihood is favored. Once the best root is selected for a given number of latent classes, information criteria to assess the number of components for the latent trajectory model can be constructed as a penalized version of the mixture classification extended quasi-likelihood. In particular, we consider 4 information criteria based on the extended quasi-likelihood under the independence model with various penalties.

- Bayesian extended quasi-likelihood under the independence model criterion (BEQC): $-2\sum_{g=1}^{C}\sum_{j=1}^{J}Q_{jg}^{+} + dlogn$
- Extended Quasi-likelihood under the independence model criterion (EQIC): $-2\sum_{g=1}^{C}\sum_{j=1}^{J}Q_{jg}^{+} + 2d$
- Classification Extended Quasi-likelihood Criterion (CEQC): $-2\sum_{g=1}^{C}\sum_{j=1}^{J}Q_{jg}^{+} + 2EN(\boldsymbol{\tau})$
- Integrated Classification Extended Quasi-likelihood Criterion (CEQ-BIC): $-2\sum_{g=1}^{C}\sum_{j=1}^{J}Q_{jg}^{+} + 2EN(\boldsymbol{\tau}) + dlogn$

As before, d refers to the number of unknown parameters in $\boldsymbol{\psi}$ and n refers to the number of subjects. The number of components can then be determined by comparing the information criteria associated with the best C-component model for $(C = 1, 2, \ldots)$. The model that minimizes the information criteria is selected and its number of components noted.

4.2.3 A Cross-Validation Approach to Mixture Classification Quasi-Likelihood

One potential limitation of using an objective function based on the quasi-likelihood as a measure of model fit for longitudinal data is that the residuals used to compute the quasi-likelihood under the independence model are correlated with the scale parameter estimated using Pearson residuals. In particular, for normal data and a posterior probability of class membership of one, note that the feature- and class-specific quasi-likelihood under the independence model simplifies to

$$\begin{aligned} Q_{jg}\left(\hat{\boldsymbol{\beta}}_{jg}\left(\mathbf{R}\right),\phi_{jg}\right) &= \sum_{i=1}^{n}\sum_{k=1}^{m_{ij}} Q_{jg}\left(\hat{\boldsymbol{\beta}}_{jg}\left(\mathbf{R}\right),\phi_{jg};\left(Y_{ik},\mathbf{X}_{ik}\right)\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{m_{ij}} Q_{gijk}/\phi_{jg} \\ &= \frac{-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{m_{ij}}\left(y_{ijk}-\mu_{ijk}\right)^{2}}{\frac{\sum_{i=1}^{n}\sum_{k=1}^{m_{ij}}\left(y_{ijk}-\mu_{ijk}\right)^{2}}{n-p}} \\ &= -\frac{1}{2}\left(n-p\right). \end{aligned}$$

As this quantity is solely dependent on the number of subjects (n) and the number of parameters in the model of the mean structure (p), it does not provide a meaningful measure of model fit. A similar dependency exists between the quasi-likelihood and the scale parameter estimated using Pearson residuals for other distributions (e.g. Poisson).

In order to overcome the correlation between the quasi-likelihood and the estimated scale parameter, we propose using a 5×2 cross-validation approach[11] that allows the unknown scale parameter and the quasi-likelihood under the independence model to be estimated independently of one another. In 5×2 cross-validation, two-fold cross validation is run five times. The data is re-stratified after each run and the 10 resulting values are averaged to obtain a single estimate. Thus, to begin, the data

is divided into two approximately evenly sized samples, say sample A and sample B. Sample A is used to estimate the regression coefficients, $\boldsymbol{\beta}_{jgA}$, and the scale parameter, ϕ_{jgA} for the feature- and class-specific GEE. The feature- and class-specific quasi-likelihood contribution of the k^{th} observation on the i^{th} subject for sample B is then computed as

$$Q_{jgB}\left(\hat{\boldsymbol{\beta}}_{jgA}\left(\mathbf{R}\right),\phi_{jgA}\right).$$

Next, sample B is used to estimate the regression coefficients and scale parameter and then the feature- and class-specific quasi-likelihood contribution of the k^{th} observation on the i^{th} subject for sample A is computed under $\boldsymbol{\beta}_{jgB}$ and ϕ_{jgB}. The two-fold cross-validation process is repeated five times and the five values of $Q_{jgB}\left(\hat{\boldsymbol{\beta}}_{jgA}\left(\mathbf{R}\right),\phi_{jgA}\right)$ and the five values of $Q_{jgA}\left(\hat{\boldsymbol{\beta}}_{jgB}\left(\mathbf{R}\right),\phi_{jgB}\right)$ are averaged together to obtain Q_{jg}. Note that the cross-validation approach described would not be necessary for binary feature variables because the scale parameter is fixed at 1 and is, therefore, uncorrelated with the residuals. The approach described in the previous section for distinguishing between multiple roots and assessing the number of components can then be implemented using the cross-validated extended quasi-likelihood. Note that the cross-validated versions of the four information criteria proposed (BEQC, EQIC, CEQC, and CEQ-BIC) will be denoted $BEQC_{CV}$, $EQIC_{CV}$, $CEQC_{CV}$, and CEQ-BIC_{CV}, respectively.

4.3 Simulation Studies

Simulation studies will be used to compare the performance of the proposed diagnostic measures in detecting the true number of components for the latent trajectory model. One-, two-, and three-class model solutions were compared using each of the possible criteria.

In several instances, models that overfit the data were found to lead to numerical problems or divergent solutions. This is not surprising since cross-sectional finite mixture models often encounter identifiability and numerical issues when too many latent classes are assumed[34]. These numerical issues arise because the regularity conditions required for the asymptotic theory of maximum likelihood to apply are sometimes violated for small data sets, mixtures with small component weights, and overfitting mixtures with too many components[30]. Thus, starting values that led to a GEE that could not be successfully estimated without error were excluded from consideration in the following simulation studies. In addition, divergent solutions were not considered. A divergent solution was deemed to be any solution for which the absolute value of one or more of the parameter estimates associated with the polytomous logistic regression model or the GEE for one of the binary feature variables exceeded 10. Finally, any root that had not converged in 100 iterations was excluded. After the aforementioned roots were excluded from consideration, weak identifiability was assessed [30]. Specifically, when considering a C-component mixture model, any class for which the mixing proportion, $\pi_g (g = 1, \ldots, C)$, was less than 0.01 was deemed an empty class and treated as a $C - 1$ class solution. Similarly, if the maximum absolute difference between elements of $\boldsymbol{\beta}_g$ and $\boldsymbol{\beta}_{g'}$ for $g \neq g' (g, g' = 1, \ldots, C)$ was less than 0.01, the two classes were deemed to be equivalent and the root was treated as a $C - 1$ class solution.

4.3.1 Normally Distributed Feature Variables with Zero Slope

The first set of simulations was performed using the data from Section 3.4.1. Recall that, in this case, 10 distinct realizations of longitudinal data were generated with five normally distributed feature variables, as well as a binary covariate. It was assumed that each of 500 subjects had six measurement occasions and no missing data. Within each latent class, the features were assumed to be independent and

data for each feature was generated separately based on an autoregressive (AR1) correlation structure with a correlation coefficient of 0.3. All feature variables were generated with a standard deviation of 5. All slopes were taken to be 0. Table 3.1 provides a summary of the intercept/mean of each feature variable for each of two classes. Two scenarios were considered. In the first, the probability of belonging to both classes was equal. In the second, the mixing proportion for belonging to class 1 was 0.80 and the mixing proportion for belonging to class 2 was 0.20.

For each scenario, a one-, two-, and three-class finite mixture model was fit to the data. In order to avoid selecting a two- or three- class local solution, the models were estimated using the EM algorithm initialized using 100 random starting values. Multiple roots were compared using the extended quasi-likelihood under the independence model and, for a given number of latent classes ($C = 1, 2, 3$), the solution that maximized the objective function was chosen as the best C-class solution. Of the best one-, two-, and three-class solutions, the root that minimized the information criteria described above was then selected as the best root and its number of classes noted. Table 4.1 provides a summary of the number of classes selected by each of the proposed criteria when the data was generated under the assumption of two latent classes with equal mixing proportions. As shown, BEQC and EQIC consistently overestimate the number of classes. This is in agreement with cross-sectional simulation studies, which have suggested that BIC and AIC tend to overestimate the number of components [54]. In contrast, CEQC and CEQ-BIC correctly select a 2-class solution in all of the runs. Further, for normal data, using a cross-validated version of the extended quasi-likelihood seems to improve the performance of BEQC. This supports the idea that using cross-validation to estimate the extended quasi-likelihood under the independence model leads to an improved measure of model fit.

Table 4.1: Simulation results for selecting the appropriate number of latent classes based on normal data generated under the assumption of two latent classes with equal mixing proportions

Number of Classes	Information Criteria							
	CEQ-BIC	CEQC	EQIC	BEQC	CEQ-BIC_{CV}	$CEQC_{CV}$	$EQIC_{CV}$	$BEQC_{CV}$
1	0	0	0	0	0	0	0	0
2*	10	10	1	2	10	10	2	8
3	0	0	9	8	0	0	8	2

Next, consider the situation where the data was generated under the assumption of two latent classes with unequal mixing proportions. Table 4.2 provides a summary of the number of latent classes selected by each criteria. In this scenario, CEQ-BIC and CEQC consistently identify the correct number of components, while EQIC and BEQC again tend to over-estimate the number of components. Further, cross-validation of the extended quasi-likelihood again seems to improve the performance of BEQC.

Table 4.2: Simulation results for selecting the appropriate number of latent classes based on normal data generated under the assumption of two latent classes with unequal mixing proportions

Number of Classes	Information Criteria							
	CEQ-BIC	CEQC	EQIC	BEQC	CEQ-BIC_{CV}	$CEQC_{CV}$	$EQIC_{CV}$	$BEQC_{CV}$
1	0	0	0	0	0	0	0	0
2*	10	10	1	1	10	10	1	3
3	0	0	9	9	0	0	9	7

Finally, data was generated under the assumption that there was one latent class. Again, 10 distinct realizations of longitudinal data were generated with five normally distributed feature variables, as well as a binary covariate. It was assumed that each of 500 subjects had six measurement occasions and no missing data. Within each latent class, the features were assumed to be independent and data for each feature was generated separately based on an autoregressive (AR1) correlation structure with a correlation coefficient of 0.3. All feature variables were generated with a standard deviation of 5. All slopes were taken to be 0. Table 4.3 presents the class-specific intercepts of the feature variables.

Table 4.3: Intercepts of five normally distributed feature variables simulated under an AR(1) correlation structure with a slope of 0, a correlation coefficient of 0.3, and a standard deviation of 5.

Feature	Intercept
Feature 1	5
Feature 2	25
Feature 3	20
Feature 4	30
Feature 5	15

Table 4.4 provides a summary of the number of times that each of the proposed criteria selected the correct number of classes when the true number of latent classes was one. EQIC and BEQC, as well as their cross-validated versions, over-estimate the true number of latent classes. In contrast, CEQ-BIC, CEQC, CEQ-BIC_{CV}, and $CEQC_{CV}$ correctly determined that there was one latent class in all 10 runs.

Table 4.4: Simulation results for selecting the appropriate number of latent classes based on normal data generated under the assumption of one latent class

Number of Classes	Information Criteria							
	CEQ-BIC	CEQC	EQIC	BEQC	CEQ-BIC_{CV}	$CEQC_{CV}$	$EQIC_{CV}$	$BEQC_{CV}$
1*	10	10	0	0	10	10	0	1
2	0	0	1	1	0	0	1	1
3	0	0	9	9	0	0	9	8

4.3.2 Discrete Feature Variables with Non-zero Slope

The second set of simulations was performed using data from Section 3.4.2. Recall that, in this case, 10 distinct realizations of longitudinal data were generated for each of 500 subjects. Each subject had six measurement occasions and no missing data. The data was composed of 2 normal, 2 binary, and 2 Poisson feature variables, as well as a binary covariate. In addition, the second simulation incorporated a non-zero slope with respect to time for a subset of the feature variables. Within each latent class, the features were assumed to be independent and data for each feature was generated separately based on an autoregressive (AR1) correlation structure with a correlation coefficient of 0.3. All feature variables were generated with a standard deviation of 5. Table 3.5 provides a summary of the intercept and slope of each feature variable for each of two classes. As was the case in the simulations based on normal feature variables, two scenarios were considered. In the first, the probability of belonging to both classes was equal. In the second, the mixing proportion for belonging to class 1 was 0.80 and the mixing proportion for belonging to class 2 was 0.20.

For each scenario, a one-, two-, and three-class finite mixture model was fit to the data. In order to avoid selecting a local solution, 100 random starting values were used for all two- and three-class models. Again, for a fixed number of latent classes, multiple roots were compared and the root that maximized the extended quasi-likelihood under the independence model was selected. Then, of these roots, the solution that minimized the information criteria being considered was selected as the best root and its corresponding number of classes noted. Table 4.5 provides a summary of the number of times that each of the proposed criteria selected a root with the correct number of classes when the mixing proportions were equal. The results suggest that CEQ-BIC performs slightly better than CEQC, EQIC, and BEQC; however, it still only correctly identifies a two-class solution in 60% of the runs.

In this simulation, there do not seem to be any improvements in the performance of the information criteria when 5×2 cross-validation is used to estimate the extended quasi-likelihood function.

Table 4.5: Simulation results for selecting the appropriate number of latent classes based on discrete and normal data generated under the assumption of two latent classes with equal mixing proportions

Number of Classes	Information Criteria							
	CEQ-BIC	CEQC	EQIC	BEQC	CEQ-BIC_{CV}	$CEQC_{CV}$	$EQIC_{CV}$	$BEQC_{CV}$
1	0	0	0	0	0	0	0	0
2*	6	5	4	4	5	5	5	5
3	4	5	6	6	5	5	5	5

Next, consider the situation in which the true number of latent classes was two and the mixing proportions were unequal with an 80-20 split. Table 4.6 provides a summary of the number of times that each of the proposed criteria selected the correct number of classes. Here, CEQ-BIC and BEQC more often select the true number of latent classes than CEQC and EQIC. The weaker performance of CEQC in this scenario is consistent with the cross-sectional literature. Specifically, Biernacki, Celeux, and Govaert(1999) [4] noted that CLC works well when the mixing proportions are restricted to being equal, but otherwise tends to over-estimate the number of components. While there is not an improvement in the performance of CEQ-BIC and BEQC, the cross-validation approach did seem to improve the performance of CEQC and EQIC.

Table 4.6: Simulation results for selecting the appropriate number of latent classes based on discrete and normal data generated under the assumption of two latent classes with unequal mixing proportions

Number of Classes	Information Criteria							
	CEQ-BIC	CEQC	EQIC	BEQC	CEQ-BIC_{CV}	$CEQC_{CV}$	$EQIC_{CV}$	$BEQC_{CV}$
1	0	0	0	0	0	0	0	0
2*	8	6	6	8	8	8	8	8
3	2	4	4	2	2	2	2	2

Finally, data was generated under the assumption that there was only one latent class. Again, 10 distinct realizations of longitudinal data were generated for six feature variables, as well as a binary covariate. It was assumed that each of 500 subjects had six measurement occasions and no missing data. Within each latent class, the features were assumed to be independent and data for each feature was generated separately based on an autoregressive (AR1) correlation structure with a correlation coefficient of 0.3. All feature variables were generated with a standard deviation of 5. Table 4.7 presents the class-specific intercepts of the feature variables.

Table 4.7: Intercepts and slopes of six feature variables simulated under an AR(1) correlation structure with a correlation coefficient of 0.3.

Feature	Distribution	Intercept	Slope
Feature 1	Poisson	0.7	1.0
Feature 2	Poisson	3.0	0.0
Feature 3	Binary	0.5	-1.0
Feature 4	Binary	-0.5	0.0
Feature 5	Normal	20.0	0.0
Feature 6	Normal	5.0	5.0

Table 4.8 provides a summary of the number of times that each of the proposed criteria selected the correct number of classes when the true number of latent classes was one. Here, all of the information criteria selected the true number of latent classes in 60% of the runs and there was no improvement when cross-validation was used.

Table 4.8: Simulation results for selecting the appropriate number of latent classes based on discrete and normal data generated under the assumption of one latent class

Number of Classes	Information Criteria							
	CEQ-BIC	CEQC	EQIC	BEQC	CEQ-BIC_{CV}	$CEQC_{CV}$	$EQIC_{CV}$	$BEQC_{CV}$
1*	6	6	6	6	6	6	6	6
2	3	3	3	3	3	3	3	3
3	1	1	1	1	1	1	1	1

4.4 Discussion

In this chapter, several diagnostic procedures for comparing multiple roots and assessing the number of components in a finite mixture model were proposed. Based on the simulation, CEQ-BIC_{CV} seems to most often select a solution with the correct number of latent classes. With that said, for discrete data, it sometimes tends to select too many latent classes. This may be because the information criteria considered rely on the extended quasi-likelihood function as a measure of model fit. The quasi-likelihood function differs from the likelihood function in that it is strictly for interior solutions. When the model is fit with too many latent classes, divergent solutions on the boundary of the parameter space are often observed and quasi-likelihood may actually reward such divergent solutions. Future research into methods for assessing the number of components, particularly when discrete feature variables are incorporated, is warranted. For example, one might consider using the projected deviance, which is more stable on the boundary of the parameter space, as an objective function to assess model fit.

Chapter 5

Identifying Subtypes of Mild Cognitive Impairment via a Latent Trajectory Model

5.1 Overview

Mild cognitive impairment (MCI) refers to an intermediate stage between normal aging and dementia [24]. Research has suggested that there is tremendous heterogeneity in the clinical presentation of MCI. As a result, the classification system for MCI encompasses several MCI subtypes based on the number and type of cognitive domains affected. More specifically, National Institutes of Health (NIH)-supported Alzheimer's disease centers classify MCI patients into the following four subtypes: 1.) Amnestic MCI-memory impairment only; 2.) Multidomain MCI-Amnestic (memory plus one or more nonmemory domains); 3.) Multidomain MCI-Non-Amnestic (more than one nonmemory domain); or 4.) Single Nonmemory MCI (one nonmemory domain). In this section, the latent trajectory methodology proposed will be used to statistically validate the presence of longitudinal MCI subtypes and to model the

progression of MCI within these distinct MCI subgroups.

The four MCI subtypes described were developed solely based on clinical observation rather than on a rigorous clustering approach. In their recent exploration of MCI subtypes, Hanfelt et al.(2011) [18] used latent class analysis to analyze cognitive, neuropsychiatric, and functional features of MCI patients. The results of their statistical analysis suggested that there are actually 7 subtypes of MCI: 1.) Minimally impaired (cognitive function indistinguishable from the cognitively normal group); 2.) Amnestic Only (subtle impairment in delayed memory); 3.) Amnestic with Functional Impairments & Neuropsychological Features (impairments in both immediate and delayed memory, difficulties performing instrumental activities of daily living (IADL); 4.) Amnestic Multidomain (impairments across cognitive domains, including episodic and semantic memory, language, and executive functioning); 5.) Amnestic Multidomain with Functional Impairment & Neuropsychological Features (impairments across a broader spectrum of cognitive domains than Amnestic Multidomain, including attention and visuomotor skills, as well as difficulties performing IADL, neuropsychiatric disturbances); 6.) Functional Impairments & Neuropsychological Features (functional and behavioral impairments with no cognitive impairment detected); and 7.) Executive Function & Language Impairments (impairments in nonmemory domain). These results support the notion that MCI is a heterogenous disorder and suggest the need for further investigation into the number and conceptualization of MCI subtypes. In particular, studying the progression of MCI over time may help researchers gain insight into the etiology of MCI, its subtypes, and its eventual outcomes.

5.2 National Alzheimer's Coordinating Center- Uniform Data Set

Data was obtained from the Uniform Data Set (UDS), a standardized assessment and data protocol maintained by the National Alzheimer's Coordination Center with 29 participating NIH-supported Alzheimer's disease centers nationwide [2, 25]. When conceptualizing MCI subtypes, 13 clinical features were considered. The mini-mental state exam (MMSE)[32] was used as a measure of overall cognitive status. In addition, measures were used to assess the following specific cognitive domains: executive function (Trail-Making Test[83]); language (Boston Naming Test[40]; category fluency[75]); attention (Digit Span and Digit Symbol subtests[84]); and episodic memory (Logical Memory, Story A[85]). Research has also indicated that neuropsychiatric and functional features unrelated to cognition may provide additional information about MCI subtypes [28, 57, 74, 76, 26, 20, 21]. Thus, the Functional Assessment Questionnaire (FAQ)[23], which measures dependence-performing IADL over the previous 4 weeks, was used to evaluate functional abilities. The FAQ assesses whether the participant has the ability to balance one's checkbook and write/pay bills, assemble tax records and other financial papers, shop alone, play complicated games/maintain a hobby, perform simple kitchen-related tasks such as heating water and turning off the stove, prepare a complicated meal, pay attention to/follow information such as a television program, and remember events and tasks such as to take medication. A count of how many of the 10 activities each participant was rated as having difficulty/needing assistance with was recorded. In addition, a count of the number of items indicated on the Neuropsychiatric Inventory Questionnaire (NPIQ)[15] was recorded for each patient. The NPI-Q evaluates problematic behavioral changes in the last month by assessing 12 behaviors including delusions, hallucinations, agitation/agression, depression/dysphoria, apathy/indifference, elation/euphoria, anxiety,

disinhibition, irritability/lability, aberrant motor behavior, nighttime behaviors, and appetite/eating. Finally, each participant was given the Geriatric Depression Scale (GDS)[73], which is a self-report on depressive symptoms. The GDS was dichotomized as 0-4 points (no depression) or 5-15 points (depression). A secondary interest in this analysis was to examine whether there was an association between cerebrovascular disease (CVD) and MCI subgroups. Thus, the Rosen Modification of Hachinksi Ischemic Score (RMHIS) [27] was used as marker for probable CVD and considered as a potential risk factor for MCI subtype classification.

Inclusion criteria required that participants had a consensus diagnosis of MCI at baseline, nonmissing information on age, years of education, and race (dichotomized as white or nonwhite), and a mini-mental state exam (MMSE) score of 22 or greater at baseline. In addition, only patients with at least two measurement occasions and nonmissing information on each neuropsychological feature on at least one of those occasions were considered. If patients had a third measurement occasion available, it was retained for analysis. Measurement occasions occurred at approximately 1 year intervals.

The neuropsychological, functional, and neuropsychiatric features described above for the 2,348 participants who met these inclusion criteria were entered into the proposed latent trajectory model. The raw cognitive test scores were standardized using baseline age, race, and education level of 5,542 cognitively normal patients from the UDS. Cognitively normal patients included in the reference group were required to have information on the necessary demographics and to have attained an MMSE score of 25 points or higher at baseline. A linear regression model with age, race, education, and the interaction between race and education as explanatory variables was fit for each of the 10 neuropsychological measures. This means that, for example, a standardized test score of -1.5 would be indicative of the fact that the MCI participant's score was 1.5 standard deviations (SDs) lower than the mean

among UDS cognitively normal subjects of the same age, education level, and race[18]. In addition, the relationship between the MCI latent classes and the RMHIS was modeled using polytomous logistic regression.

MCI participant's demographic and clinical characteristics at baseline are given in Table 5.1.

5.3 A Latent Trajectory Model for Mild Cognitive Impairment

In order to determine the appropriate number of latent classes for explaining the variability in the 13 feature variables, latent trajectory models were fit with between 1 and 6 classes. All models were fit under the fundamental assumption of local independence. An AR(1) working correlation assumption was assumed for each feature variable. In an effort to avoid local solutions, 100 random starting values were used to initialize the the EM algorithm for each model. As in the simulation studies, the stopping criteria was taken to be 100 iterations or an absolute difference in parameter estimates between the current and previous iteration of at most 1% for any parameter. Only solutions that had converged in 100 iterations were considered. For a given number of latent classes, the root which maximized the mixture extended quasi-likelihood function was selected. Of these roots, the solution that minimized CEQ-BIC_{CV} was selected as the best solution and its number of components noted.

The results suggested that a model with 2 latent classes was the most parsimonious model for explaining the heterogeneity in the observed cognitive, functional, and neuropsychiatric measures. The chosen model with 2 latent classes seemed to clearly differentiate the study population based on rate of cognitive decline. Further investigation into the model revealed that the larger of the two classes was comprised

of participants who experienced rapid cognitive decline. The average probability of an MCI patient belonging to this declining subtype was approximately 94%. The results also suggested an association between the covariate RMHIS and the empirically derived subgroups. Specifically, individuals with probable CVD, as indicated by a RMHIS score of greater than or equal to 4, were more likely to belong to the declining MCI subtype (OR = 4.816). In contrast, the smaller of the two classes corresponded to participants who displayed more stable cognitive functioning over time.

Table 5.2 presents the intercepts and slopes with respect to time for each of the cognitive, functional, and neuropsychiatric measures considered by MCI subtype (stable or decline), while Figure 5.1 shows the longitudinal trajectories for each measure by MCI subtype. In general, the declining MCI subtype experienced cognitive problems across all cognitive domains. In addition, the declining subtype consistently showed greater cognitive impairment than the stable MCI subtype at the baseline measurement occasion. At baseline, patients in the stable MCI subtype had an average MMSE score 0.194 standard deviations (SDs) higher than the mean MMSE among UDS cognitively normal subjects of the same age, education level, and race (controls). In contrast, the average MMSE score among the declining MCI subtype was 1.434 SDs lower than the mean MMSE among controls and the average MMSE score continued to decrease by 0.494 SDs for every 1 year of follow-up.

Next, differences between the declining and stable MCI subtypes by cognitive domain were considered. First, the logical and semantic memory domains were examined in greater detail. When considering logical memory, the mean score on the delayed recall assessment (Story A) in the declining subtype was 1.311 SDs lower at baseline than the mean score among controls; however, for every one year of follow-up, the average score decreased by only 0.039 SDs. In contrast, the average delayed recall score at baseline among the stable subtype was consistent with that of con-

trols (0.048 SDs higher) and did not show evidence of decline over time. Similar results were observed for the immediate recall assessment (Logical Memory). Semantic memory was assessed by the category fluency test. The stable MCI subtype had an average category fluency score at baseline that was consistent with controls(0.085 SDs higher), while the average baseline category fluency score in the declining MCI subtype was 0.925 SDs lower than controls. On average, the category fluency score for patients in the declining MCI subtype decreased by an additional 0.163 SDs for each year of follow-up. The stable MCI subtype did not show evidence of declining category fluency scores over the follow-up period.

The declining MCI subtype also showed greater cognitive impairment in the attention domain than the stable MCI subtype. Recall that positive values are indicative of greater cognitive impairment when considering the Trails A assessment. On average, the declining MCI subtype had a baseline Trails A score 0.626 SDs higher than controls and an increase in score of 0.197 SDs per year of follow-up. The Trails A and Digit Span Forward assessment did not indicate any cognitive impairment or decline in the attention domain for the stable MCI subtype.

Similar trends were observed in the language domain. More specifically, the stable MCI subtype did not show evidence of cognitive impairment at baseline or follow-up. On the other hand, the declining MCI subtype had an average baseline score on the Boston Naming Test 0.937 SDs lower than controls and the score continued to decrease, on average, by 0.176 SDs per year of follow-up. This suggests that patients in the declining MCI subtype experienced cognitive impairment in the language domain.

Finally, the declining MCI subtype also showed impairment in the executive functioning and visuomotor domains. Again, recall that positive values are indicative of greater cognitive impairment when considering the Trails B assessment. Here, the declining subtype had an average baseline Trails B score that 1.250 SDs higher than controls and the average score increased by 0.238 SDS per year. When considering

the visuomotor domain, the declining MCI subtype had an average baseline score on the Digit Symbol test that was 0.868 SDs lower than controls and the average score decreased by 0.140 SDs per year of follow-up. The stable subtype did not show evidence of cognitive impairment at baseline or follow-up in the executive functioning and visuomotor domains.

In addition to cognitive assessments, functional and neuropsychiatric assessments were considered. Patients in the declining subtype tended to have a greater odds of depression than patients in the stable subtype, as indicated by a GDS score of 5 or greater. At baseline, the odds of a patient in the declining MCI subtype experiencing depression were 0.179. For each year of follow-up, the odds of depression in the declining subtype increased by a factor of 1.069. In comparison, the odds of a patient in the stable MCI subtype experiencing depression at baseline were 0.053 and the odds increased by a factor of 0.752 per year of follow-up. In addition, it appears that patients in the declining subtype noted more problematic behavioral changes on the NPI-Q. On average, patients in the declining MCI subtype reported 1.661 problematic behavioral changes (out of 12) on the NPI-Q at baseline and the number of problematic behavioral changes reported increased by a factor of 1.096 per year of follow-up. The stable subtype reported 0.902 problematic behavioral changes at baseline and did not show evidence of increased problems over follow-up. Finally, the Functional Assessment Questionnaire (FAQ) was used to evaluate functional abilities. At baseline, the average number of tasks that the participant rated as having difficulty/needing assistance with was 2.411 among the declining MCI subtype and 0.992 among the stable MCI subtype. The stable subtype did not show evidence of decreased functional ability over time, while the average number of items rated as having difficulty/needing assistant increase by a factor of 1.285 per year among the declining subtype.

In summary, the analysis indicated that longitudinal MCI subtypes can primarily

be differentiated by the rate of cognitive decline. The latent class trajectory model indicated the presence of a stable and declining MCI subtype. Differences on the cognitive, functional, and neuropsychiatric assessments between the subtypes were observed at baseline and follow-up. Finally, the declining subgroup showed cognitive impairment across all cognitive domains.

Table 5.1: Baseline demographic and clinical characteristics of 2,348 MCI participants from the uniform data set

	$Mean \pm SD$ or n(%)
Demographic	
Age	74.3 ± 8.6
Sex: male	1168(50)
Race	
White	2003(85)
Black/African American	308(13)
Asian	37(2)
Hispanic	120(5)
Education, years	15.2 ± 3.2
MMSE	27.5 ± 2.0
Functional	
No. of IADL rated as difficult to perform/requiring assistance (10 maximum)	2.3 ± 2.6
Neuropsychiatric	
GDS ≥ 5	333(14)
No. of NPI-Q items rated as present	1.6 ± 1.9
No. of persons with an NPI-Q symptom present	
Depressed	671(29)
Irritable	619(27)
Nighttime behavior	510(22)
Anxious	486(21)
Agitated	356(16)
Apathetic	424(19)
Change in appetite	256(11)
Disinhibited	200(9)
Repetitive activities	105(5)
Euphoric	49(2)
Delusions	71(3)
Hallucinations	31(1)
Cognitive (standardized scores[a])	
MMSE	-1.3 ± 1.8
Logical memory: immediate	-1.0 ± 1.1
Logical memory: delayed	-1.2 ± 1.2
Semantic memory: category fluency	-0.9 ± 0.9
Attention: Trails A[b]	0.5 ± 1.5
Attention: Digit Span Forward	-0.3 ± 1.0
Language: Boston Naming	-0.9 ± 1.7
Executive function: Trails B[b]	1.1 ± 1.7
Executive function: Digit Span Backward	-0.4 ± 1.0
Visuomotor: Digit Symbol	-0.8 ± 1.1
Risk factors	
RMHIS ≥ 4	135(6)

Notes: Number of subjects for whom data were unavailable: GDS, N = 36; Logical memory: immediate, N = 60; Logical memory: delayed, N = 56; Category Fluency, N=36; Trails A, N = 17; Digit Span Forward, N = 24; Boston Naming, N= 35; Trails B, N = 44; Digit Span Backward, N = 24; Digit Symbol, N = 99

[a] All cognitive test scores were converted to age-, education-, and race-adjusted z scores.

[b] Positive values of Trail-Making Tests A and B indicate greater cognitive impairment.

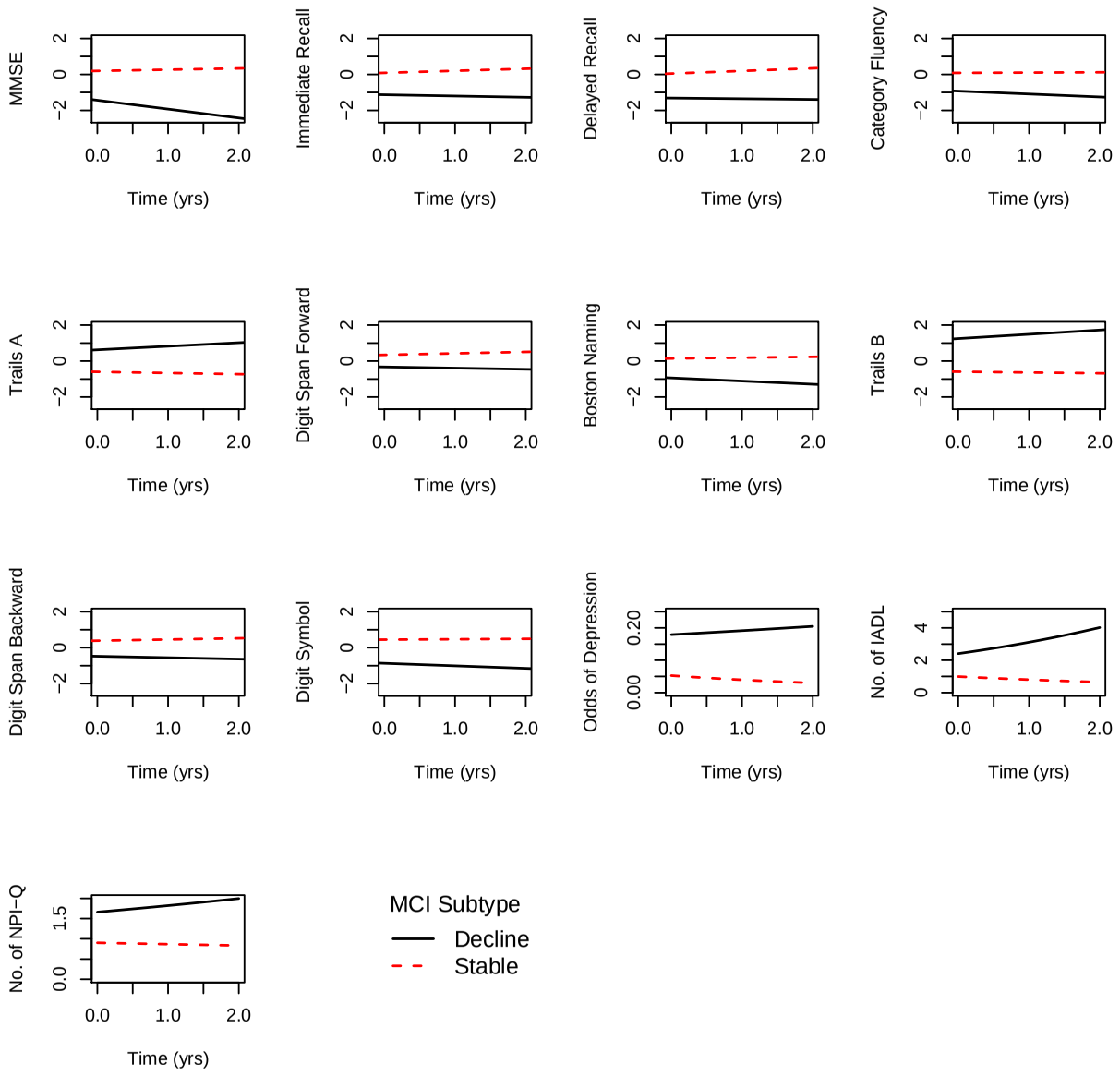
Table 5.2: Parameter estimates associated with cognitive, functional, and neuropsychiatric assessments for the two-class latent trajectory model based on 2,348 MCI patients from the uniform data set

Feature	MCI Subtype	Intercept Estimate	Slope Estimate
MMSE	Stable	0.194	0.069
	Decline	-1.434	-0.494
Logical memory: immediate	Stable	0.087	0.114
	Decline	-1.130	-0.070
Logical memory: delayed	Stable	0.048	0.145
	Decline	-1.311	-0.039
Semantic memory: category fluency	Stable	0.085	0.017
	Decline	-0.925	-0.163
Attention: Trails A[b]	Stable	-0.601	-0.064
	Decline	0.626	0.197
Attention: Digit Span Forward	Stable	0.346	0.082
	Decline	-0.325	-0.066
Language: Boston Naming	Stable	0.141	0.047
	Decline	-0.937	-0.176
Executive function: Trails B[b]	Stable	-0.593	-0.043
	Decline	1.250	0.238
Executive function: Digit Span Backward	Stable	0.389	0.067
	Decline	-0.476	-0.079
Visuomotor: Digit Symbol	Stable	0.449	0.021
	Decline	-0.868	-0.140
Depressed (GDS ≥ 5)	Stable	-2.944	-0.285
	Decline	-1.719	0.067
No. of IADL	Stable	-0.008	-0.212
	Decline	0.880	0.251
No. of NPI-Q items	Stable	-0.103	-0.038
	Decline	0.508	0.092
Polytomous Logistic Regression [a]		-2.651	-1.572

[a] The polytomous logistic regression model compares the probability of belonging to the stable MCI subtype to the probability of belonging to the decline MCI subtype. The slope is the slope with respect to the dichotomized Modified Hachinski Score.

[b] Positive values of Trail-Making Tests A and B indicate greater cognitive impairment.

Figure 5.1: Latent class trajectories associated with cognitive, functional, and neuropsychiatric assessments for stable and declining MCI subtypes based on 2,348 MCI patients from the uniform data set



5.4 Discussion

The results in the previous section support the notion that MCI is a heterogenous disorder and the association observed between the covariate RMHIS and the empirically derived subgroups of MCI is in alignment with cross-sectional results [18]. However, it was somewhat unexpected that only two subtypes were identified and that these subtypes differed primarily in the rate of decline rather than in any systematic differences in the cognitive domains affected. This is in contrast to cross-sectional results that suggest that subtypes are based on the cognitive domains affected and the degree of functional impairment. The findings from this analysis are exploratory and additional research is needed to verify the validity of the proposed model.

It should be noted that this analysis relied on data from the NACC-UDS. While the UDS represents 29 sites nationwide and has uniform definitions for each study variable, it is not a community-based sample. As Hanfelt et al. comment [18], the definition of subtypes and prevalence estimates of MCI are heavily dependent on the sample chosen (e.g. memory clinic versus population based) because UDS subjects often are motivated to participate in research based on their concerns of a family history of dementia and are not fully representative of the community. Moreover, the findings are dependent on the measures available in the national database and the inclusion of different or additional measures may have identified an alternate solution. In addition, the inclusion criteria and treatment of missing data need to be considered in greater detail. It is possible that patients with only one visit differ systematically from patients who return for a follow-up visit and that excluding these patients may have introduced bias into the analysis. Additionally, patients who received a consensus diagnosis of MCI at the baseline measurement occasion may not have maintained an MCI diagnosis for the duration of follow-up.

Future methodological research is also needed to verify the validity of the current model. First, the implementation of standard error estimation methods will allow

for statistical significance to be assessed. In addition, studies of power and sample size to determine how many subjects and measurement occasions are required to simultaneously identify discrepancies in both slopes and intercepts between subtypes are warranted. Finally, it is possible that the proposed model may be greatly influenced by the covariate used in the polytomous logistic regression model for the mixing proportions, e.g. RMHIS. In latent class analysis, there are two general approaches for handling these types of covariates. Namely, one can consider the "active covariates method" or the "inactive covariates method"[78]. The proposed methodology relies on the "active covariates method", which incorporates the covariates into a polytomous logistic regression model. In contrast, the "inactive covariates method" involves computing descriptive measures for the association between covariates and the class membership probabilities after estimating the model without covariates. In likelihood-based analysis, the decision to treat a covariate as active rather than inactive can impact the estimated model; however, the extent to which active covariates influence solutions to the proposed longitudinal latent class methodology has yet to be investigated.

Chapter 6

Summary and Future Research

6.1 Summary

This dissertation research focused on developing a robust extension of latent class methods for high-dimensional longitudinal data. Although the literature includes in depth discussions of latent class analysis and generalized estimating equations (GEEs) separately, an approach combining these topics had not yet been considered. As such, an innovative extension of latent class methods based on weighted GEEs was proposed and evaluated via simulation studies. The proposed approach can be used to model latent trajectories and select the appropriate number of latent classes.

6.2 Future Research

Due to the novel nature of this approach, there are several possible areas for future work. A subset of these areas is briefly described in the subsections below. Future work also includes: asymptotic standard error estimation for the parameters of the latent trajectory model; small dispersion asymptotic theory for the likelihood ratio approximations; a generalized method of moments approach similar to the Reboussin(2002) [70] approach to compare with the proposed approach; tests of uni-

dimensional ordered latent classes; further investigation into the use of weights to accommodate missing data when the MCAR assumption is violated; model selection approaches for determining which covariates to include in the model; additional consideration of the impact of active covariates on the estimated model; and, studies of power and sample size.

6.2.1 Empirical Likelihood

Recall that, when estimating a finite mixture of GEEs, the posterior probability of class membership is given by

$$\tau_{ig} = \frac{\pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})LR_{ig}(\boldsymbol{\theta})}{\sum_{d=1}^{C}\pi_d(\boldsymbol{x}_i;\boldsymbol{\alpha})LR_{id}(\boldsymbol{\theta})}.$$

In Chapter 3, the projection-based approach of Li(1993) [44] was used to approximate the subject-specific likelihood ratios comparing the probability that a given subject is in component $g(g = 2,\ldots,C)$ as compared to component 1. As proposed in chapter 3.2, an alternate approximation of the likelihood ratio can be determined based on a novel use of empirical likelihood[68]. Thus, future research will include estimating the finite mixture model using the empirical likelihood-based approach to approximate the likelihood ratio and comparing its performance with estimation using the projection-based approach.

6.2.2 Model Formulation

In formulating the GEE model for latent trajectories, there are three possible extensions to consider. First, one might consider nonlinear effects of time and non-temporal covariates in the mean regression model. Second, it might be useful to incorporate links other than the identity link for the scale and correlation components of the GEE [64]. Standard software packages presently offer the option of non-identity links for

the mean component of the GEE only; however, a log link is often used for the scale component and a link based on Fisher's z-transformation seems natural for the correlation component. These links would ensure that the scale is positive and that the correlation is within the range $(-1, 1)$. Finally, more complicated working correlation models can be considered. The proposed approach relies on an AR1 correlation structure, which assumes evenly spaced measurements over time. In practice, this assumption may not be valid for many applications. Thus, more flexible correlation structures (e.g. a Markov or Generalized Markov) with the ability to model unevenly spaced longitudinal measurements can be considered provided parameter estimation remains feasible.

From a clinical perspective, it is also of interest to incorporate a binary indicator of whether a patient ultimately progresses to clinically probable Alzheimer's disease (AD) into the model. Muthen and Shedden(1999) [58], as well as Proust-Lima et al.(2007) [65], incorporated a binary indicator of disease in the context of fully-parametric mixed models. For the motivating example on mild cognitive impairment (MCI), a similar extension to the proposed methodology might offer insight into the relationship between MCI subtype and long-term prognoses for progression to Alzheimer's disease. Additionally, extending this notion to incorporate multinomial clinical outcomes and competing risks would allow simultaneous investigation of the relationship between MCI and other dementia types such as frontotemporal dementia, Lewy body dementia, and vascular dementia.

6.2.3 Model Diagnostics

Even in the cross-sectional context, there is no consensus regarding the most effective approach for selecting the number of components in a finite mixture model. In the absence of a likelihood function, selecting the number of components for a finite mixture of GEEs becomes even more challenging. Although the preliminary simulation results

reported in Chapter 4 suggest that a mixture classification quasi-likelihood approach works well for normal feature variables, additional investigation into model diagnostics when there are discrete feature variables is needed. In particular, Kolaczyk(1995) [42] proposed a longitudinal information criteria known as the Empirical Information Criteria (EIC), where the objective function for assessing lack of fit was taken to be the empirical likelihood function. Thus, it is of interest to compare the performance of the information criteria described in Chapter 4 with equivalent criteria based on empirical likelihood. In addition to considering an objective function other than the quasi-likelihood under the independence model as a measure of model fit, one might also consider different penalty terms. Finally, unsupervised learning approaches for determining the number of components may be considered [30].

6.2.4 Improvements in Computational Efficiency and Numerical Issues

In the likelihood-based context, it is well known that estimation of finite mixture models can be computationally complex and lead to numerical issues. These types of challenges also arise in the proposed longitudinal extension of latent class analysis. One of the main computational issues arises due to potential for multiple roots. In the current work, multiple random starting values were used to reduce the possibility of selecting a local solution; however, this inevitably increased the computational burden of the approach. In the likelihood based context, Finch et al.(1989) [31] investigated probabilistic measures of adequacy of a numerical search for a global maximum. In other words, they estimated the probability that an iterative algorithm using a randomly selected starting point would find a solution not observed in previous random starting points. Extending this research and performing simulation studies in the current longitudinal context may help to determine the number of random starting values actually necessary to ensure convergence to a global maximum with an accept-

able level of certainty. Additionally, alternate methods for determining starting values when estimating a cross-sectional finite mixture model using the EM algorithm have been proposed. For example, one might consider initializing the algorithm by assigning subjects to classes based on a clustering algorithm, such as k-means [54]. Further, standard software packages that employ likelihood-based estimation for latent class analysis sometimes run all random starting values out for a pre-specified number of iterations and then perform additional iterations on a subset of the solutions deemed to be be the best based on evaluation of the log-likelihood function. Although this type of procedure increases computational efficiency, it does not guarantee that the global solution will be found [78]. Further, extending this type of approach to the longitudinal context may prove challenging since the quasi-likelihood function does not behave well on the boundary of the parameter space. For this reason, alternate criteria for distinguishing between multiple roots may be considered. In addition, investigation into approaches for speeding up the convergence of the EM algorithm in the absence of a likelihood function may be useful. Finally, recall that cross-sectional finite mixture models often encounter identifiability and numerical issues because the regularity conditions required for the asymptotic theory of maximum likelihood to apply are sometimes violated for small data sets, mixtures with small component weights, and overfitting mixtures with too many components [33]. Similar issues appear to arise in the proposed longitudinal extension and research into ways to address and minimize these issues would be a valuable addition to this research area.

6.2.5 Local Dependence

The key assumption underlying latent class analysis is local or conditional independence. When two or more of the feature variables measure on closely related traits, the assumption of local independence may not be satisfied. The presence of local dependence among the feature variables often results in increased lack of fit [78].

The usual way to proceed is then to increase the number of latent classes in order to improve model fit; however, this may actually lead to overfitting and result in spurious latent classes. Although Torrance-Rynard and Walter (1997)[77] note that a latent class model may supply parameter estimates reasonably close to the "true" values even when local dependence is present, this is not guaranteed. As such, in the cross-sectional context, methods have been developed to detect local dependency and, if necessary, to relax the assumption of conditional independence for cross-sectional latent analysis (see, for example, [35, 36, 78, 50]). In the future, methods should be developed to accommodate local dependence among subsets of the feature variables in the proposed longitudinal context.

Bibliography

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Transaction on Automatic Control **19** (1974), no. 6, 716–23.

[2] D.L. Beekly, E.M. Ramos, and W.W. Lee et al., *The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set*, Alzheimer Disease & Associated Disorders **21** (2007), no. 3, 249–258.

[3] C. Biernacki, G. Celeux, and G. Govaert, *Asessing a mixture model for clustering with the integrated classification likelihood*, Tech. Report 3521, Rhone-Alpes:INRIA, 1995.

[4] ———, *An improvement of the NEC criterion for assesing the number of clusters in a mixture model*, Pattern Recognition Letters **20** (1999), no. 3, 267–272.

[5] C. Biernacki and G. Govaert, *Using the classification likelihood to choose the number of clusters*, Computing Science and Statistics **29** (1997), 451–457.

[6] H. Chung, S.T. Lanza, and E. Loken, *Latent transition analysis: Inference and estimation*, Statistics in Medicine **27** (2008), no. 11, 1834–1854.

[7] H. Chung, Y. Park, and S.T. Lanza, *Latent transition analysis with covariates: Pubertal timing and substance use behaviours in adolescent females*, Statistics in Medicine **24** (2005), no. 18, 2895–2910.

[8] S.L. Crawford, *An application of the laplace method to finite mixture distributions*, Journal of the American Statistical Association **89** (1994), no. 425, 259–267.

[9] D. Dalthorp and L. Madsen, *User's guide to discsim 2.1*, `http://oregonstate.edu/dept/statistics/epa_program/user2p1.pdf`.

[10] ———, *Generating correlated count data*, Environmental and Ecological Statistics **14** (2007), 129–148.

[11] T.G. Dietterich, *Approximate statistical tests for comparing supervised classification learning algorithms*, Neural Computation **10** (1998), no. 7, 1895–1923.

[12] P.J. Diggle, K.Y. Liang, and S.L. Zeger, *Analysis of Longitudinal Data*, Clarendon Press, Oxford, 1994.

[13] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, Boca Raton, Florida, 1994.

[14] B. Winblad et al., *Mild cognitive impairment- beyond controversies, towards a consensus: Report of the international working group on mild cognitive impairment*, Journal of Internal Medicine **256** (2004), no. 3, 240–246.

[15] D.I. Kaufer et al., *Validation of the NPI-Q, a brief clinical form of the Neuropsychiatric Inventory*, Journal of Neuropsychiatry **12** (2000), 233–239.

[16] E. Dantan et al., *Pattern mixture models and latent class models for the analysis of multivariate longitudinal data with informative dropouts*, The International Journal of Biostatistics **4** (2008), no. 1.

[17] G. Fitzmaurice et al. (ed.), *Longitudinal Data Analysis*, Chapman & Hall, Boca Raton, Florida, 2009.

[18] J.J. Hanfelt et al., *An exploration of subgroups of mild cognitive impairment based on cognitive, neuropsychiatric, and functional features: Analysis of data from the National Alzheimer's Coordinating Center*, American Journal of Geriatric Psychiatry ((to appear) 2011).

[19] K. Bandeen-Roche et al., *Latent variable regression for multiple discrete outcomes*, Journal of the American Statistical Association **92** (1997), no. 440, 1375–1386.

[20] K.R. Kim et al., *Characteristic profiles of instrumental activities of daily living in different subtypes of mild cognitive impairment*, Dementia and Geriatric Cognitive Disorders **27** (2009), no. 3, 278–285.

[21] K.S. Lee et al., *Differences in neuropsychiatric symptoms according to mild cognitive impairment subtypes in the community*, Dementia and Geriatric Cognitive Disorders **26** (2008), no. 3, 212–217.

[22] M.R. Elliott et al., *Using a Bayesian latent growth curve model to identify trajectories of positive effect and negative events following myocardial infarction*, Biostatistics **6** (2005), no. 1, 119–143.

[23] R.I. Pfeffer et al., *Measurement of functional activities in older adults in the community*, Journal of Gerontology **37** (1982), no. 3, 323–329.

[24] S. Gauthier et al., *Seminar: Mild cognitive impairment*, Lancet **367** (2006), no. 9518, 1262–70.

[25] S. Weintraub et al., *The Alzheimer's Disease Centers' Uniform Data Set (UDS): the neuropsychological test battery*, Alzheimer Disease and Associated Disorders **23** (2009), no. 2, 91–101.

[26] S.T. Farias et al., *MCI is associated with deficits in everyday functioning*, Alzheimer Disease and Associated Disorders **20** (2006), no. 4, 217–223.

[27] W.G. Rosen et al., *Pathological verification of ischemic score in differentiation of dementias*, Annals of Neurology **7** (1980), no. 5, 486–488.

[28] Y.E. Geda et al., *Prevalence of neuropsychiatric symptoms in mild congitive impairment and normal cognitive aging: population-based study*, Archives of General Psychiatry **65** (2008), no. 10, 1193–1198.

[29] P. J. Farrell and K. Rogers-Stewart, *Methods for generating longitudinally correlated binary data*, International Statistical Review **76** (2008), 28–38.

[30] M. Figueiredo and A. Jain, *Unsupervised learning of finite mixture models*, IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002), no. 3, 381–396.

[31] S.J. Finch, N.R. Mendell, and H.C. Thode Jr., *Probabilistic measures of adequacy of a numerical search for a global maximum*, Journal of the American Statistical Association **84** (1989), no. 408, 1020–1023.

[32] M.F. Folstein, S.E. Folstein, and P.R. McHugh.

[33] S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models*, Springer, New York, New York, 2006.

[34] F. Garre and J. Vermunt, *Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation*, Behaviormetrika **33** (2006), no. 1, 43–59.

[35] E.S. Garrett and S.L. Zeger, *Latent class model diagnosis*, Biometrics **56** (2000), no. 4, 1055–1067.

[36] J.A. Hagenaars, *Latent structure models with direct effects between indicators: local dependence models*, Sociological Methods and Research **16** (1988), no. 3, 379–405.

[37] J.J. Hanfelt and K.Y. Liang, *Approximate likelihood ratios for general estimating functions*, Biometrika **82** (1995), no. 3, 461–477.

[38] James W. Hardin and Joseph M. Hilbe, *Generalized Estimating Equations*, Chapman & Hall, Boca Raton, Florida, 2003.

[39] R. Hathaway, *Another interpretation of the EM algorithm for mixture distributions*, Journal of Statistics & Probability Letters **4** (1986), no. 2, 53–56.

[40] E. Kaplan, H. Goodglass, and S. Weintraub, *Boston Naming Test*, 1983.

[41] B.J. Kelley and R.C. Petersen, *Alzheimer's disease and mild cognitive impairment*, Neurologic Clinics **25** (2007), no. 3, 577–609.

[42] E.D. Kolaczyk, *An information criterion for empirical likelihood with general estimating equations*, Tech. report, Department of Statistics, University of Chicago, 1995.

[43] S. Kullback and R.A. Leibler, *On information and sufficiency*, Annals of Mathematical Statistics **22** (1951), no. 1, 79–86.

[44] B. Li, *A deviance function for the quasi-likelihood method*, Biometrika **80** (1993), no. 4, 741–753.

[45] K.Y. Liang and P.J. Rathouz, *Hypothesis testing under mixture models: Application to genetic linkage analysis*, Biometrics **55** (1999), no. 1, 65–74.

[46] K.Y. Liang and S.L. Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika **73** (1986), no. 1, 13–22.

[47] K.Y. Liang, S.L. Zeger, and B. Qaqish, *Multivariate regression analyses for categorical data*, Journal of the Royal Statistical Society. Series B **54** (1992), no. 1, 3–40.

[48] B.G. Lindsay and A. Qu, *Inference functions and quadratic score tests*, Statistical Science **18** (2003), no. 3, 394–410.

[49] T.A. Louis, *Finding the observed information matrix when using the EM algorithm*, Journal of the Royal Statistical Society. Series B **44** (1982), no. 2, 226–233.

[50] J. Magidson and J.K. Vermunt, *Latent class factor and cluster models, bi-plots and related graphical displays*, Sociological Methodology **31** (2001), no. 1, 223–264.

[51] P. McCullagh, *Quasi-likelihood functions*, Annals of Statistics **11** (1983), no. 1, 59–67.

[52] G.J. McLachlan, *On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture*, Applied Statistics **36** (1987), no. 3, 318–324.

[53] G.J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions (2nd ed.)*, John Wiley & Sons, Hoboken, New Jersey, 2008.

[54] G.J. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, New York, New York, 2000.

[55] D.L. McLeish and C.G. Small, *A projected likelihood function for semiparametric models*, Biometrika **79** (1992), no. 1, 93–102.

[56] X.L. Meng and D.B. Rubin, *Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm*, Journal of the American Statistical Association **86** (1991), no. 416, 899–909.

[57] W. Muangpaisan, S. Intalapaporn, and P. Assantachai, *Neuropsychiatric symptoms in the community-based patients with mild cognitive impairment and the influence of demographic factors*, International Journal of Geriatric Pscyhiatry **23** (2008), no. 7, 699–703.

[58] B. Muthén and K. Shedden, *Finite mixture modeling with mixture outcomes using the EM algorithm*, Biometrics **55** (1999), no. 2, 463–469.

[59] J.A. Nelder and D. Pregibon, *An extended quasi-likelihood function*, Biometrika **74** (1987), no. 2, 221–232.

[60] W. Pan, *Akaike's information criterion in generalized estimating equations*, Biometrics **57** (2001), no. 1, 120–125.

[61] ———, *Model selection in estimating equations*, Biometrics **57** (2001), no. 2, 529–534.

[62] C.G. Park and D.W. Shin, *An algorithm for generating correlated random variables in a class of infinitely divisible distributions*, The Journal of Statistical Computation and Simulation **61** (1998), no. 1-2, 127–139.

[63] R.L. Prentice, *Correlated binary regression with covariates specific to each binary observation*, Biometrics **44** (1988), no. 4, 1033–1048.

[64] R.L. Prentice and L.P. Zhao, *Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses*, Biometrics **47** (1991), no. 3, 825–839.

[65] C. Proust-Lima, 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 **26** (2007), no. 10, 2229–2245.

[66] B.F. Qaqish, *CLF: C and SAS/IML modules for various computations with multivariate bernoulli distributions in general, and simulation and computation for the conditional linear family of multivariate bernoulli distributions in particular.*, `http://www.bios.unc.edu/distrib/gee/clf/README`.

[67] ———, *A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations*, Journal of the American Statistical Association **90** (2003), no. 2, 455–463.

[68] J. Qin and J. Lawless, *Empirical likelihood and general estimating equations*, The Annals of Statistics **22** (1994), no. 1, 200–325.

[69] A. Qu, B.G. Lindsay, and B. Li, *Generalised estimating equations using quadratic inference functions*, Biometrika **87** (2000), no. 4, 823–836.

[70] B.A. Reboussin, M.E. Miller, and T.R. Ten Have, *Latent class models for longitudinal studies of the elderly with data missing at random*, Applied Statistics **51** (2002), no. 1, 69–90.

[71] J.M. Robins, A. Rotnitzky, and L.P. Zhao, *Analysis of semiparameteric regression models for repeated outcomes in the presence of missing data*, Journal of the American Statistical Association **90** (1995), no. 429, 106–121.

[72] G.E. Schwarz, *Estimating the dimension of a model*, Annals of Statistics **6** (1978), no. 2, 461–464.

[73] J. Sheikh and J.A. Yesavage, *Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version, in Clinical Gerontology: A Guide to Assessment and Intervention. Edited by Brink T. L.*, The Hawthorne Press, New York, New York, 1986.

[74] J. Stepaniuk, L.J. Ritchie, and H. Tuokko, *Neuropsychiatric impairments as predictors of mild cognitive impairment, dementia, and Alzheimer's disease*, American Journal of Alzheimer Disease and Other Dementias **23** (2008), no. 4, 326–333.

[75] E. Strauss, E.M.S. Sherman, and O. Spreen, *A Compendium of Neuropsychological Tests; Administration, Norms, and Commentary (3rd ed.)*, Oxford University Press, New York, New York, 2006.

[76] E. Teng, P.H. Lu, and J.L. Cummings, *Neuropsychiatric symptoms are associated with progression from mild cognitive impairment to Alzheimer's disease*, Dementia and Geriatric Cognitive Disorders **24** (2007), no. 4, 253–259.

[77] V.L. Torrance-Rynard and S.D. Walter, *Effects of dependent errors in the assessment of diagnostic test performance*, Statistics in Medicine **16** (1997), 2157–2175.

[78] J.K. Vermunt and J. Magidson, *Technical guide for Latent GOLD 4.0: Basic and advanced*, Tech. report, Statistical Innovations Inc., 2005.

[79] H. Wang and C. Leng, *Unified LASSO estimation by least squares approximation*, Journal of the American Statistical Association **102** (2007), no. 479, 1039–1048.

[80] J. Wang, *Nonconservative estimating functions and approximate quasi-likelihoods*, Annals of the Institute of Statistical Mathematics **51** (1999), no. 4, 603–619.

[81] L. Wang and A. Qu, *Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach*, Journal of the Royal Statistical Society. Series B **71** (2009), no. 1, 177–190.

[82] Y.G. Wang and L.Y. Hin, *Modeling strategies for longitudinal data analysis: covariate, variance function and correlation structure selection*, Computational Statistics and Data Analysis **54** (2010), no. 12, 3359–3370.

[83] Washington, War Department, Adjutant General's Office, *Army US: Army Individual Test Battery*, 1944.

[84] D. Wechsler, *The Wechsler Adult Intelligence Scale-Revised*, 1981.

[85] ———, *Wechsler Memory Scale-Revised*, 1987.

[86] R.W.M. Wedderburn, *Quasi-likelihood functions, generalized linear models and the Gaussian method*, Biometrika **61** (1974), no. 3, 439–47.

[87] S.L. Zeger and K.Y. Liang, *Longitudinal data analysis for discrete and continuous outcomes*, Biometrics **42** (1986), no. 1, 121–30.

[88] ———, *Feedback models for discrete and continuous time series*, Statistica Sinica **1** (1991), 51–64.