2010

Application of the EM Algorithm for Mixture Models

Man-Kee Maggie Chu

Application of the EM Algorithm for Mixture Models

(Spine title: EM Algorithm for Mixture Models)

(Thesis format: Monograph)

by

Man-Kee Maggie Chu

Graduate Program in Epidemiology & Biostatistics

A thesis submitted in partial fulfilment
of the requirements for the degree of
Master of Science

School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

THE UNIVERSITY OF WESTERN ONTARIO
SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES

CERTIFICATE OF EXAMINATION

Supervisor

____
Dr. John Koval

Advisor

____
Dr. Duncan Murdoch

Examiners

____
Dr. Allan Donner

____
Dr. Wenqing He

____
Dr. Hao Yu

The thesis by

Man-Kee Maggie Chu

entitled

Application of the EM Algorithm for Mixture Models

is accepted in partial fulfillment of the
requirements for the degree of
Master of Science

Date ______________________________

____
Chair of the Thesis Examination Board

ABSTRACT

A developmental trajectory describes the course of behaviour over time. Identifying multiple trajectories within an overall developmental process permits a focus on subgroups of particular interest. This research introduces a SAS macro program that identifies trajectories by using the Expectation-Maximization (EM) algorithm to fit semi-parametric mixtures of logistic distributions to longitudinal binary data. For performance comparison, we consider full maximization algorithms (e.g. SAS procedure PROC TRAJ) and standard EM, as well as two other EM-based algorithms for speeding up convergence. The simulation study shows that our EM methods produce more accurate parameter estimates than the full maximization methods. The EM-based methodology is illustrated with a longitudinal data set involving adolescents smoking behaviours.

Key Words: Expectation-Maximization algorithm, Mixture models, Binary data, Longitudinal trajectories

This thesis is dedicated to my family
for their love, support and encouragement.

ACKNOWLEDGMENTS

I would like to give my endless gratitude to Dr. John Koval for his continuous guidance and encouragement in the Masters program. He has been very supportive throughout the experience and was always available for help even during his sabbatical leave. In addition, I would like to thank my advisor Dr. Duncan Murdoch for providing helpful advices throughout the research progress. This thesis would not have been finished without their support.

I want to thank my parents for their unconditional love and encouragement to pursue my interests. Also, my heartfelt love and thanks go to my friends for their help and support during the completion of my Masters degree.

Contents

List of Tables

LIST OF FIGURES

Chapter 1

Background and Introduction

1.1 Introduction

A developmental trajectory describes the course of behaviour over age or time. Such trajectories have been used by researchers in the fields of social sciences since its introduction over a decade ago. The studies of behavioural sciences in areas such as psychology, criminology and sociology often use trajectory modeling to analyze and characterize developmental processes. As the interest in the analysis of longitudinal data increases, there is a need for the development of increasingly rigorous statistical analytic methods, including trajectory modelling.

Child and youth development can be studied through trajectory modelling of the development of school-related skills, social development, risk-taking and substance use behaviour in adolescents and young adults, and the effectiveness of targeted intervention programs. One application of trajectory modelling is the analysis of how adolescents develop the habit of smoking. Longitudinal studies of adolescent smoking habits have shown that smoking uptake behaviour progresses through a sequence of developmental stages. Identifying and characterizing these adolescent smoking trajectories would improve our understanding of the factors motivating individuals to the smoking habit, thus leading to better smoking prevention and intervention programs.

1.2 Background

The focus of this thesis is on a method to analyze temporal trajectories of longitudinal binary data. Cross-sectional data are limited in the sense that hypotheses relating to change cannot be evaluated with such data. On the other hand, longitudinal studies follow the population over time with the repeated measurements that can reflect the trend of an outcome over time. In addition, longitudinal data would allow for the separation of aging effects (changes over time within subjects) from cohort effects (differences between subjects at baseline) (Diggle et al., 1994). Longitudinal study designs are becoming more popular because they can provide more efficient estimators than cross-sectional designs with the same number and pattern of observations. Subjects serve as their own controls so that between-subject variation can be excluded from the error term when examining effects of interest.

Analysis of longitudinal data is complicated by the correlation that exists in the data. Since the data consists of repeated measurements over the same subjects, this means that the observations are not independent and therefore researchers must account for the dependency in the data. Analytic methods for longitudinal data are not as well developed for more sophisticated models. Some general approaches of dealing with longitudinal data include random effects regression models, Generalized Estimating Equations (GEE) models, and other Generalized Linear Models (GLMs). Since the choice of analysis method depends on the question of interest, we focus on statistical methods suitable for analyzing mixture models: modelling longitudinal trajectories of distinct groups.

1.3 Cluster analysis and Mixture models

A mixture model is a model for analyzing data from mixture distributions, in problems when the measurements of a random variable are taken under different conditions. Mixture models assume that the data are collected from a number of subpopulations, and the data within each subgroup can be modelled using a standard statistical model. If the number of subpopulations is finite, then these models are called finite mixture models; otherwise, they are continuous mixture models. The frailty model for survival data is an example of continuous mixture models. In this thesis we will focus on the finite mixture distributions.

As an example, a mixture model can be applied to model the average height of people in a country. For instance, if we consider all Canadians as a single population, then we ignore the possibility that the individuals' heights might systematically differ, depending on their characteristics such as gender. Ignoring the group differences may lead to biased and error-prone estimates and inappropriate predictions and hypothesis tests (Nurmi, 2010). We would be able to get more accurate estimates if we divide the Canadian population into subgroups based on gender or age, within each of which a simple model would apply.

Mixture models can be regarded as a type of clustering model, where each component probability distribution corresponds to a cluster. The idea of cluster analysis is to group previously unstructured data into distinct groups containing data that are similar in some sense. Typically in mixture modelling problems, the number of subpopulation is unknown and one needs to use the data to determine the optimal number of components. The model parameters within each subgroup are also unknown and must be estimated from the data.

In some cases, there may only be hypotheses as to the number of components; the goal is to find a suitable set of subpopulations. This is usually done via model-based clustering methods, where we have a cluster model and the objective is to divide the data optimally into the clusters. Since group membership is unknown, the classification of the observations into the different components has to be carried out simultaneously with parameter estimation.

In recent years model-based clustering has appeared in statistics literature with increased frequency. Mixtures of multivariate normal densities have been considered by many researchers, including Wolfe (1970) and Day (1969). Frühwirth-Schnatter (2006) noted that some practical applications of model-based clustering using Gaussian mixtures include character recognition, minefield and seismic fault detection, clustering gene expression data and classification of astronomical data. Mixture models with non-normal components that have been studied by researchers include mixtures of the exponential (Heckman et al., 1990), Poisson (Karlis and Xekalaki, 2005), binomial (Wang and Puterman, 1998), and multinomial distributions (Jorgensen, 2004).

There have been many methods designed for parameter estimations in mixture models, ranging from Pearson's (1894) method of moments to informal graphical techniques and formal maximum likelihood approaches. Everitt and Hand (1981) discussed some generally applicable methods, including the maximum likelihood (ML) method. The ML method for estimating the parameters has desirable statistical properties: the estimators obtained by the method are consistent and they are asymptotically normally distributed (Everitt and Hand, 1981). The ML equations for parameter estimation are not usually explicitly solvable so they need to be solved using some form of iterative procedure.

For mixture modelling and model-based clustering, the iterative method usually employed for the ML estimation was first suggested by Hasselblad (1966) and Wolfe (1970), which was later called the Expectation-Maximization (EM) algorithm by Dempster, Laird, and Rubin (1977). This algorithm has two steps. In the first step, the probability of each observation belonging to each component of the mixture model is estimated. Then the second step evaluates the estimation problem, with each observation contributing to the log-likelihood with a weight given by the membership probabilities estimated in the first step. These steps are then repeated until convergence.

1.4 Applications of mixture models to longitudinal trajectories

Hierarchical modeling and latent curve analysis are two popular approaches for analyzing developmental trajectories. Nagin (1999; 2005) noted that these two standard growth curve modelling methods use unconditional models to estimate the mean and covariance structure of the population distribution of growth curve parameters, and use conditional models to explain the variability in growth throughout the population by relating the growth parameters to explanatory variables. The "semi-parametric" group-based approach proposed by Nagin (1999; 2005) focuses on identifying relatively homogeneous clusters of developmental trajectories. In summary, hierarchical and latent curve methodologies model population variability in growth with multivariate continuous distribution functions for analyzing individual-level trajectories; while the group-based method uses mixtures of suitably defined probability distributions to identify distinctive clusters of individual trajectories within the population.

The "semi-parametric" group-based method assumes that the population is composed of a mixture of distinct groups defined by their trajectories, rather than assuming a continuous distribution of trajectories within the population. The assumption of distinct subgroups may not be correct, as the development of behaviour may not follow such clear-cut categories. However, the powerful feature of the group-based approach is that, by identifying the clusters of individuals with similar trajectories, differences that may explain individual-level variability can be expressed in terms of group differences (Nagin, 2005).

The standard growth curve modelling approach is more appropriate than the group-based method in situations where the developmental process of all population members follow a common pattern of increase or decrease. Raudenbush (2001) gave examples related to language acquisition or academic learning in early childhood. For phenomena in which there may be different trajectories of change across subpopulations, such as when gang membership is the outcome of interest (Lacourse et al., 2003), the group-based method is a useful approach. This group-based approach is appropriate when the assumption that all individuals within the population follow a common trend that increases or decreases regularly may be violated, or if the objective of the analysis is to discover distinctive developmental trends in change. In general, the standard growth curve modelling is suitable for analyzing questions in terms of predictors of the outcome's developmental course, while the group-based method can answer questions in terms of the shape of the developmental course of the outcome of interest (Nagin, 2005).

In order to describe the changes in behaviour over time through developmental trajectories, the "semi-parametric" model proposed by Nagin (1999) for longitudinal data links behaviour to age or time. In summary, this group-based trajectory modeling method was designed to: (1) determine the optimal number of distinctive groups of trajectories and identify those trajectories, (2) estimate the proportion of the population that is believed to belong to each trajectory group, (3) relate the group assignments to individual characteristics, and (4) use the group membership probabilities for purposes such as creating profiles of group members.

The "semi-parametric" group-based estimation model has been implemented as a SAS based procedure, PROC TRAJ, by Jones, Nagin and Roeder (2001). This SAS procedure has been used by researchers to identify longitudinal trajectories on the development of the smoking habit. To understand the development of smoking behaviour in youth, Driezen (2001) used PROC TRAJ to analyze longitudinal data from the third Waterloo Smoking Prevention Project. The goal was to identify distinct groups of smoking initiation trajectories and regular smoking trajectories among a cohort of grade 6 students, followed for a seven year period (1990-1996). Among 2306 students who reported as non-smokers or non-regular smokers at baseline, five groups of smoking initiation trajectories were identified: never smoked, and early, mid-early, mid-late, and late onset. Likewise, among 2495 students with complete smoking histories, five distinct groups of regular smoking trajectories were identified: never regular, early uptake, mid uptake, late uptake, and dabblers.

Karp et al. (2005) have analyzed smoking trajectories of data from The McGill University Study on the Natural History of Nicotine Dependence, which consisted of a student population recruited from grade 7 classes of a sample of Montreal secondary schools and followed for seven years (1999-2005). The researchers analyzed data from the first 14 questionnaires, administered every 3 to 4 months during the first 3.5 years of follow up. The objective of the study was to describe trajectories of smoking intensity in adolescent novice smokers and to identify predictors of trajectory group membership. The statistical analysis included: (1) using individual growth modeling to uncover the overall trajectory of smoking intensity, and (2) performing the "semi-parametric" group-based modelling to classify major classes of trajectories. From the 269 novice smokers included in the analysis, four groups of smoking intensity trajectories were identified by PROC TRAJ: low-intensity, non-progressing smokers, and slow, moderate, and rapid escalators.

In order to understand the smoking behaviours in Canadian youth from late childhood to adolescence, Maggi et al. (2007) used the group-based mixture modelling method to identify smoking trajectories among participants of the Canadian National Longitudinal Survey of Children and Youth. Among children and youth from 10 to 17 years of age, the researchers examined questions regarding smoking behaviour such as trying smoking and frequency and intensity of smoking. They used PROC TRAJ to estimate growth mixture models for smoking behaviours and identified three trajectories for the probability of having tried smoking from the 2886 youths and children: early, middle, and late onset smokers. From 280 smokers regarding frequency of smoking, five distinct groups were discovered: early, and late infrequent experimenters, early frequent experimenters, as well as early, and late frequent smokers. The intensity of smoking reported by the subpopulation of 260 regular smokers could be classified into two groups: late and slow, or early and rapid escalators, with respect to the number of cigarettes smoked daily.

White et al. (2002) reported on a study which analyzed the smoking behaviours and the risk factors related to smoking among a group of 374 individuals in New Jersey. Participants were first interviewed in 1979-1981 at the age of 12, and then re-visited over the years and the fifth and final interview was conducted in 1997-1999 at the age of 30 or 31. Information regarding cigarette use was collected, such as frequency of smoking and typical quantity per day, as well as risk factors including demographic characteristics, differential association variables and intrapersonal characteristics. Using PROC TRAJ, they identified three trajectory groups with respect to cigarette use: non/experimental smokers, occasional/maturing out smokers, and heavy/regular smokers.

1.5 Objective

Nagin (1999) proposed the use of a "semi-parametric" model to identify homogeneous clusters of longitudinal developmental trajectories, and a SAS procedure called PROC TRAJ had been created to estimate parameters in this model (Jones et al., 2001). This procedure performs a maximization using the Quasi-Newton method to obtain parameter estimates, but the use of this procedure requires a careful choice of starting values to ensure convergence (Roeder et al., 1999). Some problems that have been encountered when using PROC TRAJ are: (1) the procedure sometimes fail to converge, and (2) it converges to a false maximum or a local maximum instead of global maximum (Driezen, 2001; Nawa, 2004). The EM algorithm was proposed as the solution to the convergence problems since it has been suggested as a better algorithm than the Quasi-Newton method for computing MLE's for mixtures of normal distributions (Davenport et al., 1988). Roeder et al. (1999) used the EM algorithm to model longitudinal trajectories of count data; while Nawa (2004) proposed using the EM algorithm to model binary data.

The objective of this research is to extend the EM algorithm for trajectory modelling proposed by Nawa (2004), focusing on the model for binary longitudinal data. We would like to extend the EM algorithm into a method with improved convergence properties and speed. To improve and speed up the EM convergence, we propose the use of iteratively reweighted least squares (IRLS) to fit a weighted logistic regression model at the maximization stage of the EM algorithm. We evaluate the performance of the algorithm based on measures of accuracy, in hopes of developing an algorithm with good convergence property, small estimated mean squared error of prediction, and small relative error for parameter estimates. This research also aims to provide an open source SAS/IML macro program that is publicly available for other researchers to enhance future analyses of longitudinal data.

Chapter 2

The Model and Methods of Estimation

2.1 Introduction

Mixture models can be used when it is believed that there is unobserved heterogeneity in the population, and that there exist subgroups with different parameter values within the population. In mixture modeling with longitudinal data, some models that are commonly used include latent class growth analysis (LCGA) and the growth mixture model (GMM). The "semi-parametric" group-based approach proposed by Nagin (1999) is an example of LCGA, which is the simplest longitudinal mixture model for binary or categorical measurements. This model assumes that there is no variation across individuals within a class, whereas the GMM (Muthén and Shedden, 1999) allows for within-class variation of individuals. The GMM is a more complex model where the within-class variation is represented by random effects, and it is more suitable for situations where the latent classes corresponding to one set of variables influence another set of observed variables. Muthén and Muthén (2007) described other types of longitudinal mixture models, including latent transition analysis (also referred to as hidden Markov modelling) and discrete- and continuous-time survival mixture modelling.

The group-based trajectory modeling method proposed by Nagin (1999) was designed to identify distinctive groups of individual trajectories within the population. The methodology estimates the number of groups that best fits the data and the proportion of the population following each trajectory group. The shape of the trajectory for each group is estimated and along with the group membership probabilities, profiling of the characteristics of group members can be obtained for analysis. This group-based trajectory modelling methodology is an application of finite mixture models. We will focus on the model for binary longitudinal data. The trajectories estimated by this group-based method are produced by maximum-likelihood estimation.

2.2 Mixture Models

2.2.1 Likelihood formulation for mixture models

Following Frühwirth-Schnatter (2006), we consider a population that is made up of g subgroups, mixed at random in proportion to relative group sizes. Suppose we are interested in some random feature Y which is heterogeneous across and homogenous within the subgroups. When we sample from such a population, we can record not only Y, but also the group membership indicator S, $S \in \{1, ..., g\}$.

Suppose we have the mixing proportions (or component weights) indicated by π_i, $i = 1, ..., g$, where each π_i is non-negative and $\sum_{i=1}^{g} \pi_i = 1$. We then have the probability of sampling from group S is equal to π_s.

For the population, the joint density $p(y, S)$ is given by

$$\begin{aligned} p(y, S) &= p(y|S)p(S) \\ &= \pi_s p(y|S), \end{aligned}$$

and conditional on the group S, Y is a random variable following a distribution $p(y|\theta_s)$ with θ_s being the parameter of group S. Due to heterogeneity in the population, Y has a different probability distribution $p(y|\theta_s)$ for each subgroup.

A finite mixture distribution arises if the group indicator S is not observed, that is, the population is modelled as consisting of g distinct groups (or components) in some unknown mixing proportions $\pi_1, ..., \pi_g$.

Let $Y_1, ..., Y_n$ be a random sample of size n, and $\boldsymbol{y} = (y_1, ..., y_n)^T$ denote the observed random sample where y_j is the realization of the random variable Y_j. Then a standard g-component mixture model can be expressed in the form

$$f(y_j; \boldsymbol{\psi}) = \sum_{i=1}^{g} \pi_i f_i(y_j; \theta_i),$$

where $f_i(y_j; \theta_i)$ is the component density for component i, which is the conditional density function of Y_j given group membership of the i^{th} component, and $\boldsymbol{\psi} = (\pi_1, ..., \pi_g, \theta_1, ..., \theta_g)$ is the set of model parameters from the different mixture components. The corresponding likelihood is given by

$$\begin{aligned} L(\boldsymbol{\psi}) &= \prod_{j=1}^{n} f(y_j; \boldsymbol{\psi}) \\ &= \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_i f_i(y_j; \theta_i). \end{aligned}$$

2.2.2 Complete-data likelihood for mixture models

In finite mixture models, the mixing proportions $\pi_1, ..., \pi_g$ and component parameters $\theta_1, ..., \theta_g$ are unknown and need to be estimated from the data. We denote these unknown parameters as $\boldsymbol{\psi} = (\pi_1, ..., \pi_g, \theta_1, ..., \theta_g)$. As previously mentioned, we denote $\boldsymbol{S} = (S_1, ..., S_n)^T, S_j \in 1, ..., g$, to indicate the group allocation of individual j.

Under the assumption that the allocations $\boldsymbol{S} = (S_1, ..., S_n)^T$ are observed, we can estimate parameters $\boldsymbol{\psi}$ based on the complete data $(\boldsymbol{y}, \boldsymbol{S})$. The complete-data likelihood function is equal to the sampling distribution $p(\boldsymbol{y}, \boldsymbol{S}|\boldsymbol{\psi})$ of the complete data $(\boldsymbol{y}, \boldsymbol{S})$, regarded as a function of the unknown parameter $\boldsymbol{\psi}$. It can be written as

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{S}|\boldsymbol{\psi}) &= p(\boldsymbol{y}|\boldsymbol{S}, \boldsymbol{\psi})p(\boldsymbol{S}|\boldsymbol{\psi}) \\ &= \prod_{j=1}^{n} p(y_j|S_j, \boldsymbol{\psi})p(S_j|\boldsymbol{\psi}). \end{aligned}$$

Given group i, we know that

$$p(y_j|S_j = i, \boldsymbol{\psi}) = p(y_j|\theta_i)$$

and

$$Pr(S_j = i|\boldsymbol{\psi}) = \pi_i.$$

Then the complete-data likelihood becomes

$$p(\boldsymbol{y}, \boldsymbol{S}|\boldsymbol{\psi}) = \prod_{j=1}^{n}\prod_{i=1}^{g} [\pi_i p(y_j|\theta_i)]^{I\{S_j=i\}}.$$

However, in the mixture model context we do not observe the allocations $\boldsymbol{S}$. Following Nawa (2004), we define an unobserved or missing data vector $\boldsymbol{z} = (\boldsymbol{z}_1^T, ..., \boldsymbol{z}_n^T)^T$, where $\boldsymbol{z}_j = (z_{1j}, ..., z_{gj})$ is a vector of indicator variables reflecting the group membership of individual j. We define $z_{ij} = I\{S_j = i\}$, indicating that $z_{ij} = 1$ if individual j belongs to group i and $z_{ij} = 0$ otherwise. This also implies that

$$\sum_{i=1}^{g} z_{ij} = 1.$$

Suppose we have an observed sample of size n, denoted as $\boldsymbol{y} = (y_1, ..., y_n)^T$, then the complete-data likelihood for a g-component mixture model can be expressed as

$$\begin{aligned} L_c(\boldsymbol{\psi}) &= \prod_{j=1}^{n}\prod_{i=1}^{g}(\pi_i f_i(y_j;\theta_i))^{z_{ij}} \\ &= \prod_{j=1}^{n}\prod_{i=1}^{g}\pi_i^{z_{ij}} f_i(y_j;\theta_i)^{z_{ij}}. \end{aligned}$$

2.3 Mixture models: Binary longitudinal data

2.3.1 The Likelihood

When working with binary longitudinal data, we denote $Pr(\boldsymbol{Y}_j)$ as the probability of observing a specific longitudinal sequence of binary measurements on individual j over time. The goal is to obtain a set of parameters such that the likelihood is maximized. These parameters define the shape of the trajectories and the probability of group memberships. The shape of each trajectory is described by a polynomial function of age or time, and a separate set of parameters is estimated for each group to allow the shapes of trajectories to differ across groups.

Suppose we have $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ as a random sample of size n, where $\boldsymbol{Y}_j$ is a m-dimensional vector. We have $\boldsymbol{y}_j = (y_{j1}, y_{j2}, ..., y_{jm})$ representing a longitudinal sequence of observations over m time points, where the response y_{jt} $(t = 1, ..., m)$ observed at the t^{th} time point recorded as a binary measurement.

We will assume a quadratic relationship between age (or time) and behaviour on the logit scale, and we model with the assumption that conditional on membership in group i, the probability of outcome of interest can be written as

$$Pr(Y_{jt}) = \frac{e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}{1+e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}},$$

with age_{jt} being the age of individual j at time t. The parameters β_0^i, β_1^i, and β_2^i determine the shape of the trajectories, and they are allowed to vary across the different trajectories. A positive β_1 and a negative β_2 show a single peaked trajectory, while a constant trajectory is shown if β_1 and β_2 equal to zero.

Conditional on being in group i, a subject j is assumed to have independent observations over the m time points, so we have

$$f_i(\boldsymbol{y}_j;\boldsymbol{\beta}_i) = \prod_{t=1}^{m}\left(\frac{e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}{1+e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}\right)^{y_{jt}}\left(\frac{1}{1+e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}\right)^{1-y_{jt}}.$$

The likelihood for the entire sample of n individuals is

$$\begin{aligned} L(\boldsymbol{\psi}) &= \prod_{j=1}^{n}\sum_{i=1}^{g}\pi_i f_i(\boldsymbol{y}_j;\boldsymbol{\beta}_i) \\ &= \prod_{j=1}^{n}\sum_{i=1}^{g}\pi_i\prod_{t=1}^{m}\left(\frac{e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}{1+e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}\right)^{y_{jt}}\left(\frac{1}{1+e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}\right)^{1-y_{jt}} \end{aligned}$$

and the corresponding log-likelihood is

$$l(\boldsymbol{\psi}) = \sum_{j=1}^{n}\log\left[\sum_{i=1}^{g}\pi_i\prod_{t=1}^{m}\left(\frac{e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}{1+e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}\right)^{y_{jt}}\left(\frac{1}{1+e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}\right)^{1-y_{jt}}\right].$$

The maximum likelihood estimates, $\hat{\boldsymbol{\psi}} = (\hat{\pi}_1, ..., \hat{\pi}_g, \hat{\boldsymbol{\beta}}^1, ..., \hat{\boldsymbol{\beta}}^g)$, can be obtained by maximizing the above log-likelihood.

2.3.2 The complete-data likelihood

With $\boldsymbol{y}_j = (y_{j1}, y_{j2}, ..., y_{jm})$ representing a sequence of longitudinal measurements for individual j over m time points, the complete-data likelihood for the entire sample of n individuals is

$$\begin{aligned} L_c(\boldsymbol{\psi}) &= \prod_{j=1}^{n}\prod_{i=1}^{g} \pi_i^{z_{ij}} f_i(\boldsymbol{y}_j; \boldsymbol{\beta}_i)^{z_{ij}} \\ &= \prod_{j=1}^{n}\prod_{i=1}^{g} \pi_i^{z_{ij}} \left[\prod_{t=1}^{m} \left(\frac{e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}{1+e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}\right)^{y_{jt}} \left(\frac{1}{1+e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2}}\right)^{1-y_{jt}}\right]^{z_{ij}}, \end{aligned}$$

and the corresponding log-likelihood is

$$\begin{aligned} l_c(\boldsymbol{\psi}) &= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\log\pi_i + \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\log f_i(\boldsymbol{y}_j; \boldsymbol{\beta}_i) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\log\pi_i \\ &+ \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\left[\sum_{t=1}^{m} y_{it}(\beta_0^i + \beta_1^i age_{jt} + \beta_2^i age_{jt}^2) - \log(1+e^{\beta_0^i+\beta_1^i age_{jt}+\beta_2^i age_{jt}^2})\right]. \end{aligned}$$

The EM algorithm (Dempster, Laird and Rubin, 1977) can be used to obtain maximum likelihood estimates, $\hat{\boldsymbol{\psi}}$, from the above complete-data log-likelihood.

2.4 Implementation in SAS

Jones, Nagin and Roeder (2001) developed a SAS procedure for estimating developmental trajectories, which is based on the "semi-parametric" group-based modeling strategy proposed by Nagin (1999). The procedure, called PROC TRAJ, uses a model which is a mixture of probability distributions that are specified to describe the data to be analyzed. PROC TRAJ can model three different distributions: the zero-inflated Poisson (ZIP) model for analyzing count data, the censored normal (CNORM) model for psychometric scale data, and the logistic (LOGIT) model for binary data. The user defines the input information such as the type of data to be analyzed, the number of groups to be fitted, and the shape of the trajectory to be fitted which could be a linear, quadratic or cubic function of age or time. Also, the initial starting values for each of the parameters can be specified, or the default values will be used for the model fitting. By default, the procedure uses starting values which assume constant trajectories evenly spaced through the range of the dependent variables. Parameter estimates are obtained through maximum likelihood (presented in Section 2.3.1) and performed using a Quasi-Newton method. PROC TRAJ is a compiled procedure written in the C programming language and can only be used in SAS for WINDOWS. A macro called trajplot can be used to plot the obtained trajectories, or users may make use of other software (such as MS Excel) for plotting the trajectories.

2.5 Number of groups

When one is working with finite mixture models, often the number of components or groups is unknown. Fitting too many groups would lead to the problem of overfitting, such that trajectory groups reflect only random variation. On the other hand, fitting too few groups to the data may result in a model that is not flexible enough to approximate the true underlying distribution. There are several statistical tools for determining the optimal number of distinct groups in a mixture model.

The chi-square likelihood ratio statistic could be used to determine the most appropriate number of distinct groups (Everitt and Hand, 1981); however, it is not suitable for mixture modeling because a g-component model is not nested in the interior of the parameter space of a $(g+1)$-component model. In the group-based trajectory modeling context, the problem is caused by the null hypothesis. The null hypothesis (i.e. g groups) is on the boundary of the parameter space, because we set the probability of being in the $(g+1)^{st}$ group to zero. The classical asymptotic results that underly the likelihood test would not hold under such a situation (Nagin, 1999). Since the regularity conditions of the test statistics are not met, the null distribution of the likelihood ratio statistic does not converge to a chi-square distribution and the calculated p-value obtained would not be correct (Nylund et al., 2007). McLachlan and Peel (2000, Section 6.4 and 6.5) and Frühwirth-Schnatter (2006, Section 4.4) provide more details and reviews of relevant literature. Simulation studies conducted by Everitt (1981; 1988) and Nylund et al. (2007) have shown the inappropriateness of using the chi-square likelihood ratio test when working with mixture models. Nylund et al. (2007) noted that, although the chi-square difference test in the form of the likelihood ratio test cannot be used for mixture model selection, there are alternative likelihood ratio tests that may be appropriate.

Other methods for determining the number of groups in mixture models include the Akike's Information Criterion (AIC) (Akaike, 1973) and the Bayesian Information Criterion (BIC) (Schwarz, 1978). Both measure the goodness-of-fit based on the log likelihood of a fitted model, while penalizing for model complexity and/or sample size. Since they use different penalties, the two criteria may point to a different number of groups as the best model.

The AIC is defined as

$$AIC = -2\log(L) + 2k$$

where k is the number of free parameters in the model. This depends on the number of groups and the function used for describing the shape of trajectories. For example, for a three-groups model with trajectories described by quadratic functions, there are eleven parameters (nine parameters for the three trajectories and two for the mixing proportions). Akaike (1973) suggested choosing the model which gives the smallest AIC over the set of models considered.

The BIC is defined as

$$BIC = \log(L) - 0.5k\log(n)$$

where k is again the number of free parameters in the model and n is the sample size. The model with the maximum BIC value, i.e. least negative number (since BIC is always negative), is recommended as the best finite mixture model. The BIC criterion can be used for comparison of both nested and non-nested models (Kass and Raftery, 1995).

It has been shown that the use of either AIC or BIC as a criteria for mixture model selection would not underestimate the true number of groups in the population (Leroux, 1992). The use of BIC is often preferred over AIC because the BIC is consistent as a selection criterion, whereas the AIC has been shown to be not consistent (Bozdogan, 1987). In particular, the probability that the BIC will select the true model approaches one as the sample size becomes large, while the AIC tends to choose more complex models as the sample size increases. Researchers had performed simulation studies to evaluate the various model choice criteria including AIC and BIC. Keribin (2000) found that BIC can determine the optimal number of groups in finite mixture models, with it being consistent (avoiding over- or underestimation) under correct specification of the group density families. Nylund et al. (2007) conducted a Monte Carlo simulation study that examined the performance of likelihood ratio tests and several Information

Criterion (ICs) used for determining the number of groups in mixture models. Comparing the performance of AIC, consistent AIC (CAIC), BIC, and adjusted BIC across different mixture models and sample size specifications, they showed that the BIC is the best of the ICs considered. They found that AIC is not a good criterion for identifying the correct model for any of the modelling settings being considered. Also, the accuracy of AIC decreased as sample size increased, reflecting a known problem with AIC because there is no adjustment for sample size. These results are in agreement with previous research indicating the AIC is not a good indicator for determining the optimal number of groups (Celeux and Soromenho, 1996; Yang, 2006), and that BIC performs well in the context of mixture models (Keribin, 2000). PROC TRAJ implemented the calculation of both BIC and AIC. In our proposed EM approach, we choose a model which maximizes BIC among the different component models.

2.6 Newton-type optimization methods

Parameter estimations in statistical problems often involve the maximization or minimization of an objective function. For mixture models, maximum likelihood (ML) estimation has been the approach most widely considered in the literature. To find the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}$ of a parameter $\boldsymbol{\theta}$, we use the log-likelihood function $l(\boldsymbol{\theta}, \boldsymbol{y})$ and find the value of $\boldsymbol{\theta}$ that maximizes the log-likelihood function. The MLE can be found by differentiating the log-likelihood and equating the derivative with zero. This derivative is called the score function, $S(\boldsymbol{\theta}, \boldsymbol{y})$, so that we have

$$S(\boldsymbol{\theta}, \boldsymbol{y}) = \frac{\partial l}{\partial \boldsymbol{\theta}} = 0.$$

The score function is often nonlinear, thus requiring iterative root-finding algorithms to obtain the solution. Different types of iterative algorithms are used to perform nonlinear optimization, namely the Newton-Raphson, Fisher Scoring, and Expectation-Maximization (EM) algorithms. Starting with some initial value as the parameter estimate for $\boldsymbol{\theta}$, the estimate gets updated through iterations and eventually converges to the MLE of interest, $\hat{\boldsymbol{\theta}}$. In this section, we focus on Newton-type optimization methods. An advantage of the Newton-type optimization methods compared with EM is that the Newton-type methods provide estimates of the standard errors for the MLE's as a by-product of the maximization process. The covariance matrix of the estimated parameters can be obtained using the Hessian matrix.

2.6.1 Newton-Raphson method

The Newton-Raphson method is one of the best known methods for numerically evaluating roots of complex functions. The Newton sequence is

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - H^{-1}(\boldsymbol{\theta}^{(k)})S(\boldsymbol{\theta}^{(k)}),$$

where $\boldsymbol{\theta}^{(k)}$ is the ML estimate at the k^{th} iteration, $H^{-1}(\boldsymbol{\theta}^{(k)})$ is the inverse of the Hessian matrix, and S is the score function. The Hessian matrix H is the matrix of the second derivatives of the log-likelihood, $\frac{\partial^2 l}{\partial \boldsymbol{\theta}^2}$, and is the negative of the observed information matrix.

Some disadvantages of the Newton-Raphson method include its sensitivity regarding starting values and slow convergence. The initial estimate for starting the iterations should not be a "guess", as it should be selected such that it has as many properties of the solution as possible (Kelley, 2003 pg 15). When it works, Newton-Raphson can find the solution rapidly. However, if the initial estimate is not close enough to the solution, the Newton-Raphson method may not converge, or may converge to the wrong root, such as converging to a local maximum instead of the global maximum.

In some situations, the iteration may fail to converge to a root, when either the iteration becomes unbounded or the Hessian matrix is non-invertible (Kelley, 2003 pg 18). One of the main drawbacks of the Newton-Raphson method is that the Hessian may become numerically singular when iterations are far from the maximum point. The other problem is that the calculation of the Hessian matrix might be very computational intensive for high dimensions, leading to very slow convergence. There have been alternatives proposed to speed up the convergence of this algorithm.

2.6.2 Fisher scoring method

The Fisher scoring algorithm is similar to the Newton-Raphson method, except the Fisher's information matrix (the expected information matrix) is used instead of the observed information matrix (the negative of the Hessian matrix). For generalized linear models, the two methods are the same if the canonical link function is used, that is, the expected value and the actual value of the Hessian matrix are equivalent for the canonical link (McCullagh and Nelder, 1989). The Fisher scoring method is more reliable than the Newton-Raphson method in the sense that, for a well-defined model, the expected information matrix is more likely to be positive definite than the negative Hessian matrix. Also, compared to the observed information matrix, the expected information matrix is more robust to possible outliers; thus leading to a better estimate of the approximate standard errors at the final iteration (Demidenko, 2004 pg 86). However, one disadvantage of the Fisher scoring method is that, in some cases it is difficult to evaluate the Fisher's information matrix analytically.

2.6.3 Quasi-Newton method

Since Newton-Raphson and Fisher scoring methods require the calculation of the second order derivatives of the log-likelihood with respect to the parameters; this computation may be very difficult and is often very slow. For solving nonlinear-equation systems with n-dimensions, the Newton methods require n^2 second derivative evaluations, n first derivative evaluations and a matrix inverse before even the linear search can be attempted (Nash, 1990 pg 187). The Quasi-Newton algorithm has been proposed as a solution to this slow computation, because the Quasi-Newton method uses an approximation to the Hessian to update the nonlinear iteration sequence. The inverse of the Hessian matrix can be approximated directly from the first derivative information at each step of the iteration, so that the calculation of the second partial derivatives can be avoided (Nash, 1990 pg 187). This greatly reduces the amount of computation needed to obtain the Hessian matrix and its inverse. This method reduces the tendency of the Newton-Raphson method to lead to local minima or maxima by forcing the approximate Hessian to be negative definite; however, there is still no guarantee of global convergence (McLachlan and Krishnan, 2008). As well, the Quasi-Newton method still suffers from being too sensitive to the initial iterate estimates, because initially it approximates the Hessian by the identity, which may be a poorly scaled approximation to the estimation problem (McLachlan and Krishnan, 2008 pg 6).

2.7 EM algorithm

2.7.1 Introduction

Since the EM algorithm was presented in a paper by Dempster, Laird, and Rubin (1977), it has become a popular algorithm for ML estimation in a wide variety of situations. McLachlan and Krishnan (2008) noted that the EM algorithm is the most suitable method for handling parameter estimations in incomplete-data problems such as missing data, truncated distributions and censored or grouped observations. EM is the preferred approach in these situations, where the Newton-type methods may be more complicated due to the absence of some part of the data. Another application of the EM algorithm is in the optimization of the likelihood function when that likelihood is analytically intractable, but the likelihood function can be simplified by assuming the values for additional parameters as missing. In other words, the incompleteness of the data is not natural or evident. It would then depend on the statistician to formulate the incompleteness in an appropriate manner to facilitate the application of the EM algorithm.

Each iteration of the EM algorithm consists of two steps: the Expectation step (E-step) and the Maximization step (M-step). During the E-step, the algorithm finds the expected value of the complete-data log-likelihood with respect to the unknown data, given the observed data and the current parameter estimates. The M-step of the algorithm would then maximize the expected log-likelihood obtained in the first step and update the parameter estimates. Starting from some initial values, the E- and M-steps are repeated until some convergence criterion is satisfied. Each iteration is guaranteed to increase the log-liklihood and thus the algorithm is guaranteed to converge to a local maximum of the ML function. The EM algorithm for the mixture modeling problem has been studied by several authors (Hathaway, 1986; McLachlan and Peel, 2000; Meng, 1997); Redner and Walker (1984) noted that the algorithm has been found, in most instances, to have the advantages of reliable global convergence, low cost per iteration and ease of programming.

A main drawback of the EM algorithm is that it can be very slow to converge in some situations. Researchers have been developing modified versions of the algorithm in attempt to solve this problem, as well as other simulation-based methods and extensions. To speed up the estimation procedure, authors such as Redner and Walker (1984) and Aitkin and Aitkin (1996) have proposed the use of hybrid algorithms such as combining the EM algorithm with Newton's method. Another criticism of EM algorithm is that the covariance matrix of the estimated parameters is not produced as an end-product of the algorithm, but there are methods for obtaining approximate standard errors from EM algorithms (Louis, 1982; McLachlan and Krishnan, 2008; Meng and Rubin, 1991). In the context of mixed logistic regression models, Wang and Puterman (1998) reported the use of a hybrid algorithm for speeding up the convergence and obtaining approximate standard errors for estimates. They performed the EM algorithm for parameter estimates until some convergence criteria has been met, and then switched to the Quasi-Newton method so that approximations of standard errors could be obtained as a by-product of the Newton maximization approach.

2.7.2 EM estimation for longitudinal trajectory models

The EM algorithm can be used to obtain MLE's for the group-based trajectory models by maximizing the complete-data log-likelihood previously described in

Section 2.3.2, with the inclusion of a missingness component. The EM algorithm is implemented by treating the unknown group membership of the mixture population as missing data, so that the data is augmented with indicators of group membership.

In the EM framework, starting from some initial value for $\boldsymbol{\psi}$, say $\boldsymbol{\psi}^{(0)}$, the E-step involves the calculation of the expectation of the complete data log-likelihood, conditional on the observed data and the initial estimate $\boldsymbol{\psi}^{(0)}$. Since $\boldsymbol{y}$ and $\boldsymbol{\psi}^{(0)}$ are constants, the conditional expectation depends only on the expectation of Z_{ij}.

The **E-step** of the $(k+1)^{th}$ iteration involves the evaluation of

$$\begin{aligned} E(Z_{ij}|\boldsymbol{y}_j;\boldsymbol{\psi}^{(k)}) &= \frac{\pi_i^{(k)} f_i(\boldsymbol{y}_j;\boldsymbol{\beta}_i^{(k)})}{f(\boldsymbol{y}_j;\boldsymbol{\psi}_i^{(k)})} \\ &= \frac{\pi_i^{(k)} f_i(\boldsymbol{y}_j;\boldsymbol{\beta}_i^{(k)})}{\sum_{i=1}^{g} \pi_i^{(k)} f(\boldsymbol{y}_j;\boldsymbol{\beta}_i^{(k)})} \\ &= \hat{z}_{ij}^{(k)}. \end{aligned}$$

The resulting estimate is the posterior probability that individual j belongs to group i.

The **M-step** then determines the value of $\boldsymbol{\psi}$ that maximizes the complete-data log-likelihood with each z_{ij} replaced by the corresponding posterior probability, that is, the evaluation of

$$\boldsymbol{\psi}^{(k+1)} = argmax_{\boldsymbol{\psi}} E[\log L(\boldsymbol{\psi}|\boldsymbol{y};\boldsymbol{\psi}^{(k)})],$$

which is given by

$$\hat{\pi}_i^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}$$

$$\hat{\boldsymbol{\beta}}_i^{(k+1)} =$$

$$argmax_{\boldsymbol{\beta}_i} \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \left[\sum_{t=1}^{m} y_{it}(\beta_0^i + \beta_1^i age_{jt} + \beta_2^i age_{jt}^2) - \log(1 + e^{\beta_0^i + \beta_1^i age_{jt} + \beta_2^i age_{jt}^2}) \right].$$

Starting from some initial parameter value $\boldsymbol{\psi}^{(0)}$, the E- and M-steps are repeated until convergence. In the M-step, there is no closed form solution for the evaluation of $\boldsymbol{\beta}$ so this maximization requires iteration. We can use optimization procedures such as Newton-Raphson or Quasi-Newton methods. Another alternative to these maximization methods is to fit a weighted logistic regression model and perform ML estimation via iteratively reweighted least squares (IRLS).

2.7.3 Weighted logistic regression and IRLS

We note that our longitudinal trajectory model is a mixture of weighted logistic distributions. Consider the estimation for component i. When we perform the maximization step in the EM algorithm for this group, we can estimate the parameters $\boldsymbol{\beta}^i = (\beta_0^i, \beta_1^i, \beta_2^i)$ by treating the model as a weighted logistic regression. That is, for each group i we have the model with log-likelihood written in the form

$$l(\boldsymbol{\psi}; \boldsymbol{y}) = \sum_{j=1}^{n} z_{ij} \left[\sum_{t=1}^{m} y_{it}(\beta_0^i + \beta_1^i age_{jt} + \beta_2^i age_{jt}^2) - \log(1 + e^{\beta_0^i + \beta_1^i age_{jt} + \beta_2^i age_{jt}^2}) \right].$$

Using matrix notation, we have the following parameters in our longitudinal model.

$$\boldsymbol{\beta}^i = \begin{pmatrix} \beta_0^i \\ \beta_1^i \\ \beta_2^i \end{pmatrix}_{3\times1}$$

where β^i describes the shape of the trajectory for a particular group. Let

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}_{mn\times3}, \boldsymbol{x}_j = \begin{pmatrix} 1 & age_{j1} & age_{j1}^2 \\ & \vdots & \\ 1 & age_{jm} & age_{jm}^2 \end{pmatrix}_{m\times3},$$

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix}_{mn\times1}, \boldsymbol{y}_j = \begin{pmatrix} y_{j1} \\ \vdots \\ y_{jm} \end{pmatrix}_{m\times1},$$

where $\boldsymbol{X}$ denotes the covariate (age) information for the n individuals over m time points, and $\boldsymbol{Y}$ denotes the binary responses of the n individuals observed over m time points.

The probability that individual j belongs to group i is denoted by z_{ij}. Let the vector $\boldsymbol{Z}$ represent the group membership probabilities for all n individuals, so that

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_n \end{pmatrix}_{mn\times1}, \boldsymbol{z}_j = \begin{pmatrix} z_{ij} \\ \vdots \\ z_{ij} \end{pmatrix}_{m\times1}.$$

Note that we are estimating the parameters for group i only, so that all m elements in $\hat{\boldsymbol{z}}_j$ are the same, namely $\hat{z}_{ij}$.

The logistic model is a generalized linear model with the logit link as the canonical link, that is,

$$\eta = logit(\boldsymbol{\mu}) = log\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right) = \boldsymbol{X}^T\boldsymbol{\beta}$$

and

$$\boldsymbol{\mu} = \frac{exp(\boldsymbol{X}^T\boldsymbol{\beta})}{1+exp(\boldsymbol{X}^T\boldsymbol{\beta})}.$$

For our longitudinal model, the parameter μ in matrix notation is

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_n \end{pmatrix}_{mn\times 1}, \mu_j = \frac{e^{x_j^T\beta}}{1+e^{x_j^T\beta}}.$$

McCullagh and Nelder (1989, pg 114-117) described the method for parameter estimation for binary data. We follow the same steps using the parameters defined above, but we fit the weighted logistic model rather than a classic linear logistic model. We need to consider the group membership probabilities of the individuals when we fit the model. We can incorporate them into the weight matrix, thus we have W as a diagonal matrix of weights given by

$$\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{\mu}(1-\boldsymbol{\mu})$$

and the score function becomes

$$\partial l/\partial\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Z}(\boldsymbol{Y}-\boldsymbol{\mu}).$$

We can estimate $\boldsymbol{\beta}$ using the iterative Fisher's scoring procedure, where at the $(t+1)^{th}$ stage we have

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + i^{-1}(\hat{\boldsymbol{\beta}}^{(t)})S(\hat{\boldsymbol{\beta}}^{(t)}),$$

where

$$i^{-1}(\hat{\boldsymbol{\beta}}^{(t)}) = \boldsymbol{X}\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(t)})\boldsymbol{X}^T$$

is the information matrix and

$$S(\hat{\boldsymbol{\beta}}^{(t)}) = \boldsymbol{X}\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(t)})(\boldsymbol{Y}-\boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(t)}))\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}$$

is the score function.

This is called the IRLS procedure because the weight matrix changes for each iteration, depending on the previous parameter estimates. We will use this procedure to perform the maximization (M-step) in the EM algorithm for each group. We start the IRLS procedure using the parameter estimates for this group i from the previous EM iteration, and repeat the iterative step until convergence produces MLE's of $\boldsymbol{\beta}$ for this particular group. We perform the same method to obtain parameter estimates for all the groups in our mixture model, and repeat the EM iteration steps until we reach the EM convergence criteria.

2.8 Limitations of ML estimation methods

Although ML is the most widely used estimation approach for mixture models, there are some practical difficulties associated with this type of estimation methods. Some of the common problems researchers may encounter when dealing with estimations of mixture models include issues related to model identification, convergence and sample sizes (Frühwirth-Schnatter, 2006). A model is non-identifiable when more than one set of parameter values correspond to the same model, such that there is no way of knowing which set of parameters contains the true values (Casella and Berger, 1990 pg 511). Model identification may be a problem for mixture modelling, as a g-component model may have $g!$ ways of assigning the g sets of parameters to g components, leading to a total of $g!$ equivalent solutions (Bishop, 2006 pg 434). Muthén and Muthén (2007) noted that not all growth mixture models are identifiable, and the Hessian matrix in a non-identifiable model may be singular. In this case, standard errors cannot be computed and estimation may not converge or may not produce interpretable estimates for all of the model parameters.

Convergence failures may also occur when mixture components are not well separated or when the sample size is small. There is no guarantee that the ML methods will fit a model successfully, as the estimation procedures may fail to find a solution or only converge to a local maximum instead of global maximum. For example, Finch et al. (1989) performed simulation studies on two-component normal mixtures, which are considered computationally easy compared to mixtures with more components, but their results showed that it was difficult to get the global maximum with a high degree of reliability. Also, mixture likelihoods may be unbounded and have many local spurious modes. In these situations, the search methods will usually converge to a local maximum rather than the global maximum (Frühwirth-Schnatter, 2006). Finally, McLachlan and Peel (2008) noted that sample sizes of mixture models have to be very large before asymptotic theory of ML can be applied. Working with mixtures with small data sets or small mixing proportions, or overfitting mixtures with too many components may lead to violation of regularity conditions. Mixture models are complex statistical models, and researchers need to be cautious when using ML estimations to fit mixture models.

Chapter 3

Simulation Study

3.1 Introduction

This research introduces a SAS/IML macro program that identifies trajectories by using the EM algorithm to fit mixtures of logistic distributions to longitudinal binary data. To try to speed up the EM convergence, we proposed the use of iteratively reweighted least squares (IRLS) to fit a weighted logistic regression model at the maximization stage. We performed simulation studies to investigate the properties of PROC TRAJ and EM-based algorithms under a variety of parameter combinations in mixtures with different numbers of components.

3.2 Data generation

Simulations were designed to compare six estimation algorithms when the population consists of two or three mixture components, fitting various trajectory shapes. Consider the trajectories shown in Figure 3.1, with time plotted against the probability of the group having the characteristics of interest. Suppose that we are interested in the probability of smoking for individuals. For each trajectory, the trend being described and the corresponding parameters are displayed in Table 3.1. The different trends reflect how an individual's smoking habit may change over time. For example, trajectory 1 shows how an individual may have tried to quit but then resumed to smoking regularly later on, while trajectories 3 and 4 show that individuals may become regular smokers at different rates. The simulation cases consisted of observations generated from the different combinations of the trajectories. We generated data involving five time points, denoted as age of individuals ($age_{j1} = 1, \ldots, age_{j5} = 5$), and we assumed that the responses (y_{jt}) were independent across time. The binary response for the j^{th} individual in group i at t^{th} time point was generated by a binomial distribution with probability of success as

$$p = \frac{e^{\beta_0^i + \beta_1^i age_{jt} + \beta_2^i age_{jt}^2}}{1 + e^{\beta_0^i + \beta_1^i age_{jt} + \beta_2^i age_{jt}^2}}.$$

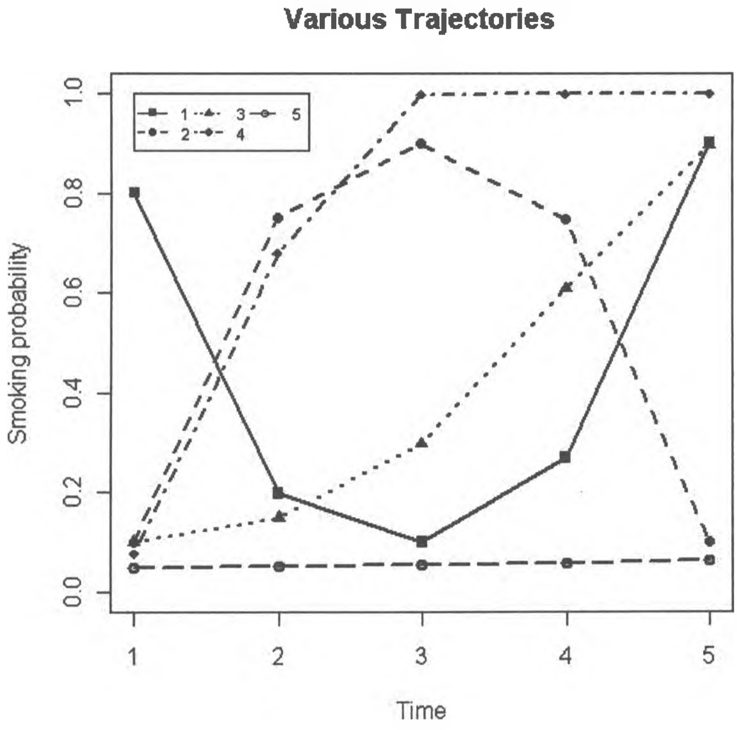


Figure 3.1: Trajectories designed for simulation

Table 3.1: Descriptions and parameter values for the various trajectories

		Parameters		
Trajectory	Description	β_0	β_1	β_2
1	Temporarily quitting then resumed smoking	6.170	-5.780	0.997
2	Stopped smoking	-7.690	6.590	-1.099
3	Gradual onset	-2.240	-0.170	0.210
4	Early onset	-3.050	-0.800	1.350
5	Never smoked	-3.000	0.010	0.010

3.3 Comparing different estimation algorithms

We considered six different algorithms for maximum likelihood estimation:

1. EM with the IRLS method (EM-IRLS):

   Given specific initial values, the estimation was performed using the EM algorithm, with the use of the IRLS method at the maximization step;

2. EM with mixed maximization method (EM-Mixed):

   Given specific initial values, the estimation was performed using EM with IRLS, and then switched to the EM-NLPQN if the IRLS estimation failed;

3. EM with the Quasi-Newton method (EM-NLPQN):

   Given specific initial values, the estimation was performed using the EM algorithm, with the use of the Quasi-Newton method to perform the maximization step;

4. PROC TRAJ 1:

   Estimation was performed using the SAS procedure PROC TRAJ with specified starting values;

5. PROC TRAJ 2:

   Estimation was performed using PROC TRAJ with the procedure's default starting values; and

6. Full maximization using Quasi-Newton method (FullMax):
   Given specific initial values, the estimation was performed using the Quasi-Newton approach. This is an attempt to replicate the algorithm being used by PROC TRAJ 1.

We refer to the PROC TRAJ and FullMax algorithms as full maximization approaches and the other three algorithms as the EM-based methods.

The FullMax and the EM-based algorithms were implemented as SAS macro programs using the SAS/IML language. The IRLS method within the maximization step of the EM-IRLS algorithm was programmed using code adapted from the example on logistic and probit regression for binary response models given in the SAS/IML 9.2 User's Guide (SAS, 2008). We have implemented the EM-IRLS using the portion of code corresponding to the logistic regression model; it is mathematically equivalent to the IRLS method described in Chapter 2. The EM-NLPQN method used the SAS/IML function NLPQN (nonlinear optimization by Quasi-Newton method) to perform the Quasi-Newton maximization (SAS, 2008), and FullMax was also implemented using this optimization subroutine.

The iterative estimation procedures stop when convergence is reached, which suggests that the log-likelihood reaching a maximum. PROC TRAJ stops iterating when the log-likelihood stops increasing or when it decreases. Following Nawa (2004), the EM algorithm was implemented to stop when the log-liklihood stops increasing (defined as having a difference in successive values of the log-likelihoods of 10^{-8}) or if it reaches a specified maximum number of 1000 iterations. When the EM algorithms failed to converge within 1000 iterations, the estimates obtained from the last iteration were taken to be the final estimate. PROC TRAJ often converges within small number of iterations (less than 100 iterations) and since FullMax performs the same maximization procedure as PROC TRAJ, we decided to use the default setting of 200 iterations in the optimization subroutine NLPQN as the maximum number of iterations for the FullMax algorithm. Standard errors of the estimates from the EM algorithms were calculated using a closed formula (Nawa, 2004 Section 3.2), whereas approximate standard errors of estimates from PROC TRAJ and FullMax were obtained as a by-product of the Quasi-Newton approach (Jones et al., 2001).

To analyze each case of simulated data, the same set of initial values was used for the different algorithms (except for PROC TRAJ 2, which used the procedure's own default starting values). We simplified the choice of starting values by only specifying the intercept component of the $\boldsymbol{\beta}$'s, that is, we specify the initial trajectories to be constant trajectories. The initial values used for the simulation cases are displayed in the corresponding tables in Appendices A and B. We considered using equal proportions as starting values for the mixing proportions, which is the default setting for PROC TRAJ. We compared the algorithms in terms of: the number of converged samples, estimated mean squared error of prediction (EMSEP) of the trajectories, mean number of iterations required until convergence, and run-time required.

1. Number of converged samples

It was discovered that the estimation methods did not always converge and sometimes produced unreasonable estimates of standard errors, i.e. very large values. For mixture modelling, model identification can be difficult. The term non-identified is used for models without reliable estimates for its parameters (Muthén and Muthén, 2007). When a model cannot be identified, standard errors cannot be computed due to a singular Fisher information matrix (i.e. non-invertible matrix). Another problem that may arise is obtaining estimates with large standard errors, which is associated with trajectories being over-parameterized (Nagin, 1999). Due to these issues, we have excluded from our results summary those realizations where variances were unable to be calculated, or estimated negative variances or standard errors larger than 100. We defined the remaining samples as the "converged samples" and we compared the number of such converged samples in each case across the different algorithms.

2. Estimation of the mixing proportions

To assess how well the algorithms could identify the distinct trajectories in each case, we can look at the estimated mixing proportions. If the estimated mixing proportions are different from the true values, this indicates that some observations in the samples have been misclassified into the wrong group. We calculated the relative errors of the mixing proportion estimates in each case, and generated box plots for display of the error characteristics. Relative errors (RE) are calculated as

$$RE = \frac{Estimate - Theoretical\ value}{Theoretical\ value}.$$

The ideal case would be $RE = 0$.

3. Estimated mean squared error of prediction (EMSEP)

We calculated the EMSEPs to reflect the closeness of the estimated shapes to the true shapes of the trajectories, rather than of the parameter estimates. We focused on the trajectory shapes because the practical meaning of trajectory modelling depends on how the trajectory describes the development of behaviour, and we note that different estimates of the parameters (intercept, linear and quadratic components) may describe the same shape of a trajectory. For example, consider two sets of parameter estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ as (-3.05, -0.8, 1.35) and (-5.80, 3.28, -0.02). The two groups of parameter estimates are distinctively different but they lead to the trajectory curves shown in Figure 3.2, which are describing the same developmental trend on the behaviour of interest. Since we want to use the trajectories to characterize behaviour changes over time, it was decided that we should focus on a measure for indicating the difference between the estimated and the theoretical trajectory curves.

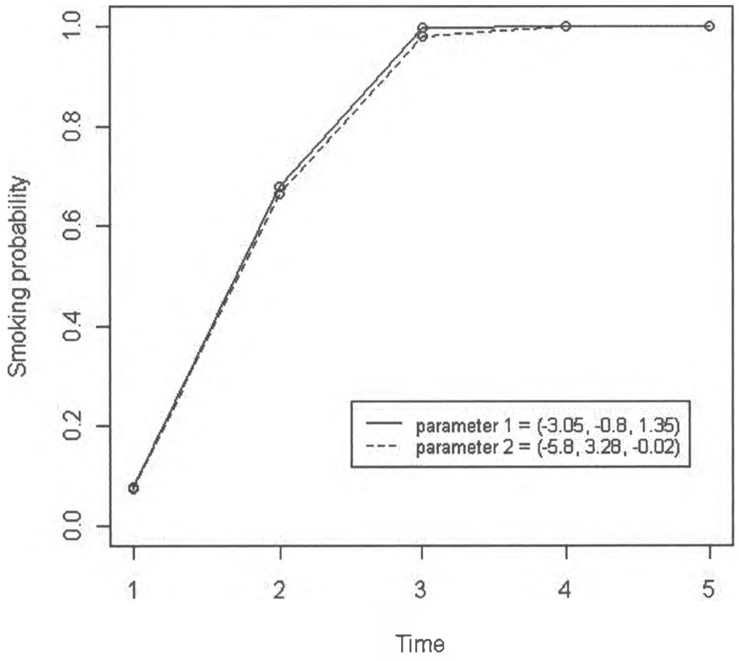


Figure 3.2: Example: Same trajectory shape may be described by different parameter values

Based on our definition of EMSEP as a measure of difference in trajectory shapes, the EMSEP based on one simulation case is given by

$$EMSEP = \frac{1}{n_s}\sum_{i=1}^{n_s}[(\hat{p}_{i1} - p_1)^2 + (\hat{p}_{i2} - p_2)^2 + \ldots + (\hat{p}_{i5} - p_5)^2],$$

where n_s is the number of converged samples in that simulation case, p_t denotes the theoretical smoking probability at time t and p_{it} denotes the estimated smoking probability at time t for the i^{th} sample.

4. Number of iterations

Since the estimations by the EM-based algorithms and the full maximization algorithms are based on different models (maximizing different log-likelihoods), it was not appropriate to compare the number of iterations across the two types of approaches. Thus we focused on the EM methods and only compared the number of EM iterations required to reach convergence for each case (averaged over the number of converged samples) across the three EM-based algorithms: EM-IRLS, EM-NLPQN and EM-Mixed.

5. Run-time for each case

PROC TRAJ is a compiled program written in C but the other algorithms were implemented using SAS/IML macro programs, so time efficiency is not a fair measure for comparison across all estimation algorithms. We therefore only evaluated the run-time required between the implemented programs of the EM-based algorithms and FullMax.

3.4 Results

3.4.1 Two-component mixtures

For mixtures of two components, we considered six different sets of parameter configurations and for each set we simulated 50 samples of 500 observations from a mixture with mixing proportions $\pi_1 = 0.32$ and $\pi_2 = 0.68$ (leading to 160 observations in group one and 340 observations in group two). This is similar to the data set simulated by Nawa (2004, Section 3.2) but fitting different combinations of trajectories. The trajectory groups being simulated for the six cases are displayed in Figure 3.3.

Converged samples

Table 3.2 shows the number of converged samples for the different methods, that the EM approach produced more acceptable results than PROC TRAJ and FullMax, with the EM-IRLS and EM-Mixed being the more reliable methods than EM-NLPQN. EM-Mixed was able to improve the result for case 2 by reaching convergence in one more sample than EM-IRLS. Across all six cases, the numbers of converged samples using FullMax appeared to be more stable than using EM-NLPQN or the two PROC TRAJ methods, in that it produced relatively high number of converged samples across the cases. Performances of EM-NLPQN and the two PROC TRAJ algorithms were comparable, with the PROC TRAJ procedures having more converged samples in some cases while EM-NLPQN was superior in other cases.

We note that the number of converged samples may be considered as a measure for how well the algorithms can estimate the parameters for each mixture

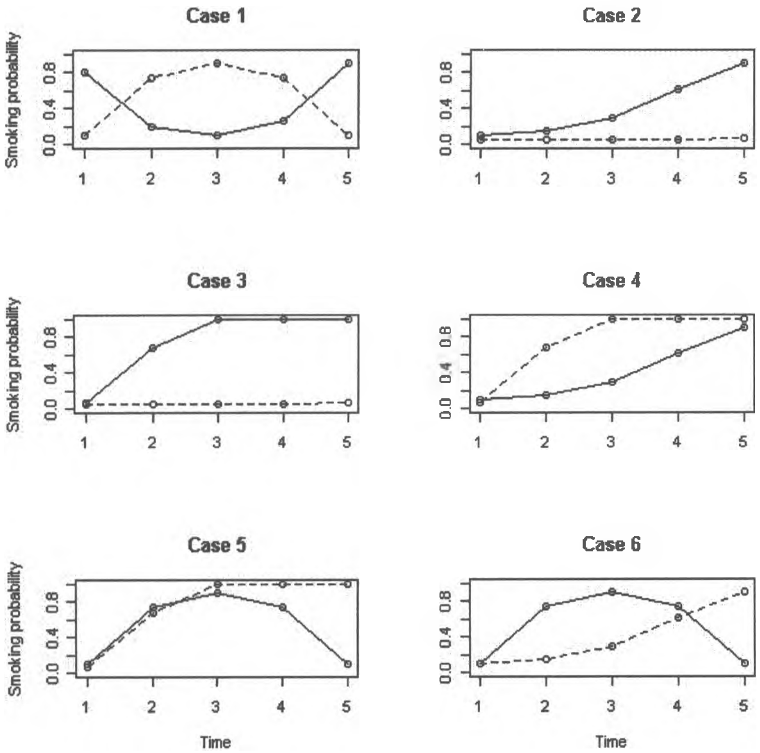


Figure 3.3: Mixtures of two components: Trajectories simulated in each case (Trajectory 1: solid line; Trajectory 2: dashed line)

component. Based on the values of the parameters describing the theoretical curves (see Table 3.1), we can expect that combinations of the gradual onset, early onset, and never smoking trajectories will be the most difficult situations for parameter estimation, since those three trajectories are described by parameter values that are very similar. Thus for the two-component mixtures, we would expect that the convergence properties of the algorithms would be worse in cases 2, 3 and 4 compared to the other two cases. The results in Table 3.2 indeed indicated that case 3 appeared to be the most difficult situation for the algorithms to model the trajectories correctly, especially for EM-NLPQN and the PROC TRAJ algorithms.

Table 3.2: Mixtures of two components: Number of converged samples

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
EM-IRLS	50	48	50	50	50	50
EM-Mixed	50	49	50	50	50	50
EM-NLPQN	50	47	14	34	21	50
PROC TRAJ 1	50	50	18	23	28	42
PROC TRAJ 2	50	50	19	25	28	46
FullMax	46	46	47	35	42	48

Parameter estimates

Table 3.3 shows the estimated mixing proportions from the different algorithms for all cases of two-component mixtures, averaged over the converged samples in each case. The results for case 1 were expected to be excellent, due to the very distinct trajectory shapes being simulated. Across the six cases being considered, we have expected the classification performance of algorithms to be worst for analyzing data in case 2 and case 5 due to the close resemblance in the trajectory shapes in these two cases. Case 4 was also considered to be a difficult case for identifying the two distinct trajectories since both trajectories were describing an increasing trend (i.e. describing the onset of behaviour).

By inspecting Table 3.3, we can see that the estimation results agree with what we have expected. The estimated mixing proportions have values furthest from the true values for case 4, which consisted of the two trajectories describing the onset of behaviour. Also as predicted, the algorithms classified some observations into the wrong groups in cases 2 and 5 as well. Estimates across the different algorithms do not differ much, but we can see that EM-IRLS, EM-Mixed, and FullMax had the estimates that are closer to the true values compared to other algorithms. If we focus on cases 4 and 5, FullMax appeared to have produced

Table 3.3: Mixtures of two components: Estimates of mixing proportions

Method	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Group 1	**0.32**					
EM-IRLS	0.3208	0.3390	0.3194	0.3511	0.3310	0.3231
EM-Mixed	0.3208	0.3374	0.3194	0.3511	0.3310	0.3231
EM-NLPQN	0.3208	0.3404	0.3189	0.3613	0.3408	0.3231
PROC TRAJ 1	0.3218	0.3447	0.3190	0.3575	0.3279	0.3194
PROC TRAJ 2	0.3218	0.3449	0.3198	0.3568	0.3279	0.3201
FullMax	0.3215	0.3413	0.3198	0.3381	0.3199	0.3215
Group 2	**0.68**					
EM-IRLS	0.6792	0.6610	0.6806	0.6489	0.6690	0.6769
EM-Mixed	0.6792	0.6626	0.6806	0.6489	0.6690	0.6769
EM-NLPQN	0.6792	0.6596	0.6811	0.6387	0.6592	0.6769
PROC TRAJ 1	0.6782	0.6553	0.6810	0.6425	0.6721	0.6806
PROC TRAJ 2	0.6782	0.6551	0.6802	0.6432	0.6720	0.6799
FullMax	0.6785	0.6587	0.6802	0.6619	0.6801	0.6785

estimates closest to the true values for the two mixing proportions, although they were averaged over only 35 and 42 converged samples for the two cases respectively (see Table 3.2 for number of converged samples). All the estimates were within the 95% confidence interval, which were (0.19, 0.45) and (0.55, 0.81) for the mixing proportions 0.32 and 0.68 respectively. We note that these confidence intervals are wide due to the sample size of 50. If the sample size was doubled, the corresponding confidence intervals would become (0.23, 0.41) and (0.59, 0.77). The estimates would still fall within this narrower 95% confidence interval for the sample size of 100, resulting to the same conclusions with the assumption of observing similar results from the algorithms.

The box plots showing the relative errors for the estimates of mixing proportions (described in section 3.3) are displayed in Appendix A.1. The box plots for the mixing proportion estimates across the different cases show that the relative errors for the estimates of mixing proportions by FullMax are distinctively far away from zero in all cases. Across all the simulation cases of two-component mixtures, FullMax consistently over-estimated the mixing proportions of group 1. This means the FullMax algorithm estimated the observations as more equally distributed among the two mixtures (that is, mixing proportions close to 0.5) than they actually were. The inter-quartile range (H-spread) of a box plot represents the middle 50% of the data; for FullMax, this range of data is often further away from zero than the outliers produced by the other methods. Hence, it cannot be concluded that FullMax has better classification performance than the other algorithms. For the other estimation algorithms, PROC TRAJ 1 and 2 had wider H-spread for cases 2, 4 and 6 but they performed best for case 1. Note that the EM-based algorithms produced outliers for case 5 while the PROC TRAJ methods did not; however the PROC TRAJ methods only converged for 28 samples while the EM-IRLS and EM-Mixed converged for all 50 samples in this simulation case.

The parameter estimation results from the different algorithms are presented in Appendix A.2. Except for FullMax, the estimates obtained from all other methods were close to the theoretical ones and estimated the correct shapes of the trajectories, with the exception of the early onset trajectory. If we focus on the early onset trajectory in the simulation cases 3, 4 and 5, we can see that all methods produced estimates far away from the theoretical values, with standard errors larger than those of other trajectory estimates. However, although the estimates were not similar to the theoretical values, they still described the same trajectory shape as the theoretical curve. For FullMax, the estimates for this trajectory curve were very different from the estimates obtained by the other methods in all three cases. FullMax was able to produce estimates very close to the theoretical values for the early onset trajectory in case 4, but for the other two cases it obtained estimates that are quite different from both the true values and the results from the other methods.

Estimated mean squared error of prediction (EMSEP)

Table 3.4 summarizes the EMSEP values for the simulation cases. The EMSEPs we present here are summed over the two trajectories in each case. The EM-based algorithms outperformed the full maximization algorithms in most cases except for case 5. FullMax produced estimates that described trajectories with shapes rather different from the true trajectories for all cases, especially worse for cases 1 and 3. The other algorithms had the most trouble identifying the shapes of the trajectories in case 2, but the EMSEPs for this case were still very small, indicating that the algorithms were able to estimate trajectories that were very similar to the correct trajectory shapes even in the worst case.

Table 3.4: Mixtures of two components: EMSEP

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
EM-IRLS	0.005	0.012	0.002	0.009	0.011	0.006
EM-Mixed	0.005	0.012	0.002	0.009	0.011	0.006
EM-NLPQN	0.005	0.012	0.002	0.011	0.012	0.006
PROC TRAJ 1	0.005	0.015	0.002	0.010	0.004	0.007
PROC TRAJ 2	0.005	0.015	0.002	0.010	0.004	0.007
FullMax	4.616	2.122	6.118	1.862	1.797	2.674

To summarize the accuracy of the different estimation algorithms, Figure 3.4 shows the plot of number of converged samples and EMSEPs for all six cases and all six algorithms. The ideal situation would be to obtain all the points in the upper left region of the plot, indicating high number of converged samples and EMSEPs close to zero. However, we see that the points representing results from

FullMax are spread out across the top of the plot, showing the varying EMSEPs obtained. All other algorithms produced estimates with very small EMSEPs across various numbers of converged samples. EM-IRLS and EM-Mixed had the best performances; points representing these two algorithms are concentrated in the upper left corner, showing low EMSEPs with large numbers of converged samples.

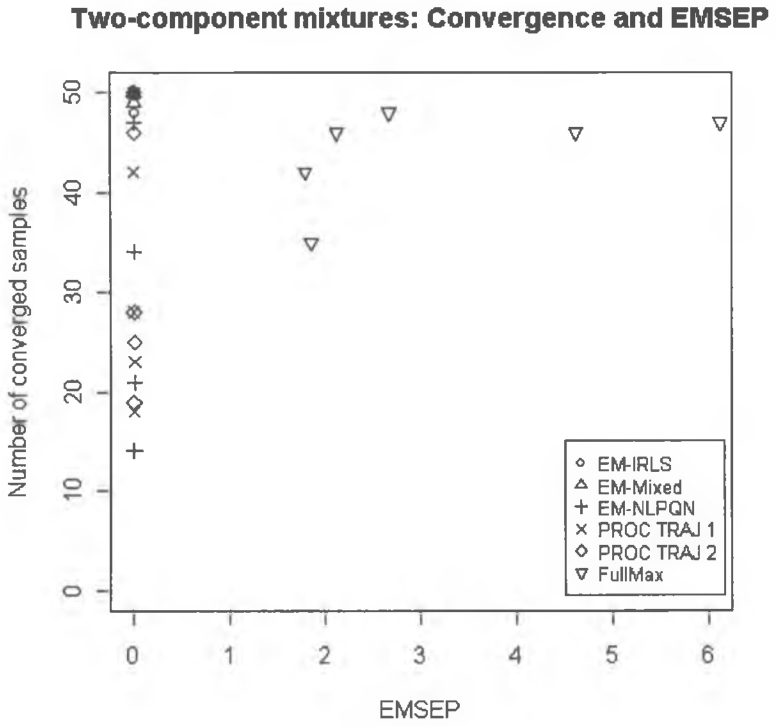


Figure 3.4: Mixtures of two components: Convergence and EMSEP

EM iterations and run-time

The number of EM iterations and run-time required for the different algorithms are presented in Table 3.5 and Table 3.6. We note that in most cases, EM-NLPQN requires less number of iterations but more time to reach convergence when compared to EM-IRLS and EM-Mixed. For the samples in case 3, EM-NLPQN required very small number of iterations to converge (approximately 21 iterations), while the other two EM methods required the maximum number of iterations (1000 iterations) to reach convergence in a lot of the samples, thus EM-NLPQN obtained estimates for this case much quicker than the other two EM methods. The full maximization algorithms can produce estimates faster than the EM methods (Roeder et al., 1999); Table 3.6 shows that FullMax was much faster than the EM-based algorithms in parameter estimation.

Table 3.5: Mixtures of two components: EM iterations required

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
EM-IRLS	36.12	468.42	955.18	778.12	938.82	110.66
EM-Mixed	36.12	496.9	955.18	778.12	938.82	110.66
EM-NLPQN	29.46	301.91	21.07	215.29	252.52	80.96

Table 3.6: Mixtures of two components: Run-time required in (hrs):mins:secs

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
EM-IRLS	29:48	6:50:03	12:42:23	10:49:54	13:18:40	1:28:26
EM-Mixed	29:53	8:16:45	12:44:20	10:51:50	13:22:09	1:28:49
EM-NLPQN	1:37:47	15:03:10	1:26:59	11:09:45	13:06:26	4:10:43
FullMax	27:32	52:37	49:35	44:40	56:49	45:09

3.4.2 Three-component mixtures

We considered 5 different sets of parameter configurations for the three-component mixtures and the trajectories being simulated for each case are presented in Figure 3.5. For each case, we simulated 50 sets of 800 observations from a three group model with proportions $\pi_1 = 0.2$, $\pi_2 = 0.425$ and $\pi_3 = 0.375$ (leading to 160 observations in group one, 340 in group two and 300 in group three). This follows from Nawa's (2004, Section 3.2) data generation procedure.

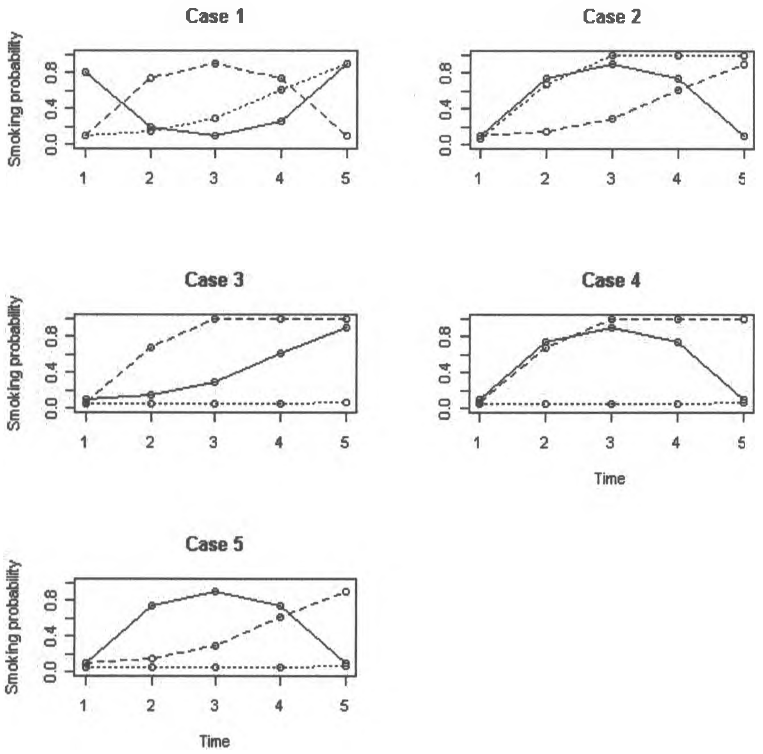


Figure 3.5: Mixtures of three components: Trajectories simulated in each case (Trajectory 1: solid line; Trajectory 2: dashed line; Trajectory 3: dotted line)

Converged samples

We expected case 1 to be the easiest case for the algorithms to handle due to the distinct shapes of the trajectories being simulated. Case 3 was expected to be the hardest case for algorithms to reach convergence since it consisted of the combination of the three trajectories with very similar parameter values: the early onset, gradual onset, and never smoking groups. All other cases included combinations of any two of these three trajectory shapes, thus the performances of the algorithms in such cases were predicted to be worse than those in case 1.

The number of converged samples in each case are presented in Table 3.7. As expected, case 1 had the most number of converged samples across all algorithms, and most algorithms performed worst in case 3. Across all cases, EM-IRLS and EM-Mixed have the best performance, while PROC TRAJ 2 and FullMax generally performed worse than other methods. It is unclear why EM-IRLS and EM-Mixed converged in more number of samples in case 3 compared to case 2, but they still performed well and was able to reach convergence in at least 43 out of the total 50 samples. The performances of FullMax in case 1 and case 5 were much worse than those of the other algorithms, with only 15 and 13 converged samples respectively. Overall, it is noted that EM-based algorithms performed better than the full maximization algorithms in terms of number of converged samples.

Table 3.7: Mixtures of three components: Number of converged samples

	Case 1	Case 2	Case 3	Case 4	Case 5
EM-IRLS	50	43	47	50	50
EM-Mixed	50	43	48	50	50
EM-NLPQN	50	35	31	27	47
PROC TRAJ 1	44	21	21	24	43
PROC TRAJ 2	24	16	9	21	37
FullMax	15	17	25	22	13

Parameter estimates

Next, we examined the estimates of the mixing proportions for each case to evaluate each algorithm's ability to classify the observations into the correct trajectory groups. Table 3.8 shows the results from the six algorithms across the five cases. Despite its good performance in the two-component mixtures, it is clear that FullMax was not able to handle the more complex three-component mixtures. All other algorithms performed well, but small amounts of misclassifications occurred in cases 2, 3 and 5 similarly across the five algorithms. For case 2 and case 5, the algorithms were more likely to misclassify observations between groups 1 and 2, while for case 3 the algorithms misclassified observations between groups 2 and 3. FullMax produced estimates of the mixing proportion for group 1 falling outside the 95% confidence interval (0.09, 0.31) for cases 1 and 5. If the sample size was doubled, the estimate of group 1 proportion for case 2 will also be outside the 95% confidence interval (0.12, 0.28) for a sample size of 100. The 95% confidence intervals for the groups 2 and 3 mixing proportions for 50 samples are (0.29, 0.56) and (0.24, 0.51) respectively. The estimates for the groups 2 and 3 mixing proportions produced by all algorithms were within these 95% confidence intervals, as well as the intervals for a sample size of 100, which are (0.33, 0.52) and (0.28, 0.47) for the groups 2 and 3 proportions.

The box plots showing the relative errors of the estimated mixing proportions (described in section 3.3) are shown in Appendix B.1. We can see that besides FullMax, the relative errors of estimates from all other algorithms have medians close to zero but varying H-spread widths in most cases. In some cases, the relative errors from EM-based methods had narrower H-spreads, while in other cases the PROC TRAJ methods produced relative errors with smaller H-spreads. For case 1, the algorithms (except for FullMax) had wide H-spreads for the mixing proportions of groups 1 and 3 but narrow H-spreads for the group 2 proportion. Excluding FullMax, PROC TRAJ 2 had relatively wide H-spreads for cases 1 and 2 while EM-NLPQN had wide H-spreads for cases 2 and 3 when compared to the other algorithms. We can see that almost all algorithms produced outliers in case 2, indicating that it was a difficult combination of trajectories for the

Table 3.8: Mixtures of three components: Estimates of mixing proportions

Method	**Case 1**	**Case 2**	**Case 3**	**Case 4**	**Case 5**
Group 1	**0.200**				
EM-IRLS	0.2043	0.222	0.1994	0.2015	0.2212
EM-Mixed	0.2043	0.222	0.2005	0.2015	0.2212
EM-NLPQN	0.2043	0.2433	0.1991	0.2008	0.2241
PROC TRAJ 1	0.1985	0.2406	0.1832	0.2076	0.2265
PROC TRAJ 2	0.2484	0.2294	0.2013	0.1986	0.2305
FullMax	0.0732	0.1089	0.2365	0.1835	0.5037
Group 2	**0.425**				
EM-IRLS	0.4224	0.4144	0.4418	0.4318	0.4025
EM-Mixed	0.4224	0.4144	0.4413	0.4318	0.4025
EM-NLPQN	0.4223	0.4143	0.442	0.4344	0.3994
PROC TRAJ 1	0.4251	0.3957	0.4429	0.4312	0.4041
PROC TRAJ 2	0.4222	0.36	0.4425	0.4254	0.4031
FullMax	0.4476	0.5492	0.3851	0.4317	0.1515
Group 3	**0.375**				
EM-IRLS	0.3734	0.3636	0.3588	0.3667	0.3763
EM-Mixed	0.3734	0.3636	0.3582	0.3667	0.3763
EM-NLPQN	0.3734	0.3424	0.3589	0.3648	0.3766
PROC TRAJ 1	0.3764	0.3638	0.3739	0.3612	0.3694
PROC TRAJ 2	0.3296	0.4106	0.3562	0.376	0.3665
FullMax	0.4792	0.3419	0.3784	0.3848	0.3448

algorithms to identify. Across all cases, FullMax had relative errors with medians further away from zero and with H-spreads wider than those of the other algorithms. For some of the simulation samples, FullMax could only identify two components instead of three, thus leading to such unreasonable mixing proportion relative errors estimates.

The parameter estimates from the different algorithms are presented in Appendix B.2. FullMax was unable to provide estimates similar to the theoretical values, leading to incorrect trajectory shapes being estimated. All other algorithms performed well, with parameter estimates close to true values, except for the early onset trajectories in cases 2, 3 and 4. Similar to the results obtained for the two-component mixtures, the algorithms (except FullMax) produced biased parameter estimates for the early onset trajectory, but the estimated parameters described the same trajectory shape as the true curve. The distances between the estimated and the true trajectories can be described using the EMSEPs.

Estimated mean squared error of prediction (EMSEP)

From the EMSEPs shown in Table 3.9, we can see that the EM-based algorithms and PROC TRAJ 1 performed better than the other two full maximization algorithms across all cases compared to the full maximization algorithms. The EM methods were able to estimate trajectories closer to the true curves in all cases except case 2. All algorithms had trouble modelling the trajectories in case 2 in terms of the number of converged samples, but the EM algorithms had approximately twice as many converged samples than the PROC TRAJ 1 method (see Table 3.7).

Table 3.9: Mixtures of three components: EMSEP

	Case 1	Case 2	Case 3	Case 4	Case 5
EM-IRLS	0.041	0.398	0.016	0.006	0.023
EM-Mixed	0.041	0.398	0.016	0.006	0.023
EM-NLPQN	0.041	0.638	0.017	0.006	0.023
PROC TRAJ 1	0.110	0.021	0.201	0.006	0.070
PROC TRAJ 2	0.300	0.743	0.239	0.304	0.201
FullMax	3.553	3.391	1.930	1.715	1.754

The accuracy of the estimations in terms of number of converged samples and EMSEPs are shown in Figure 3.6. Compared to the simulation cases of two-component mixtures (see Figure 3.4), the algorithms obtained less precise estimations in the three-component mixture cases. FullMax had the worst performance among all algorithms, as the FullMax converged in low number of samples and produced EMSEPs far from zero in all cases. For the three-component mixture cases, the PROC TRAJ procedures resulted in estimates with larger EMSEP values compared to the estimates from two-component mixtures. The points representing EM-IRLS and EM-Mixed are close to the upper left region, indicating their excellent performance in terms of both small EMSEPs and high convergence.

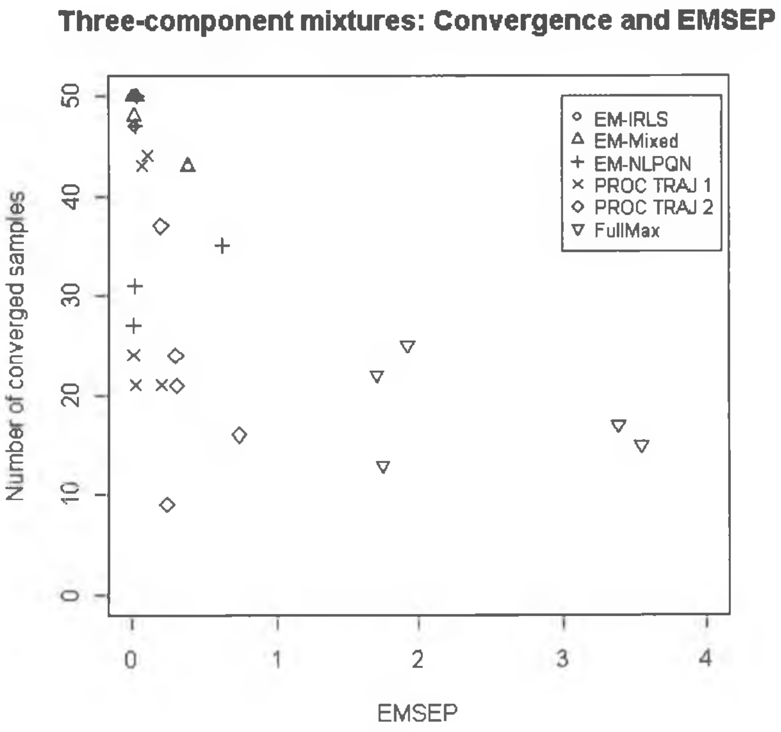


Figure 3.6: Mixtures of three components: Convergence and EMSEP

EM iterations and run-time

The results regarding number of EM iterations and run-time required are summarized in Table 3.10 and Table 3.11, and the same trends are observed here as those seen in the two-component mixture cases. Across the EM-based algorithms, EM-IRLS required the least amount of time for parameter estimation but required more iterations than EM-NLPQN. FullMax was much faster than the EM algorithms but its performance based on the other characteristics (number of converged samples and EMSEPs) showed that it is not a reliable method despite its speed. Overall, all algorithms required more time to perform parameter estimation in the three-component mixtures than the two-component cases. This is expected, as the parameter estimation process becomes more complicated when the model complexity increases (an increase in parameters and uncertainty in group allocation).

Table 3.10: Mixtures of three components: EM iterations required

	Case 1	Case 2	Case 3	Case 4	Case 5
EM-IRLS	829.56	955.74	941.09	963.80	587.74
EM-Mixed	829.56	955.74	942.31	963.80	587.74
EM-NLPQN	610.92	697.91	488.35	159.07	374.40

Table 3.11: Mixtures of three components: Run-time required in hrs:mins:secs

	Case 1	Case 2	Case 3	Case 4	Case 5
EM-IRLS	25:24:22	31:53:46	31:35:59	34:37:58	18:20:17
EM-Mixed	26:53:30	39:59:43	41:14:44	34:38:12	18:21:25
EM-NLPQN	69:09:48	81:46:43	53:40:38	19:20:46	44:30:29
FullMax	5:18:08	8:24:44	4:42:09	6:41:07	4:37:02

3.5 Summary

The complexity of mixture models increases as the number of components increases. This is shown through our simulation results, as the performances of all methods were more stable in the two-component mixtures compared to the three-component mixtures. In general, the full maximization algorithms (FullMax and PROC TRAJ 1 and 2) are faster than the EM-based algorithms (EM-IRLS, EM-Mixed and EM-NLPQN) in parameter estimation. However, the convergence properties of the EM-based methods are much more dependable, in that they converged for high number of samples in most cases. Results showed that the FullMax algorithm was not able to handle the more complex cases of three-component models.

In terms of the number of converged samples, the EM-based algorithms outperformed the full maximization algorithms in almost all the simulation cases. This is an important characteristic of the algorithms since it reflects how dependable the estimation methods are in regards to trajectory modelling. EM-IRLS converged for 43 out of the total 50 samples in the worst situation (case 2 of the three-component mixtures). However, the full maximization methods obtained lower numbers of converged samples in some cases, such as PROC TRAJ 1 converging for only 18 samples (case 3 of the two-component mixtures) in the worst situation, and PROC TRAJ 2 and FullMax had even worse performances in terms of convergence.

Based on the converged samples, all algorithms except FullMax were able to obtain the correct estimation results. The focus is on the estimated trajectory shapes rather than on the estimated parameters $\boldsymbol{\beta}$ because different parameters may describe the same trajectory curve. Unlike the coefficients from linear regression, the parameter estimates do not have interpretation by themselves, one must look at the curves described by the parameters to understand the changes in the behaviour of interest. The EMSEPs showed that EM-based algorithms and PROC TRAJ were able to estimate the correct trajectory shapes even when some parameter estimates were quite different from the true parameter values. However, FullMax was unable to achieve the same results, because their estimated trajectories were very far from the true curves in some cases.

The full maximization algorithms need to perform the optimization of the log of sum, which is difficult to evaluate (see, for example, Bilmes (1998 pg 3)). This may explain why the FullMax and PROC TRAJ methods have difficulties with the more complex mixture models. Although the FullMax estimation and PROC TRAJ were supposed to be using the same algorithm, sometimes they produced distinctively different results. For the FullMax algorithm, the results obtained for the two-component mixtures were remarkably different from those for the three-component mixtures. For the two-component mixtures, FullMax produced results with relatively high number of converged samples but with large EMSEPs. For the three-component mixtures, FullMax converged to the false maximum in many situations, and for some samples the procedure could only identify two mixture components instead of three. Note that the default maximum of 200 iterations was used as one of the stopping criteria for the FullMax algorithm; it may be able to produce better estimates if more iterations were allowed. However, since the PROC TRAJ procedures often converged within 100 iterations in the simulation cases; there is no strong evidence for this suggestion.

With regards to the maximization method in the EM framework, simulation results showed that the EM-IRLS algorithm was able to obtain correct convergence in a less amount of time compared to the EM-NLPQN method. It is noted that the Quasi-Newton method is slow to converge but each iteration step is scaled to ensure an increase in the optimization function. Also, at each M-step of an EM iteration, the SAS function NLPQN call would perform a maximum of 200 iterations by default; while the IRLS was implemented such that a maximum of 20 iterations would be performed. These reasons may explain why EM-NLPQN required less number of iterations but more time to converge when compared to EM-IRLS in most situations.

It was anticipated that EM-NLPQN would have good performance but slow speed, while the performance of the faster algorithm EM-IRLS was unsure. The EM-Mixed approach was implemented such that it may improve the performance of EM algorithm in the sense of allowing the failed samples from EM-IRLS to have a second chance. From our simulation results, we did observe improvements in some cases when EM-Mixed algorithm was used, although the method was not necessarily needed in most of the simulation cases due to the excellent performance of the EM-IRLS method. The poor behaviour of EM-NLPQN in the simulation results was unexpected; namely it produced a small number of converged samples in some simulation cases.

One concern of our simulation study is the sample size. The confidence level of simulation output depends on the size of data set; the larger the number of runs, the higher is the associated confidence. In our simulation study, increasing the simulation runs would have allowed for narrower confidence intervals of our estimates. However, larger simulation sample sizes also require more effort and resources. With the use of a single computer, our current sample size of 50 required the systems to run non-stop for approximately 31 days to finish all the simulation cases. Doubling the sample size to 100 would required approximately two months to complete the simulation study. Future simulation studies may consider increasing the sample size and compare the performances of the algorithms with different parameters and starting values.

In conclusion, our simulation results showed that EM-IRLS is a reliable method with better convergence and estimation properties than other algorithms. On average, the FullMax algorithm required 25% of the time required for EM-IRLS and EM-Mixed, while the EM-NLPQN algorithm required almost the same time or even more time than the other two EM-based algorithms. Although the full maximization methods were faster in parameter estimation, we may not be able to draw conclusions from their results due to having non-identified models. Compared to the full maximization algorithms, the EM-based algorithms have superior performance in terms of convergence. The PROC TRAJ 1, PROC TRAJ 2, and FullMax algorithms converged for only 66%, 57% and 63% of the samples respectively, while EM-IRLS and EM-Mixed converged for over 98% of the samples and EM-NLPQN converged for 74% of samples on average. For future trajectory modelling research, the use of the EM-IRLS algorithm is recommended in order to avoid convergence problems and produce precise estimations.

Chapter 4

Application: smoking data

4.1 Introduction

We have applied the group-based trajectory modelling methods discussed in previous chapters to the data set from the Third Waterloo Smoking Prevention Project (WSPP3) (Brown et al., 2002). The main purpose of our analysis was to determine the number of distinct smoking trajectories to allow for profiling the characteristics of the identified groups. In this chapter, we will give a short description of the study design of the WSPP3; a more detailed description of the study is given by Driezen (2001). Then we will discuss the results obtained by applying the longitudinal trajectory model to identify the different smoking trajectories within the study sample.

4.2 Description of the longitudinal study

The objective of the longitudinal study WSPP3 was to evaluate a high school tobacco control intervention program (Brown et al., 2002). The study followed a cohort of more than 4000 students over a seven year period (1990-1996), and examined their smoking behaviour and the long term effectiveness of the smoking prevention program. The study was carried out in three phases: evaluate the social influences smoking prevention program at the elementary school level (grades 6 to 8), at the secondary school level (grades 9 and 10), and then a follow-up assessment when students were in grades 11 and 12.

The first phase of the study consisted of a randomized trial with 100 elementary schools from seven school boards. The goal of this phase was to evaluate the effectiveness of self-preparation materials and having teachers as the providers of the social influences programs. Six of the school boards agreed to continue their participation into the second phase of the study, and 30 schools were eligible and willing to take part in this next phase. In the second phase, the schools were matched within school board and then randomized within pairs to intervention and control groups. The intervention program was provided to grade 9 and grade 10 students within the intervention schools, and the program consisted of involving as many students as possible in smoking prevention and cessation activities (Brown et al., 2002). The final phase of the study consisted of a follow up survey of the participated students when they were in grade 11 and grade 12, in order to assess the long term impacts of the interventions provided (Driezen, 2001).

4.3 Description of the data set

We considered the data set that was used by Driezen (2001) to identify regular smoking trajectories. The sample consisted of 2495 students for whom smoking status was recorded at all seven time points, but may have missing age information. This sample included students who did not smoke at baseline (grade six) as well as students who reported as smokers. Smoking status was originally recorded into five categories: never smoker, tried once, quitter, experimental, or regular. Following Driezen (2001), we wanted to focus on how youths develop the habit of smoking regularly, thus smoking status at each time point was dichotomized as regular or non-regular (never smoker, tried once, quitter or experimental). Table 4.1 summarizes the number of students in each category at each time point, and we can see that the proportion of regular smokers increases as the students grew older.

Table 4.1: Number of non-regular and regular smokers at each time point

Grade	Non-regular smoker	Regular smoker	% Regular smoker
6	2483	12	0.48
7	2463	32	1.28
8	2349	146	5.85
9	2151	344	13.79
10	1893	602	24.13
11	1775	720	28.86
12	1649	846	33.91

4.4 Longitudinal trajectories

The current analysis aims to identify developmental trajectories of smoking in a sample of adolescents. We estimated several trajectory models based on the 2495 students, and each model used a quadratic term in age to describe the relationship between age and youths' smoking behaviours. We applied the EM-IRLS, PROC TRAJ, and FullMax algorithms to fit the three-, four-, five- and six-component models without covariates. We used BIC for model selection, choosing the model with the maximum BIC as the optimal model for each algorithm.

Logistic regression starting values proposed by Nawa (2004) were used to start the estimation procedures; he has shown that these starting values may reduce convergence problems. The procedure for obtaining the starting values consisted of dividing the data into groups according to responses at some chosen time points, then fitting a logistic regression model in each groups and use the obtained parameter estimates as initial values to start the trajectory modelling algorithms. The method is given in detail by Nawa (2004, Section 3.2.4).

The logistic regression starting values were obtained and used to start the EM-IRLS and FullMax algorithms for the different component models. However, floating point exceptions occurred in PROC TRAJ when Nawa's starting values were specified, so the PROC TRAJ analyses were performed using the procedure's default starting values. It was not necessary to use the stringent convergence criterion of having log-likelihood values correct to five decimal places; therefore, the EM-IRLS convergence criterion was changed to 10^{-3}, meaning the iterative procedure would stop when the difference between successive log-likelihood values was less than 10^{-3}.

4.4.1 Models

Three group model

Figure 4.1 shows the results from fitting a mixture of three components using EM-IRLS, PROC TRAJ and FullMax respectively. The corresponding mixing proportions are shown within each figure; the results obtained by EM-IRLS and PROC TRAJ were very similar. Although FullMax obtained the same trajectory shapes, the mixing proportions obtained were different from the other two. Also, FullMax obtained negative variances for the estimates in the model, indicating that it is a non-identifiable model and conclusions cannot be made based on these estimated trajectories. The EM-IRLS and PROC TRAJ models showed that about 69% of students remained as non-smokers throughout the seven years.

About 12% of students started smoking early while 19% of the students had a later smoking onset, but both groups escalated to regular smoking by the age of seventeen.

Four group model

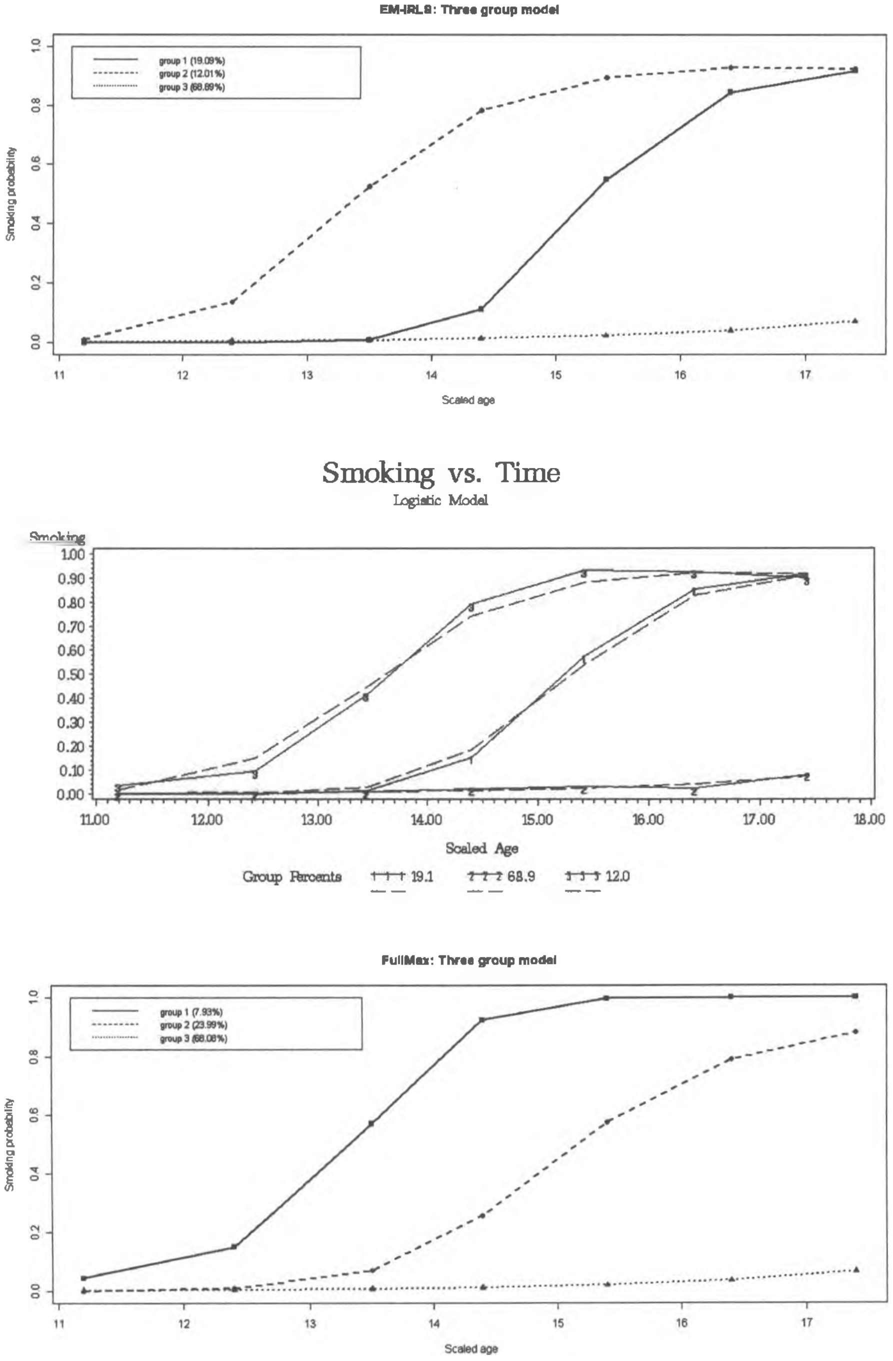
The four group models fitted using the three algorithms are presented in Figure 4.2. Note that the trajectories and mixing proportions estimated by PROC TRAJ and FullMax appeared to be the same, but standard errors could not be calculated for the parameter estimates (FullMax obtained a singular observation matrix). The model fitted by EM-IRLS showed that three distinct smoking trajectories were identified, with different smoking onset patterns. The largest group was the non-smoker group of students, consisted of approximately 69% of the study sample. The two smoking groups from the three-component model were split into the three smoking trajectories shown in this model, and the early onset trajectory showed that only 3.6% of the students started smoking at the early age of twelve.

Five group model

The five group models fitted are displayed in Figure 4.3. PROC TRAJ and FullMax had the similar problem of being unable to calculate the standard errors, with FullMax producing negative variances for the parameter estimates. FullMax obtained the same trajectories and mixing proportions as those estimated by EM-IRLS, but PROC TRAJ estimated two non-smoking groups of students (the two groups were split from the non-smoking group estimated in the four group model). The five group model fitted by EM-IRLS is of interest, as the additional trajectory in this model compared to the four-component model showed a group of students with decreased smoking probability by the age of seventeen. Although this group of "quitters" consisted of only 5.6% of the study population, it is of public health interest to characterize this group of students to better understand the factors that led them to smoking reduction.

Six group model

Figure 4.4 shows the six group models fitted by the three algorithms. Again, PROC TRAJ and FullMax suffered from the problem of non-identifiable models, with FullMax obtaining negative variances for the estimates. PROC TRAJ was only able to identify four distinct trajectories while FullMax estimated two non-smoking trajectories in this model. EM-IRLS produced six distinct trajectories, with the largest group of student as the non-smokers group and the smallest group being the early onset smokers. Compared to the five-component model, the group of "quitters" in this model remained as approximately 5.5% of the study population and the additional group identified was a group of late onset smokers, consisting of about 7.14% of the study population.



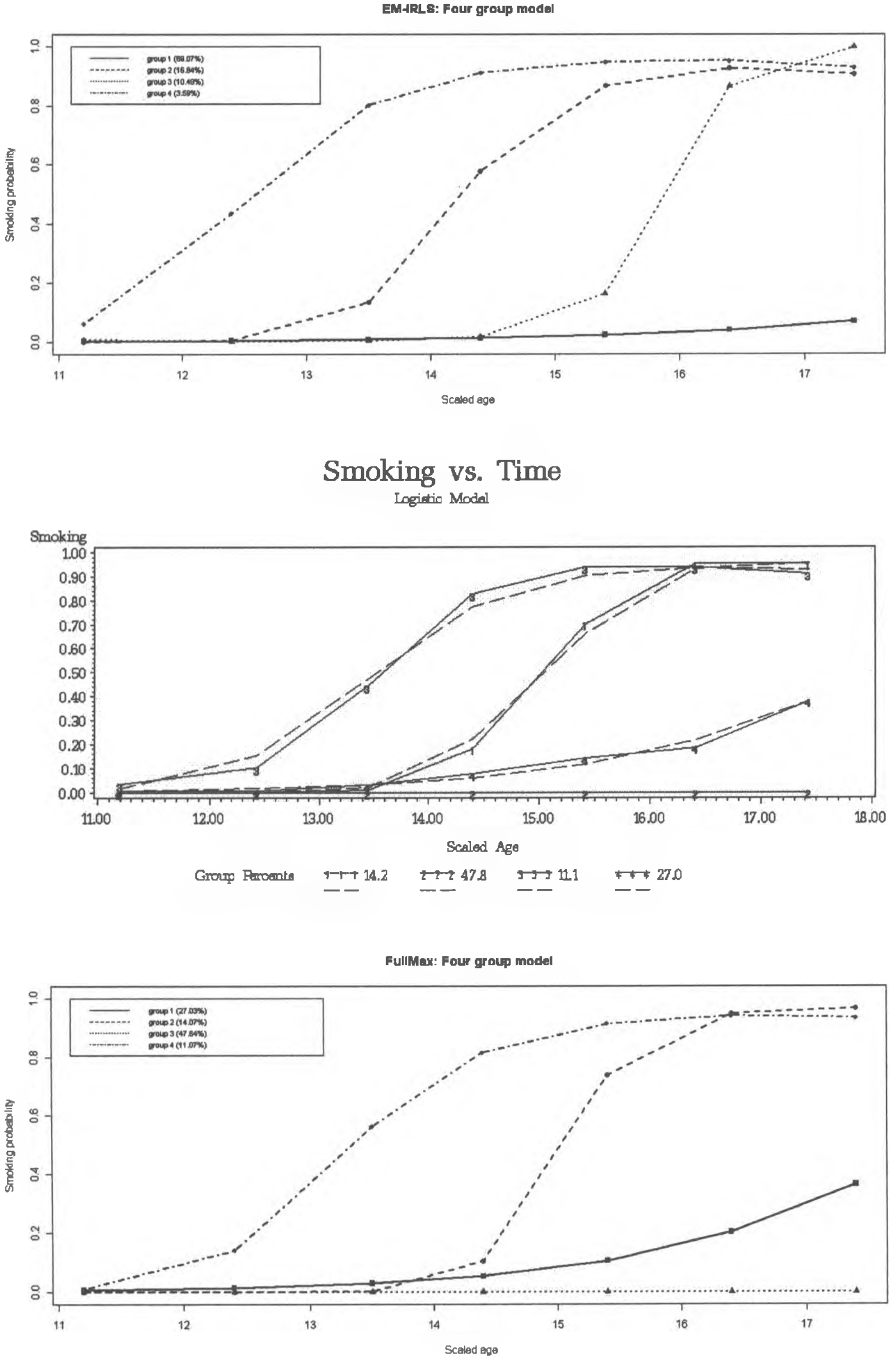
EM-IRLS: Three group model

Smoking vs. Time

Logistic Model

FullMax: Three group model

Figure 4.1: Three group model fitted by EM-IRLS, PROC TRAJ, and FullMax



EM-IRLS: Four group model

Smoking vs. Time
Logistic Model

FullMax: Four group model

Figure 4.2: Four group model fitted by EM-IRLS, PROC TRAJ, and FullMax

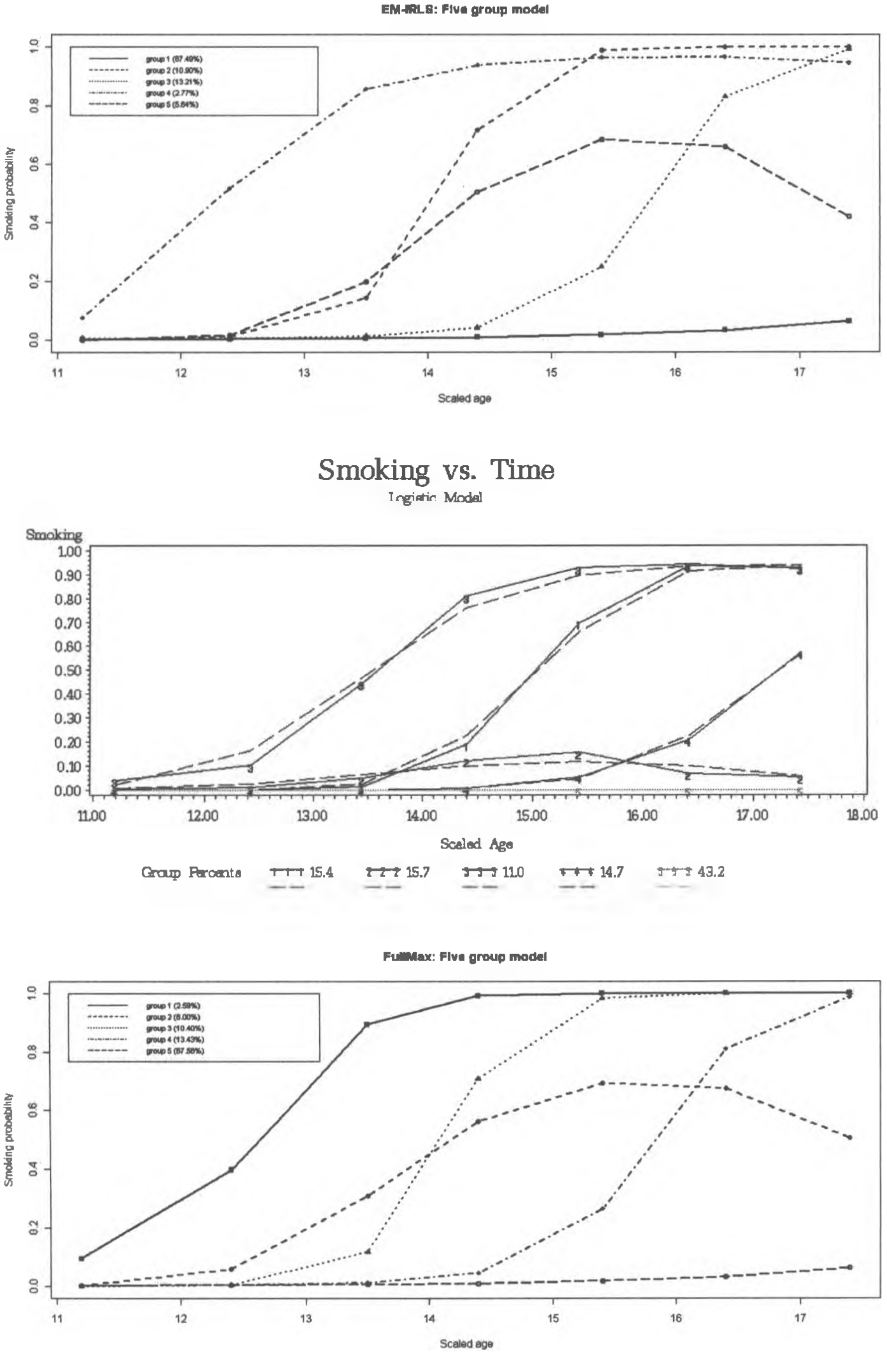


Figure 4.3: Five group model fitted by EM-IRLS, PROC TRAJ, and FullMax

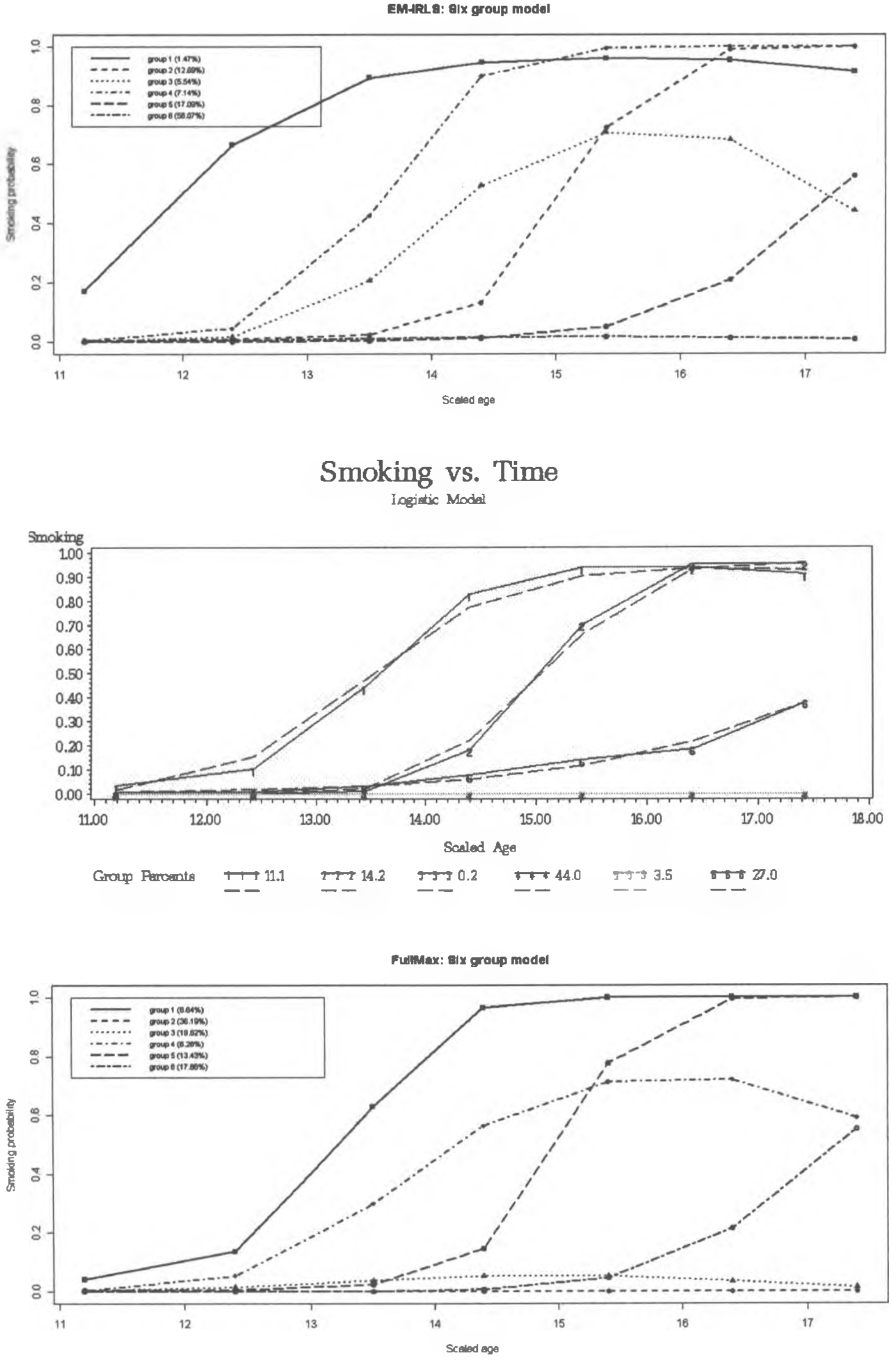


Figure 4.4: Six group model fitted by EM-IRLS, PROC TRAJ, and FullMax

4.4.2 Discussion

For model selection, BIC was used to choose the optimal model for each algorithm. The BIC values, shown in Table 4.2, indicate that the four group model was the best fitting model chosen for EM-IRLS while the five group model was optimal for the other two algorithms. We cannot draw conclusions based on the models fitted by PROC TRAJ and FullMax, because the approximate standard errors could not be calculated for the estimates in almost all of the models. The EM-IRLS algorithm converged without such problems in all models.

Table 4.2: BIC values for models

	EM-IRLS	PROC TRAJ	FullMax
3 groups	-4937.96	-4535.91	-4535.84
4 groups	-4932.54	-4530.39	-4506.53
5 groups	-5053.95	-4520.36	-4481.21
6 groups	-5556.99	-4561.88	-4491.79

The four group model obtained by EM-IRLS (see Figure 4.2) showed that four developmental trajectories related to smoking behaviour were identified, of which three were distinct smoking trajectory groups that led to regular smoking at age sixteen to seventeen. The remaining trajectory consisted of the non-smoking group of students, which was the largest subgroup within the study sample. Although the group of "quitters" identified in the five- and six-component models is of public health interest, there was only a small percentage of students within this group, indicating that there were not enough students following a reduced smoking pattern that was distinct from the other frequent smoking patterns.

Using PROC TRAJ with the procedure's default starting values, Driezen (2001, Section 3.3) fitted the same component models without covariates to this data set in order to determine the optimal number of trajectory groups. However, he obtained different trajectory models from those estimated by PROC TRAJ in our study. Driezen's (2001) results suggested that the five group model was optimal and he further investigated the effect of baseline risk factors by fitting the five group model with the risk factors as covariates. The discrepancies between our results and his results may be explained by the different versions of SAS and PROC TRAJ used. Driezen's analyses were performed using the version of PROC TRAJ tailored for SAS version 8, while we used the version of PROC TRAJ designed for SAS version 9. Some extensions of the methodology included the ability to calculate group membership probabilities as a function of time-stable covariates and fitting the dual-trajectory model (Jones and Nagin, 2007).

Nawa (2004) proposed the algorithm of starting the parameter estimation with the EM algorithm and then switch to PROC TRAJ after a few iterations. He applied this algorithm to fit the same component models without covariates to the WSPP3 data set and chose the six group model as the best model. Due to the complete case analysis approach he employed, the data set he analyzed only contained the 2394 students with complete response measurements at all seven time points and complete baseline risk factors values. This is a smaller sample from the one we analyzed; this may explain the difference between our results and those he obtained.

Nawa (2004) continued his analysis by fitting the three- to six-component models with three risk factors as covariates, and the four group model was chosen as the optimal model. The final model with covariates selected by Nawa (2004, Section 5.3.2) for this complete data set showed the same four trajectory as those in our final model produced by EM-IRLS. However, the mixing proportions were not presented so it is not certain that we have identified the same four trajectory groups.

Chapter 5

Conclusions and future work

5.1 Summary and conclusions

This thesis focused on trajectory modelling of longitudinal binary data. We considered the group-based "semi-parametric" trajectory modelling method proposed by Nagin (1999; 2005), which identifies multiple trajectories within a population using a mixture modelling approach. In the case of longitudinal binary data, the model consists of mixtures of logistic regressions, in which the regression coefficients for each group determine the shape of the group trajectory. The number of groups is unknown and has to be inferred from the available data, along with the mixing proportions and the logistic regression parameter estimates. A procedure in SAS called PROC TRAJ had been created to estimate the parameters in this trajectory model (Jones et al., 2001). This procedure employs the Quasi-Newton method for parameter estimation, but it has been shown to be very sensitive to starting values and have some convergence problems, so that the procedure sometimes fails to converge or converges to a false maximum. It has been suggested that using the EM algorithm may solve the problems (Nawa, 2004).

The EM algorithm can be implemented to perform maximization using different optimization methods or by fitting a weighted logistic regression model. To speed up the EM convergence, we proposed the use of the iteratively reweighted least squares method (denoted as EM-IRLS) to fit the weighted logistic model at the maximization step. The simulation study shows that EM-based methods produced estimates which described the correct trajectory shapes with fewer convergence problems compared to full maximization methods such as PROC TRAJ. We found that the full maximization algorithms had a higher chance of resulting in non-identifiable models where parameter estimates are unreliable. The EM-IRLS method outperformed the EM method implemented with the Quasi-Newton maximization step in terms of convergence properties and speed. When we applied the various trajectory modelling algorithms to smoking data, the results were consistent with our simulation study.

5.2 Future work

The longitudinal trajectory model we considered can be extended in several ways. We note that we have only considered models without covariates, and that models may be more stable if covariates are included (such as risk factors related to smoking behaviour). Also, this model assumes independence over time points and between individuals, which may not be true in clustered data. Since the model is fitted to longitudinal data, the independence assumption may be violated due to correlation between observations over time. The regression coefficients of our model are most likely estimated without bias, however, the estimates of standard errors may be overestimated by ignoring the dependency (see, for example, Rodriguez and Goldman (1995) and Donner and Klar (2000 pg 96)). By including covariate information and correlation structures into the model, the subgroups within the population would become more distinct with less variation. The trajectories identified from such an improved model would then be more reliable.

The current model assumes that there is a quadratic relationship between age (time) and behaviour, but the relationship can be represented using other polynomial functions as well. Our model can analyze data with missing covariates but requires complete response information from the individuals. It is of interest to extend this model to one where missingness in the responses can be handled using sophisticated methods such as multiple imputation. This can be done, for example, using SAS PROC MI (SAS, 2009) and the efficiency can be evaluated. Information loss by considering only the complete data or available data can then be reduced and the analysis would result in improved parameter estimation.

Another concern for this group-based trajectory approach is related to model selection. For mixture models, model selection is complicated and there is not one commonly accepted statistical tool for choosing the optimal number of mixture components in the models. BIC is one of the popular model selection tools often used for mixture models and model-based clustering, but researchers can also consider many other instruments. Statistical efforts need to focus on how to choose the most efficient and appropriate model selection criteria depending on the scientific problem of interest.

The implementation of any iterative procedure requires a choice of convergence criterion. Our methods were implemented such that the iterations would stop when the difference between two successive log-likelihood values is less than a specified value. It was argued that this condition is actually a lack of progress criterion rather than a convergence criterion, and that it might underestimate the correct log-likelihood value (McNicholas et al., 2010). An adjustment that can be made to our EM algorithm is to make use of an Aitken's accelerationbased convergence criterion. This condition considers the estimated value of log-likelihood that the algorithm will converge to asymptotically, based on the last three iterations, and iteration would stop when the difference between this estimated value and the current log-likelihood value is small (McNicholas et al., 2010).

Future work should focus on these issues to improve the current group-based trajectory modelling methodology. This study has been concerned with developing and evaluating methods of trajectory modelling of longitudinal binary data. Additional research can include evaluating the methodology for modelling other types of data, such as count data. Furthermore, the methods should be evaluated for the case of more than three mixture components as researchers may be interested in identifying more distinct patterns within a developmental process.

Appendix A

Results for two-component mixtures

A.1 Relative errors of mixing proportions estimates

Algorithms compared: EM-IRLS, EM-Mixed, EM-QN (represents EM-NLPQN), PT1 (represents PROC TRAJ 1), PT2 (represents PROC TRAJ 2), and FullMax

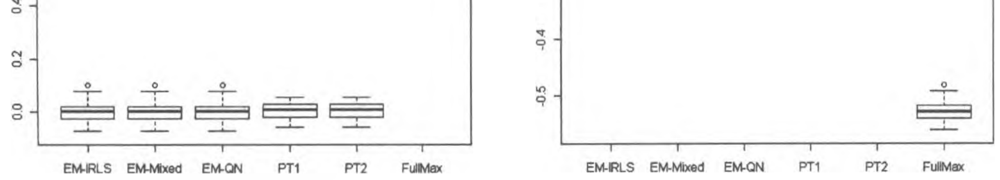


Figure A.1: Mixtures of two components: Relative errors of mixing proportions estimates in Case 1
(Trajectory 1: Temporarily quitting then resumed smoking; Trajectory 2: Stopped smoking)

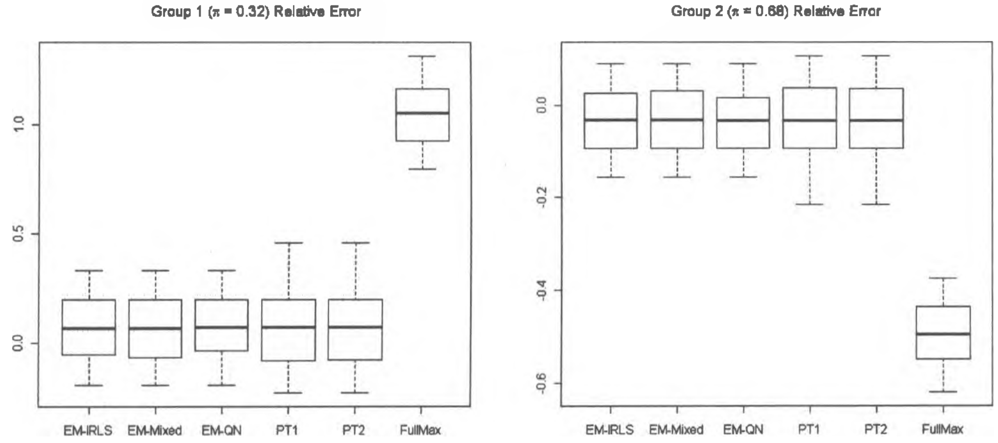


Figure A.2: Mixtures of two components: Relative errors of mixing proportions estimates in Case 2
(Trajectory 1: Never smoked; Trajectory 2: Gradual onset)

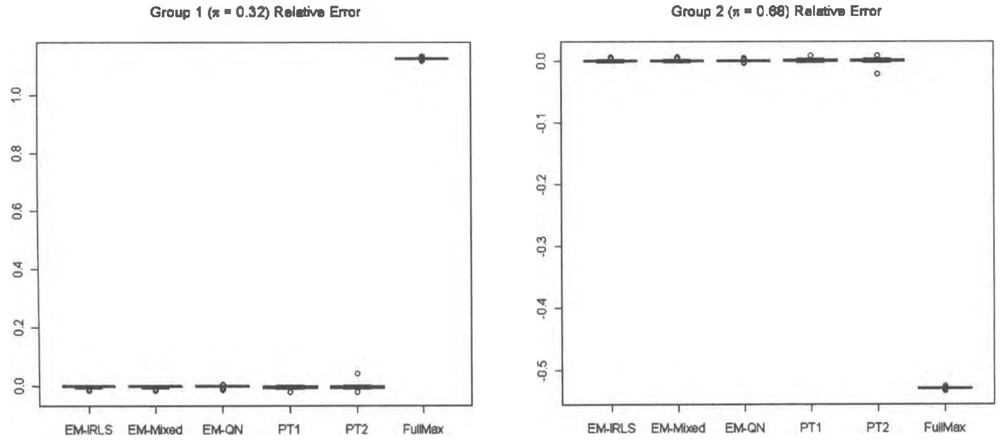


Figure A.3: Mixtures of two components: Relative errors of mixing proportions estimates in Case 3
(Trajectory 1: Never smoked; Trajectory 2: Early onset)

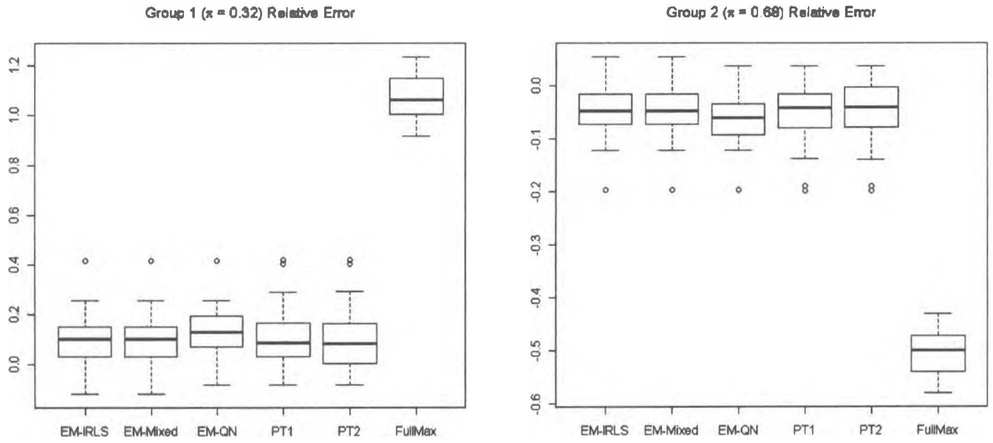


Figure A.4: Mixtures of two components: Relative errors of mixing proportions estimates in Case 4
(Trajectory 1: Early onset; Trajectory 2: Gradual onset)

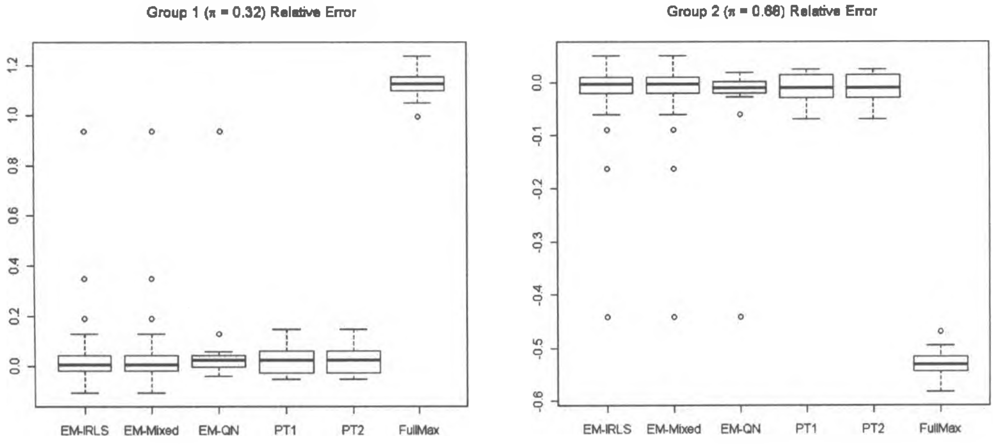


Figure A.5: Mixtures of two components: Relative errors of mixing proportions estimates in Case 5
(Trajectory 1: Early onset; Trajectory 2: Stopped smoking)

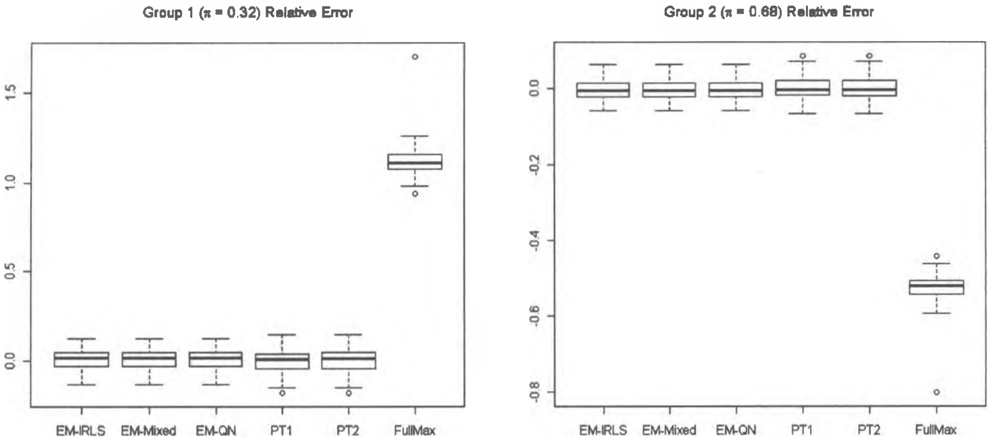


Figure A.6: Mixtures of two components: Relative errors of mixing proportions estimates in Case 6
(Trajectory 1: Gradual onset; Trajectory 2: Stopped smoking)

A.2 Parameter estimates for trajectories

Table A.1: Mixtures of two components: Parameter and standard error (SE) estimates for trajectories in Case 1

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**6.17**		**-5.78**		**0.997**	
	EM-IRLS	6.2447	0.5760	-5.8406	0.4843	1.0074	0.0835
	EM-Mixed	6.2447	0.5760	-5.8406	0.4843	1.0074	0.0835
	EM-NLPQN	6.2447	0.5760	-5.8406	0.4843	1.0074	0.0835
	PROC TRAJ 1	6.1460	0.5680	-5.7672	0.4783	0.9959	0.0825
	PROC TRAJ 2	6.1459	0.5680	-5.7672	0.4783	0.9959	0.0825
	FullMax	6.1669	0.5669	-5.7771	0.4764	0.9952	0.0820

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-7.69**		**6.59**		**-1.099**	
	EM-IRLS	-7.7649	0.4423	6.6745	0.3417	-1.1156	0.0567
	EM-Mixed	-7.7649	0.4423	6.6745	0.3417	-1.1156	0.0567
	EM-NLPQN	-7.7649	0.4423	6.6745	0.3417	-1.1156	0.0567
	PROC TRAJ 1	-7.8369	0.4503	6.6897	0.3450	-1.1129	0.0569
	PROC TRAJ 2	-7.8369	0.4503	6.6897	0.3450	-1.1129	0.0569
	FullMax	-7.8164	0.4486	6.6984	0.3455	-1.1178	0.0572

*Starting values of β: (-2, 0, 0), (-1, 0, 0)

*Trajectory 1: Temporarily quitting then resumed smoking

*Trajectory 2: Stopped smoking

*Parameter and SE estimates are averaged over the number of converged samples.

Table A.2: Mixtures of two components: Parameter and standard error (SE) estimates for trajectories in Case 2

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**-3.00**		**0.01**		**0.01**	
	EM-IRLS	-3.4036	1.4367	0.3110	1.3131	-0.0607	0.2773
	EM-Mixed	-3.4097	1.4799	0.3382	1.3872	-0.0730	0.3032
	EM-NLPQN	-3.2973	1.5474	0.2183	1.5086	-0.0427	0.3502
	PROC TRAJ 1	-2.8259	1.2073	-0.1202	1.1545	0.0133	0.2507
	PROC TRAJ 2	-2.8242	1.2062	-0.1229	1.1528	0.0141	0.2502
	FullMax	-3.3160	1.6136	0.2214	1.5854	-0.0384	0.3720

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-2.24**		**-0.17**		**0.21**	
	EM-IRLS	-2.2045	0.4174	-0.1620	0.3185	0.2103	0.0574
	EM-Mixed	-2.2186	0.4167	-0.1531	0.3174	0.2088	0.0571
	EM-NLPQN	-2.2190	0.4200	-0.1561	0.3204	0.2100	0.0578
	PROC TRAJ 1	-2.3028	0.4290	-0.1350	0.3263	0.2102	0.0589
	PROC TRAJ 2	-2.3018	0.4292	-0.1360	0.3265	0.2104	0.0590
	FullMax	-2.2066	0.4199	-0.1636	0.3213	0.2112	0.0580

*Starting values of β: (-2, 0, 0), (-1, 0, 0)

*Trajectory 1: Never smoked

*Trajectory 2: Gradual onset

*Parameter and SE estimates are averaged over the number of converged samples.

Table A.3: Mixtures of two components: Parameter and standard error (SE) estimates for trajectories in Case 3

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**-3.00**		**0.01**		**0.01**	
	EM-IRLS	-2.9762	0.7940	-0.0504	0.5956	0.0204	0.0961
	EM-Mixed	-2.9762	0.7940	-0.0504	0.5956	0.0204	0.0961
	EM-NLPQN	-3.0230	0.7886	0.0135	0.5881	0.0100	0.0950
	PROC TRAJ 1	-2.9984	0.7924	-0.0151	0.5937	0.0135	0.0962
	PROC TRAJ 2	-3.0707	0.8034	0.0918	0.6116	-0.0143	0.1016
	FullMax	-2.9788	0.7909	-0.0500	0.5918	0.0217	0.0954

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-3.05**		**-0.80**		**1.35**	
	EM-IRLS	-4.1026	1.2403	0.7916	1.6732	0.8231	0.5301
	EM-Mixed	-4.1026	1.2403	0.7916	1.6732	0.8231	0.5301
	EM-NLPQN	-4.1239	1.5064	0.8190	2.0821	0.8049	0.6671
	PROC TRAJ 1	-3.9757	2.1488	0.5805	3.0536	0.8910	0.9933
	PROC TRAJ 2	-3.9329	2.0988	0.5573	2.9802	0.8946	0.9690
	FullMax	4.5067	3.3414	-12.0924	4.9466	5.1107	1.6418

*Starting values of β: (-2, 0, 0), (-1, 0, 0)

*Trajectory 1: Never smoked

*Trajectory 2: Early onset

*Parameter and SE estimates are averaged over the number of converged samples.

Table A.4: Mixtures of two components: Parameter and standard error (SE) estimates for trajectories in Case 4

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**-3.05**		**-0.80**		**1.35**	
	EM-IRLS	-6.0465	1.8059	3.6679	2.2676	-0.1545	0.6900
	EM-Mixed	-6.0465	1.8059	3.6679	2.2676	-0.1545	0.6900
	EM-NLPQN	-5.8994	2.0178	3.5958	2.5703	-0.1674	0.7841
	PROC TRAJ 1	-6.0140	2.0786	3.7050	2.7182	-0.1795	0.8436
	PROC TRAJ 2	-5.7970	5.1369	3.3578	7.3353	-0.0540	2.3881
	FullMax	-3.6014	3.1224	-0.0519	4.4737	1.1068	1.4750

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-2.24**		**-0.17**		**0.21**	
	EM-IRLS	-2.1929	0.4192	-0.2407	0.3202	0.2209	0.0518
	EM-Mixed	-2.1929	0.4192	-0.2407	0.3202	0.2209	0.0518
	EM-NLPQN	-2.2043	0.4293	-0.2529	0.3290	0.2236	0.0532
	PROC TRAJ 1	-2.2697	0.4321	-0.2678	0.3273	0.2312	0.0528
	PROC TRAJ 2	-2.2148	0.4268	-0.2850	0.3236	0.2324	0.0523
	FullMax	-2.2435	0.4027	-0.1617	0.2994	0.2051	0.0480

*Starting values of β: (-1, 0, 0), (1, 0, 0)

*Trajectory 1: Early onset

*Trajectory 2: Gradual onset

*Parameter and SE estimates are averaged over the number of converged samples.

Table A.5: Mixtures of two components: Parameter and standard error (SE) estimates for trajectories in Case 5

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**-3.05**		**-0.80**		**1.35**	
	EM-IRLS	-5.9616	2.2902	2.7495	2.5394	0.3038	0.7022
	EM-Mixed	-5.9616	2.2902	2.7495	2.5394	0.3038	0.7022
	EM-NLPQN	-4.2215	5.9679	1.0534	8.7071	0.7053	2.8638
	PROC TRAJ 1	-4.6839	5.5193	1.5702	8.0482	0.5836	2.6486
	PROC TRAJ 2	-4.6953	4.9202	1.5874	7.1480	0.5779	2.3483
	FullMax	1.5555	5.3244	-7.6308	7.8310	3.6142	2.5858

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-7.69**		**6.59**		**-1.099**	
	EM-IRLS	-7.7008	0.4349	6.6452	0.3428	-1.1135	0.0603
	EM-Mixed	-7.7008	0.4349	6.6452	0.3428	-1.1135	0.0603
	EM-NLPQN	-7.7703	0.4495	6.7397	0.3620	-1.1355	0.0648
	PROC TRAJ 1	-7.9076	0.4458	6.7912	0.3585	-1.1363	0.0645
	PROC TRAJ 2	-7.9077	0.4453	6.7912	0.3579	-1.1363	0.0644
	FullMax	-7.7703	0.4269	6.6886	0.3404	-1.1189	0.0608

*Starting values of β: (-1, 0, 0), (1, 0, 0)

*Trajectory 1: Early onset

*Trajectory 2: Stopped smoking

*Parameter and SE estimates are averaged over the number of converged samples.

Table A.6: Mixtures of two components: Parameter and standard error (SE) estimates for trajectories in Case 6

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**-3.05**		**-0.80**		**1.35**	
	EM-IRLS	-2.2102	0.6573	-0.1781	0.5553	0.2102	0.1010
	EM-Mixed	-2.2102	0.6573	-0.1781	0.5553	0.2102	0.1010
	EM-NLPQN	-2.2101	0.6573	-0.1781	0.5553	0.2102	0.1010
	PROC TRAJ 1	-2.1679	0.6706	-0.2679	0.5701	0.2294	0.1038
	PROC TRAJ 2	-2.1773	0.6657	-0.2439	0.5656	0.2240	0.1029
	FullMax	-4.7424	1.8854	1.4211	0.8606	-0.0293	0.1598

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-7.69**		**6.59**		**-1.099**	
	EM-IRLS	-7.7809	0.4608	6.6983	0.3933	-1.1202	0.0697
	EM-Mixed	-7.7809	0.4608	6.6983	0.3933	-1.1202	0.0697
	EM-NLPQN	-7.7808	0.4608	6.6983	0.3933	-1.1202	0.0697
	PROC TRAJ 1	-7.8018	0.4587	6.6618	0.3865	-1.1083	0.0679
	PROC TRAJ 2	-7.8234	0.4609	6.6772	0.3882	-1.1100	0.0682
	FullMax	-7.5932	0.4580	6.4817	0.3884	-1.0797	0.0686

*Starting values of β: (-1, 0, 0), (1, 0, 0)

*Trajectory 1: Gradual onset

*Trajectory 2: Stopped smoking

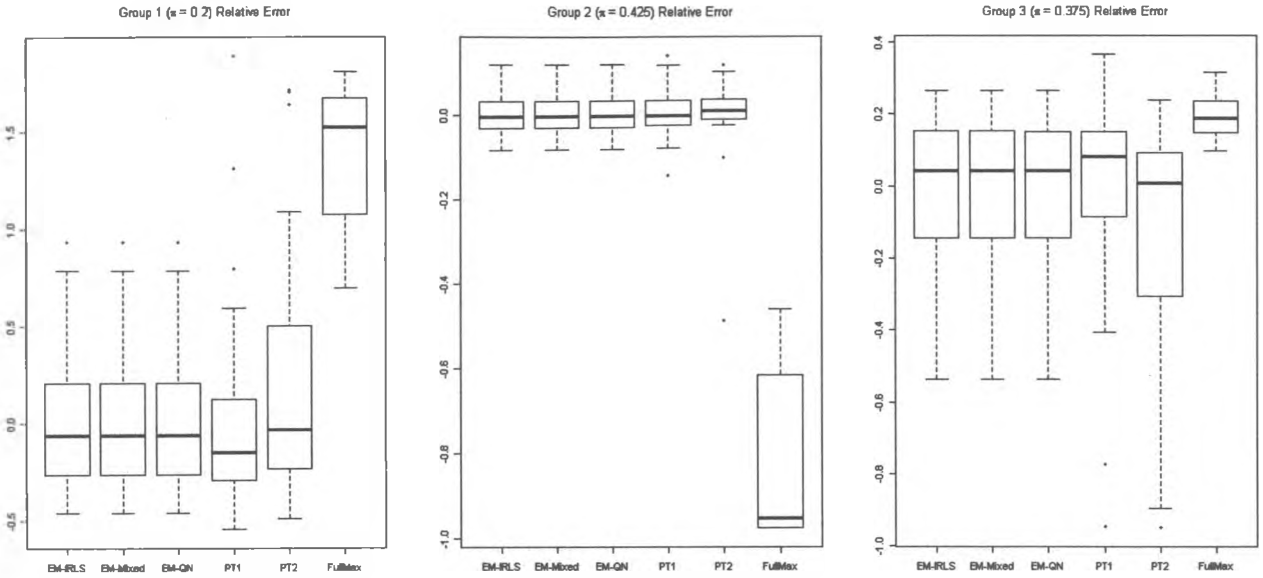
*Parameter and SE estimates are averaged over the number of converged samples.

Appendix B

Results for three-component mixtures

B.1 Relative errors of mixing proportions estimates

Algorithms compared: EM-IRLS, EM-Mixed, EM-QN (represents EM-NLPQN), PT1 (represents PROC TRAJ 1), PT2 (represents PROC TRAJ 2), and FullMax



Group 1 (π = 0.2) Relative Error

Group 2 (π = 0.425) Relative Error

Group 3 (π = 0.375) Relative Error

Figure B.1: Mixtures of three components: Relative errors of mixing proportions estimates in Case 1 (Trajectory 1: Temporarily quitting then resumed smoking; Trajectory 2: Stopped smoking; Trajectory 3: Gradual onset)

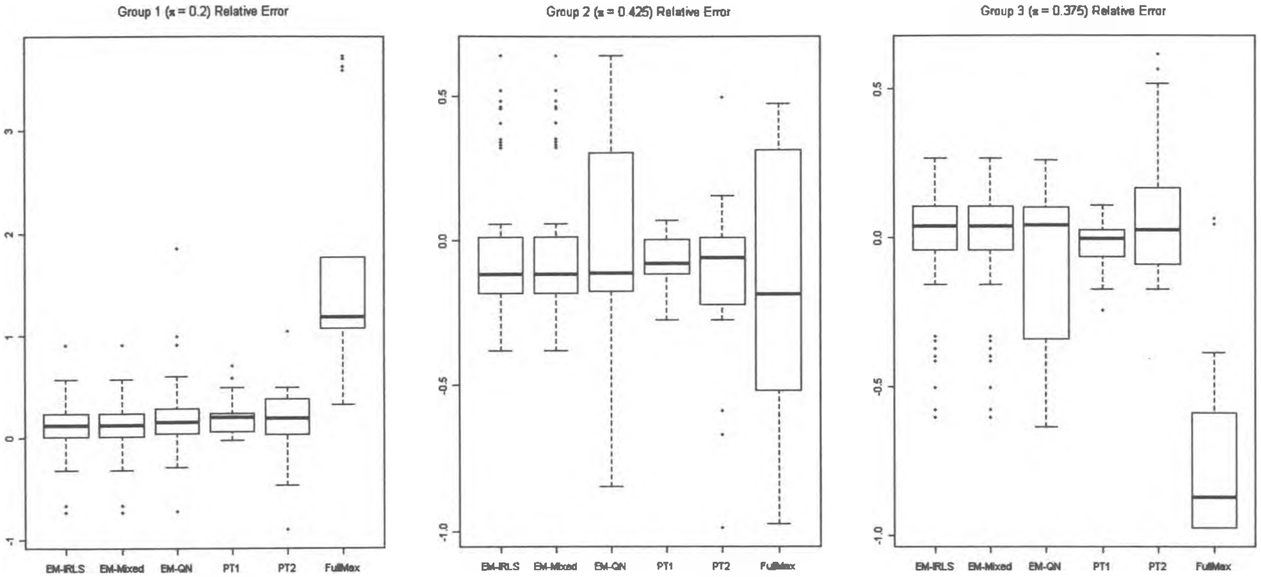


Figure B.2: Mixtures of three components: Relative errors of mixing proportions estimates in Case 2 (Trajectory 1: Early onset; Trajectory 2: Gradual onset; Trajectory 3: Stopped smoking)

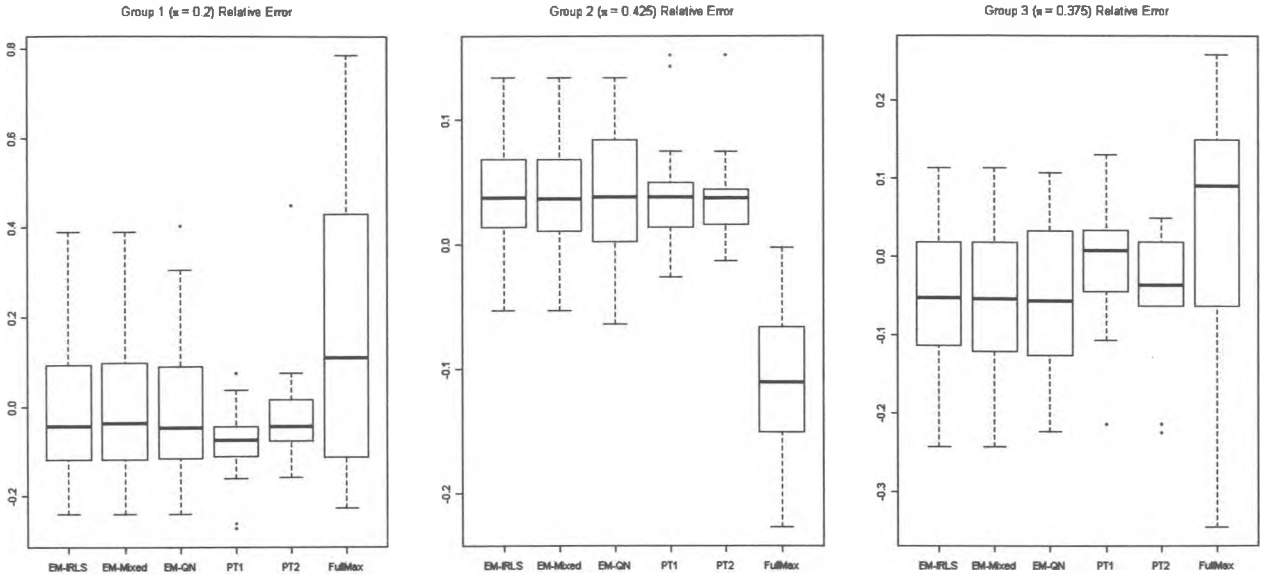


Figure B.3: Mixtures of three components: Relative errors of mixing proportions estimates in Case 3 (Trajectory 1: Never smoked; Trajectory 2: Early onset; Trajectory 3: Gradual onset)

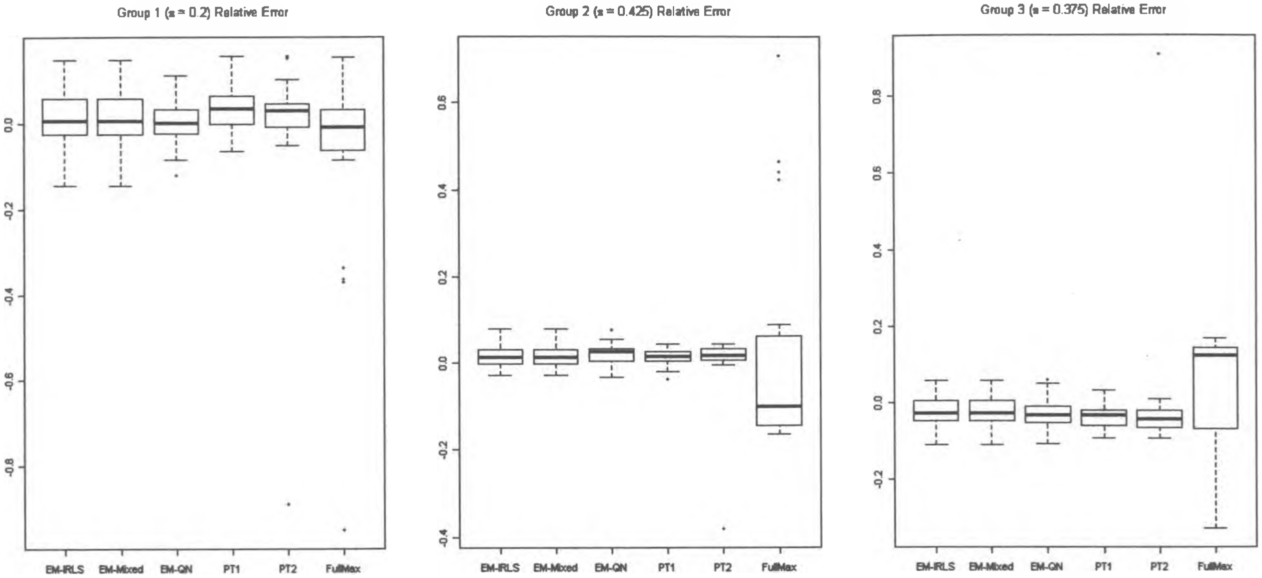


Figure B.4: Mixtures of three components: Relative errors of mixing proportions estimates in Case 4 (Trajectory 1: Never smoked; Trajectory 2: Early onset; Trajectory 3: Stopped smoking)

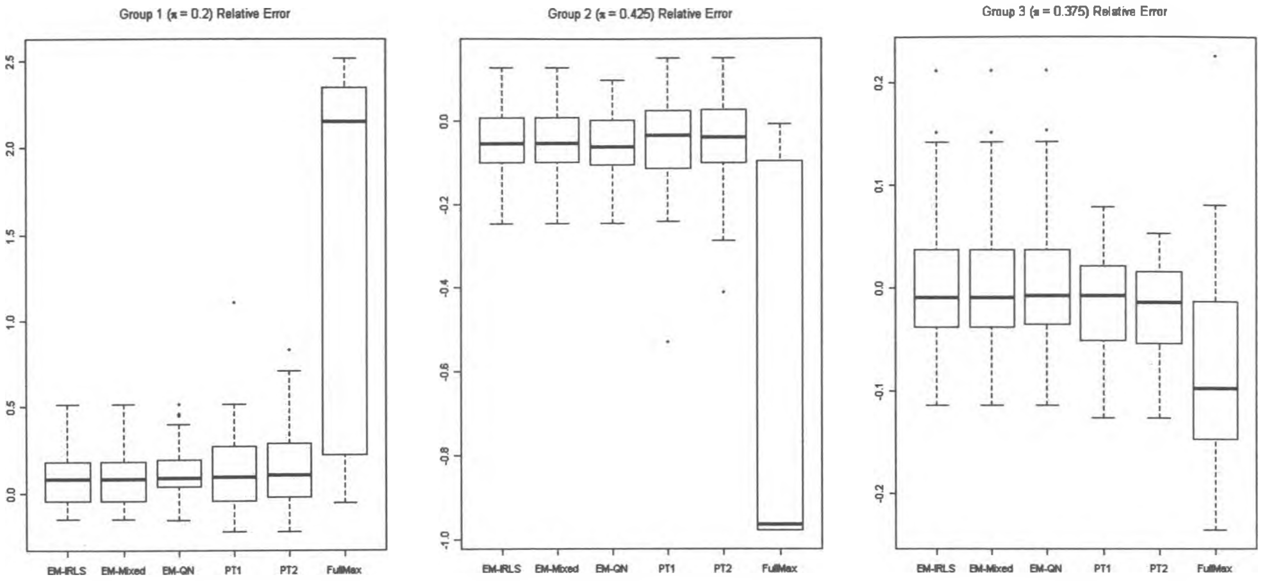


Figure B.5: Mixtures of three components: Relative errors of mixing proportions estimates in Case 5 (Trajectory 1: Never smoked; Trajectory 2: Gradual onset; Trajectory 3: Stopped smoking)

B.2 Parameter estimates for trajectories

Table B.1: Mixtures of three components: Parameter and standard error (SE) estimates for trajectories in Case 1

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**6.17**		**-5.78**		**0.997**	
	EM-IRLS	6.6065	3.0975	-6.1384	2.1327	1.0501	0.3098
	EM-Mixed	6.6065	3.0975	-6.1384	2.1327	1.0501	0.3098
	EM-NLPQN	6.6051	3.1380	-6.1374	2.1590	1.0499	0.3134
	PROC TRAJ 1	7.4092	3.5154	-6.6626	2.4711	1.1272	0.3656
	PROC TRAJ 2	6.1253	2.3614	-5.6926	1.6220	0.9784	0.2349
	FullMax	3.2042	5.6249	-1.9946	4.4331	-0.3504	3.2960

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-7.69**		**6.59**		**-1.099**	
	EM-IRLS	-7.7751	0.5181	6.6901	0.4471	-1.1178	0.0799
	EM-Mixed	-7.7751	0.5181	6.6901	0.4471	-1.1178	0.0799
	EM-NLPQN	-7.7751	0.5181	6.6901	0.4471	-1.1178	0.0799
	PROC TRAJ 1	-7.9494	0.5300	6.7787	0.4463	-1.1274	0.0784
	PROC TRAJ 2	-8.1221	0.6754	6.7813	0.5466	-1.1142	0.0924
	FullMax	-7.6487	0.5298	6.4363	0.4364	-1.0601	0.0753

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
3	**Theoretical**	**-2.24**		**-0.17**		**0.21**	
	EM-IRLS	-2.4015	1.7643	-0.0428	1.1499	0.1920	0.1700
	EM-Mixed	-2.4015	1.7643	-0.0428	1.1499	0.1920	0.1700
	EM-NLPQN	-2.4018	1.7885	-0.0426	1.1654	0.1919	0.1722
	PROC TRAJ 1	-2.1914	1.9116	-0.2278	1.1092	0.2294	0.1923
	PROC TRAJ 2	-2.6334	1.8671	0.4620	1.4352	0.0745	0.2289
	FullMax	0.6187	0.8922	-2.1574	0.6664	0.5100	0.1084

*Starting values of β: (-2, 0, 0), (-1, 0, 0), (2, 0, 0)

*Trajectory 1: Temporarily quitting then resumed smoking

*Trajectory 2: Stopped smoking

*Trajectory 3: Gradual onset

*Parameter and SE estimates are averaged over the number of converged samples.

Table B.2: Mixtures of three components: Parameter and standard error (SE) estimates for trajectories in Case 2

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**-3.05**		**-0.80**		**1.35**	
	EM-IRLS	-6.3523	1.9586	3.9648	1.8948	-0.3570	0.4664
	EM-Mixed	-6.3523	1.9586	3.9648	1.8948	-0.3570	0.4664
	EM-NLPQN	-5.2830	3.0544	2.9047	3.7849	-0.1590	1.1287
	PROC TRAJ 1	-6.1273	2.3703	3.8106	3.1282	-0.2110	0.9869
	PROC TRAJ 2	-5.6193	1.7600	3.2682	2.0025	-0.2455	0.5706
	FullMax	-5.2471	12.1678	0.5842	9.6299	5.3895	15.8710

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-2.24**		**-0.17**		**0.21**	
	EM-IRLS	-1.9168	0.7820	-0.6168	0.6914	0.2991	0.1260
	EM-Mixed	-1.9168	0.7820	-0.6168	0.6914	0.2991	0.1260
	EM-NLPQN	-1.7585	1.0029	-0.8040	0.9169	0.3446	0.1762
	PROC TRAJ 1	-2.2499	0.5298	-0.3055	0.4605	0.2385	0.0806
	PROC TRAJ 2	-0.3019	8.4694	-1.3147	7.5900	0.3957	1.4306
	FullMax	-3.7483	0.5087	1.4278	0.4627	-0.0569	0.0987

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
3	**Theoretical**	**-7.69**		**6.59**		**-1.099**	
	EM-IRLS	-7.6393	0.7187	6.4801	0.6973	-1.0588	0.1446
	EM-Mixed	-7.6393	0.7187	6.4801	0.6973	-1.0588	0.1446
	EM-NLPQN	-7.9022	0.8071	6.6679	0.7861	-1.0728	0.1613
	PROC TRAJ 1	-7.7521	0.5974	6.7091	0.5514	-1.1279	0.1057
	PROC TRAJ 2	-7.6860	0.6769	6.5365	0.6344	-1.0528	0.1106
	FullMax	-6.7069	2.0202	5.5208	1.2035	-0.5190	0.8468

*Starting values of β: (-2, 0, 0), (-1, 0, 0), (1, 0, 0)

*Trajectory 1: Early onset

*Trajectory 2: Gradual onset

*Trajectory 3: Stopped smoking

*Parameter and SE estimates are averaged over the number of converged samples.

Table B.3: Mixtures of three components: Parameter and standard error (SE) estimates for trajectories in Case 3

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**-3.00**		**0.01**		**0.01**	
	EM-IRLS	-3.8923	1.7860	0.8085	1.5681	-0.1688	0.3366
	EM-Mixed	-3.8668	1.7685	0.7771	1.5502	-0.1613	0.3322
	EM-NLPQN	-3.6334	1.4652	0.7706	1.4454	-0.1881	0.3408
	PROC TRAJ 1	-3.6342	1.9603	0.9570	2.3083	-0.2738	0.6408
	PROC TRAJ 2	-3.1276	1.3505	0.5101	1.4055	-0.1193	0.3281
	FullMax	-3.4621	1.4931	0.1183	2.3242	0.1457	0.7220

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-3.05**		**-0.80**		**1.35**	
	EM-IRLS	-5.7396	1.1551	3.2811	1.5176	-0.0153	0.4822
	EM-Mixed	-5.7172	1.2674	3.2488	1.6963	-0.0041	0.5445
	EM-NLPQN	-5.0633	9.1377	2.2452	13.5774	0.3359	4.5198
	PROC TRAJ 1	-5.2361	2.9553	2.6943	4.1972	0.0826	1.3697
	PROC TRAJ 2	-5.7331	1.5767	3.2956	2.1044	-0.0355	0.6592
	FullMax	-6.3667	8.8728	-1.2030	5.3900	2.6280	1.9163

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
3	**Theoretical**	**-2.24**		**-0.17**		**0.21**	
	EM-IRLS	-2.2268	0.5299	-0.2189	0.4280	0.2184	0.0736
	EM-Mixed	-2.2261	0.5309	-0.2182	0.4312	0.2182	0.0738
	EM-NLPQN	-2.2662	0.5268	-0.2003	0.4295	0.2168	0.0716
	PROC TRAJ 1	-2.1377	4.6742	-0.6107	6.6828	0.4447	2.1728
	PROC TRAJ 2	-2.2518	0.5693	-0.2838	0.4467	0.2134	0.0780
	FullMax	-2.3859	0.6163	0.0928	0.4874	0.1780	0.0795

*Starting values of β: (-2, 0, 0), (-1, 0, 0), (1, 0, 0)

*Trajectory 1: Never smoked

*Trajectory 2: Early onset

*Trajectory 3: Gradual onset

*Parameter and SE estimates are averaged over the number of converged samples.

Table B.4: Mixtures of three components: Parameter and standard error (SE) estimates for trajectories in Case 4

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**-3.00**		**0.01**		**0.01**	
	EM-IRLS	-2.9701	1.3124	-0.1158	1.2560	0.0321	0.2071
	EM-Mixed	-2.9701	1.3124	-0.1158	1.2560	0.0321	0.2071
	EM-NLPQN	-2.9011	1.1233	-0.1491	1.0335	0.0382	0.1698
	PROC TRAJ 1	-3.3741	0.9386	0.4440	0.8056	-0.0627	0.1323
	PROC TRAJ 2	-2.9315	1.3719	0.3886	0.9705	-0.0665	0.1457
	FullMax	-2.4986	1.4869	-0.1383	1.6163	-0.0740	0.8057

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-3.05**		**-0.80**		**1.35**	
	EM-IRLS	-4.7172	1.1627	1.7032	1.5352	0.5284	0.4802
	EM-Mixed	-4.7172	1.1627	1.7032	1.5352	0.5284	0.4802
	EM-NLPQN	-4.1755	4.3674	0.8829	6.4152	0.8045	2.1189
	PROC TRAJ 1	-4.4190	1.7646	1.1841	2.4866	0.6955	0.8054
	PROC TRAJ 2	-4.5608	3.5362	1.3828	5.1143	0.6099	1.6749
	FullMax	-7.1540	3.7793	2.3678	2.7014	0.7335	1.0097

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
3	**Theoretical**	**-7.69**		**6.59**		**-1.099**	
	EM-IRLS	-7.8646	0.4971	6.7891	0.4193	-1.1395	0.0747
	EM-Mixed	-7.8646	0.4971	6.7891	0.4193	-1.1395	0.0747
	EM-NLPQN	-7.9336	0.5239	6.8384	0.4473	-1.1484	0.0805
	PROC TRAJ 1	-7.9666	0.5185	6.8929	0.4446	-1.1564	0.0800
	PROC TRAJ 2	-8.0649	0.5319	6.9542	0.4510	-1.1631	0.0800
	FullMax	-7.6069	0.4536	6.5106	0.3706	-1.0565	0.0626

*Starting values of β: (-2, 0, 0), (-1, 0, 0), (1, 0, 0)

*Trajectory 1: Never smoked

*Trajectory 2: Early onset

*Trajectory 3: Stopped smoking

*Parameter and SE estimates are averaged over the number of converged samples.

Table B.5: Mixtures of three components: Parameter and standard error (SE) estimates for trajectories in Case 5

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
1	**Theoretical**	**-3.00**		**0.01**		**0.01**	
	EM-IRLS	-3.2160	1.9600	0.0060	1.8592	0.0164	0.3435
	EM-Mixed	-3.2160	1.9600	0.0060	1.8592	0.0164	0.3435
	EM-NLPQN	-2.7353	1.4051	-0.2868	1.4100	0.0670	0.2774
	PROC TRAJ 1	-3.2954	1.4959	0.4949	1.5925	-0.1086	0.3817
	PROC TRAJ 2	-3.2313	1.3913	0.3531	1.4142	-0.0575	0.3164
	FullMax	-3.2094	0.7037	0.6079	0.6272	-0.0079	0.1203

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
2	**Theoretical**	**-2.24**		**-0.17**		**0.21**	
	EM-IRLS	-2.2274	0.5329	-0.2045	0.4665	0.2254	0.0896
	EM-Mixed	-2.2274	0.5329	-0.2045	0.4665	0.2254	0.0896
	EM-NLPQN	-2.2104	0.5372	-0.2195	0.4708	0.2287	0.0907
	PROC TRAJ 1	-2.3797	0.5501	-0.1216	0.4760	0.2089	0.0914
	PROC TRAJ 2	-2.4114	0.5503	-0.0447	0.4819	0.1875	0.0924
	FullMax	-1.4061	4.9705	2.2696	6.2296	-0.8355	3.3652

		β_0		β_1		β_2	
Group		Estimate	SE	Estimate	SE	Estimate	SE
3	**Theoretical**	**-7.69**		**6.59**		**-1.099**	
	EM-IRLS	-7.6928	0.5356	6.6179	0.4773	-1.1061	0.0857
	EM-Mixed	-7.6928	0.5356	6.6179	0.4773	-1.1061	0.0857
	EM-NLPQN	-7.7038	0.5372	6.6265	0.4787	-1.1076	0.0860
	PROC TRAJ 1	-7.7877	0.5359	6.7222	0.4766	-1.1251	0.0851
	PROC TRAJ 2	-7.6895	0.5442	6.5768	0.4832	-1.0952	0.0867
	FullMax	-8.0079	0.7092	6.9550	0.5833	-1.1552	0.0998

*Starting values of β: (-2, 0, 0), (-1, 0, 0), (1, 0, 0)

*Trajectory 1: Never smoked

*Trajectory 2: Gradual onset

*Trajectory 3: Stopped smoking

*Parameter and SE estimates are averaged over the number of converged samples.

Appendix C

SAS/IML Macro

```
/***************************************************************\
    SAS/IML macro for identifying group-based trajectories
    (EM-IRLS for two-group model)

    dataset: data set to be analyzed
    cov: time/age information at each time point
  (e.g. age1 age2 age3 age4 age5)
    dep: response variables at each time point
  (e.g. smk1 smk2 smk3 smk4 smk5)
    groups: number of mixture components

\***************************************************************/
%macro traj_model(dataset, cov, dep, groups);
proc iml;

  use &dataset;
  read all var{&cov} into time;
  read all var{&dep} into resp;
  n = nrow(resp);
  m = ncol(resp);
  g = &groups;

  /* check if have missing covariates */
  miss=0;
  do indiv = 1 to n while (miss=0);
    do tp = 1 to m;
      if time[indiv,tp] = . then miss = 1;
    end;
  end;

  /* define variables */
  loglik = 0;
  loglik_m = 0;
```

```
iter = 1;

/* set up vectors for parameters */
/* equal proportions as initial values for mixing proportions */
default_pi = 1/g;
pi = shape(default_pi, g, 1);
pi_m = shape(0, g, 1);
zj = shape(0,n,g);
err = 0;

/* starting values for beta parameters */
beta = shape(0,3,g);
do i = 1 to g;
  k = -3 + i;
  beta[1,i] = k;
end;
beta_m = shape(0,3,g);

do until(iter > 1000| (diffloglik <= 0.0001));

  /* E-step */
  do indiv = 1 to n;
    compdensity = shape(1,g,1);
    f1 = 1;
    f2 = 1;
    z_est = shape(0,g,1);
    do tp = 1 to m;
      if time[indiv,tp]^=. then do;
        timej = 1 || time[indiv,tp] || time[indiv,tp]#2;
        yjt = resp[indiv,tp];

        do grp = 1 to g;
          expA = exp(beta[1,grp])
                  #exp(timej[2]#beta[2,grp])#exp(timej[3]#beta[3,grp]);
          if yjt = 1 then Tt = expA/(1+expA);
          else Tt = 1/(1+expA);
          compdensity[grp] = compdensity[grp]#Tt;
          if grp = 1 then f1 = f1#Tt;
          if grp = 2 then f2 = f2#Tt;
        end;
      end;
  end;

  density = 0;
  do grp = 1 to g;
    density = density + pi[grp]#compdensity[grp];
  end;

  do grp = 1 to g;
    z_est[grp] = (pi[grp]#compdensity[grp])/density;
    zj[indiv,grp] = z_est[grp];
  end;
end;

/* M-step */
/* estimation of mixing proportions */
do grp = 1 to g;
  pi_m[grp] = sum(zj[,grp])/n;
  if pi_m[grp] = 0 then pi_m[grp] = 0.000001;
end;

/* IRLS */
/* estimation of beta parameters */
do grp = 1 to g while (err = 0);
  err = 0;
  b = 0;
  newb = beta[,grp];
  total = m#n;
  index = 1;

  X = shape(1, total, 3);
  do indiv = 1 to n;
    do tp = 1 to m;
      if time[indiv,tp] ^=. then do;
        a = time[indiv,tp];
        X[index,1] = 1;
      end;
      else do;
        a = 0;
        X[index,1] = 0;
      end;
      X[index,2] = a;
    X[index,3] = a##2;
    index = index+1;
  end;
end;
Y = shape(1, total, 1);
Z = shape(1, total, 1);
index = 1;
do indiv = 1 to n;
  do tp = 0 to m-1;
    Z[index+tp] = zj[indiv,grp];
    Y[index+tp] = resp[indiv,tp+1];
  end;
  index = index+m;
end;

fz = 0;
do looptime = 1 to 20 while(max(abs(newb-b)) > 1e-8);
  b = newb;
  fz = X*b;
  fpi = shape(0,nrow(fz),1);
  do k = 1 to total while(err = 0);
    if fz[k] > 700 then fpi[k] = 1/(1+exp(-fz[k]));
    else fpi[k] = exp(fz[k])/(1+exp(fz[k]));
  end;

  if err = 0 then do;
    fpi = choose(fpi=0, 0.0000001, fpi);
    fpi = choose(fpi=1, 0.9999999, fpi);

    W = Z/(fpi#(1-fpi));
    xx = fpi#fpi#exp(-fz)#X;
    info_mat = t(xx)*(W#xx);
  end;

  if det(info_mat) = 0 then err = 1;
  if err = 0 then do;
    info = inv(info_mat);
    D = Y-fpi;
    score = t(xx)*(W#D);
    newb = b + info*score;
  end;
```

```
    else do;
      newb = b;
    end;
  end;

  beta_m[,grp] = b;
end;

/* calculate log-likelihood */
if err = 0 then do;
  z = shape(0,n,g);
  do indiv = 1 to n;
    do grp = 1 to g;
      z[indiv,grp] = zj[indiv,grp]#log(pi_m[grp]);
    end;
  end;

  part1 = 0;
  do grp = 1 to g;
    part1 = part1 + sum(z[,grp]);
  end;

  z = shape(0,n,g);
  do indiv = 1 to n;
    do grp = 1 to g;
      T = 0;
      b = beta_m[,grp];
      do tp = 1 to m;
        if time[indiv,tp] ^= . then do;
          timej = 1 || time[indiv,tp] || time[indiv,tp]##2;
          yjt = resp[indiv,tp];
          A = b[1] + (timej[2]#b[2]) + (timej[3]#b[3]);
          T = T + (yjt#A) - A - log(1+exp(-A));
        end;
      end;
      z[indiv,grp] = zj[indiv,grp]#T;
    end;
  end;

  part2 = 0;
  do grp = 1 to g;
```

```
      part2 = part2 + sum(z[,grp]);
    end;

    loglik_m = part1 + part2;
  end;
  else do;
    loglik_m = loglik;
  end;

  /* initialization for the next iteration */
  diffloglik = abs(abs(loglik)-abs(loglik_m));
  pi = pi_m;
  beta = beta_m;
  loglik = loglik_m;
  iter = iter+1;

end;
iter = iter-1;

  /* SE calculation (For two group model)*/
if err = 0 then do;
  Ic_pi = sum(zj[,1]/(pi[1]##2) + zj[,2]/(pi[2]##2));

  do gp = 1 to 2;
    za = shape(0,n,1); zb = shape(0,n,1);
    zc = shape(0,n,1); zd = shape(0,n,1);
    ze = shape(0,n,1);
    do indiv = 1 to n;
      aa = 0; bb = 0;
      cc = 0; dd = 0;
      ee = 0;
      if gp = 1 then b = beta[,1]; if gp = 2 then b = beta[,2];
      do tp = 1 to m;
        agejt = time[indiv,tp];
        agej = 1 || agejt || agejt#agejt;
        expAB = exp(b[1])#exp(agej[2]#b[2])#exp(agej[3]#b[3]);
        denom = (1+expAB)##2;
        aa = aa + expAB/denom;
        bb = bb + (agejt#expAB)/denom;
        cc = cc + ((agejt##2)#expAB)/denom;
        dd = dd + ((agejt##3)#expAB)/denom;
      ee = ee + ((agejt##4)#expAB)/denom;
    end;
    if gp = 1 then z = zj[,1]; if gp = 2 then z = zj[,2];
    za[indiv] = z[indiv]#aa; zb[indiv] = z[indiv]#bb;
    zc[indiv] = z[indiv]#cc; zd[indiv] = z[indiv]#dd;
    ze[indiv] = z[indiv]#ee;
  end;
  if gp = 1 then
    Ic_beta1 =  (sum(za)||sum(zb)||sum(zc))//
                (sum(zb)||sum(zc)||sum(zd))//
                (sum(zc)||sum(zd)||sum(ze));
  if gp = 2 then
    Ic_beta2 =  (sum(za)||sum(zb)||sum(zc))//
                (sum(zb)||sum(zc)||sum(zd))//
                (sum(zc)||sum(zd)||sum(ze));
end;

covsc_pi =  sum((zj[,1]#(1-zj[,1]))/(pi[1]##2) +
                (zj[,2]#(1-zj[,2]))/(pi[2]##2) +
                (2#zj[,1]#zj[,2])/(pi[1]#pi[2]));

A0j_1 = shape(0,n,1); A1j_1 = shape(0,n,1); A2j_1 = shape(0,n,1);
A0j_2 = shape(0,n,1); A1j_2 = shape(0,n,1); A2j_2 = shape(0,n,1);
do indiv = 1 to n;
  do gp = 1 to 2;
    if gp = 1 then b = beta[,1]; if gp = 2 then b = beta[,2];
    a0 = 0; a1 = 0; a2 = 0;
    do tp = 1 to m;
      agejt = time[indiv,tp];
      agej = 1 || agejt || agejt#agejt;
      yjt = resp[indiv,tp];
      expAB = exp(b[1])#exp(agej[2]#b[2])#exp(agej[3]#b[3]);
      a0 = a0 + yjt - (expAB/(1+expAB));
      a1 = a1 + (yjt#agejt) - (agejt#expAB)/(1+expAB);
      a2 = a2 + (yjt#(agejt##2)) - ((agejt##2)#expAB)/(1+expAB);
    end;
    if gp = 1 then do;
      A0j_1[indiv] = a0; A1j_1[indiv] = a1; A2j_1[indiv] = a2;
    end;
    if gp = 2 then do;
      A0j_2[indiv] = a0; A1j_2[indiv] = a1; A2j_2[indiv] = a2;
      end;
    end;
  end;

  one = zj[,1]#((1-zj[,1])/pi[1] + zj[,2]/pi[2]);
  covsc_pi_beta1 = sum(A0j_1#one)||sum(A1j_1#one)||sum(A2j_1#one);

  two = zj[,2]#(zj[,1]/pi[1] + (1-zj[,2])/pi[2]);
  covsc_pi_beta2 = (-sum(A0j_2#two))||(-sum(A1j_2#two))||(-sum(A2j_2#two));

  one = zj[,1]#(1-zj[,1]);
  covsc_beta1 =
      ((sum((A0j_1##2)#one)||sum(A0j_1#A1j_1#one)||sum(A0j_1#A2j_1#one))//
   (sum(A0j_1#A1j_1#one)||sum((A1j_1##2)#one)||sum(A1j_1#A2j_1#one))//
   (sum(A0j_1#A2j_1#one)||sum(A1j_1#A2j_1#one)||sum((A2j_1##2)#one)));

  two = zj[,2]#(1-zj[,2]);
  covsc_beta2 =
      ((sum((A0j_2##2)#two)||sum(A0j_2#A1j_2#two)||sum(A0j_2#A2j_2#two))//
   (sum(A0j_2#A1j_2#two)||sum((A1j_2##2)#two)||sum(A1j_2#A2j_2#two))//
   (sum(A0j_2#A2j_2#two)||sum(A1j_2#A2j_2#two)||sum((A2j_2##2)#two)));

  k = zj[,1]#zj[,2];
  covsc_beta1_beta2 =
      -((sum(A0j_1#A0j_2#k)||sum(A0j_1#A1j_2#k)||sum(A0j_1#A2j_2#k))//
   (sum(A1j_1#A0j_2#k)||sum(A1j_1#A1j_2#k)||sum(A1j_1#A2j_2#k))//
   (sum(A2j_1#A0j_2#k)||sum(A2j_1#A1j_2#k)||sum(A2j_1#A2j_2#k)));

  Jm = ((covsc_pi||covsc_pi_beta1||covsc_pi_beta2)//
     (t(covsc_pi_beta1)||covsc_beta1||covsc_beta1_beta2)//
     (t(covsc_pi_beta2)||t(covsc_beta1_beta2)||covsc_beta2));

  A = shape(0,1,3); B = shape(0,3,3);

  Ic = ((Ic_pi||A||A)//
     (t(A)||Ic_beta1||B)//
     (t(A)||B||Ic_beta2));

  info_mat = Ic-Jm;
end;
if det(info_mat) = 0 then err = 1;
if err = 0 then do;
  cov_mat = inv(info_mat);
  var_est = diag(cov_mat);
end;
do w = 1 to 7 while (err = 0);
  if var_est[w,w] < 0 then err = 1;
end;
se_est = 0;
if err = 0 then do;
  se_est = sqrt(var_est);
end;
se_est = vecdiag(se_est);
pi_se = se_est[1];
beta1_se = se_est[2]//se_est[3]//se_est[4];
beta2_se = se_est[5]//se_est[6]//se_est[7];
beta_se = beta1_se||beta2_se;

/* Module RMISS */
/* http://www.psych.yorku.ca/lab/sas/iml.htm */
/* Remove rows with missing observations from matrix*/
start rmiss(mat1, mat2, miss);
   if nrow(miss)=0 then miss={.};
   badpos=loc(mat1=miss);
   badrow=ceil(badpos/ncol(mat1));
   keeprow=remove(1:nrow(mat1),badrow);
   mat2=mat1[keeprow,];
finish;

/* Find column averages for time */
avg = shape(0,m,1);
if miss = 1 then do;
  run rmiss(time, time_cc, miss);
end;
else time_cc = time;

do tp = 1 to m;
  do indiv = 1 to nrow(time_cc);
    avg[tp] = avg[tp] + time_cc[indiv,tp];
  end;
end;
avg = avg/n;
```

```
  /* Final results */
  if err = 1 then print 'Unsuccessful optimization termination';

  print 'Number of iterations:' iter;
  print 'Log-likelihood:' loglik;

  traj_curve = shape(0,m,g);
  do grp = 1 to g;
    do tp = 1 to m;
      t = 1 || avg[tp] || (avg[tp])##2;
      traj_curve[tp,grp] = exp(t*beta[,grp])/(1+exp(t*beta[,grp]));
    end;
  end;

  do group = 1 to g;
    print group;
    group_proportion = pi[group];
    group_proportion_SE = pi_se;
    beta_values = t(beta[,group]);
    beta_standard_error = t(beta_se[,group]);
    trajectory = t(traj_curve[,group]);
    print group_proportion;
    if group = 1 then print group_proportion_SE;
    print beta_values;
    print beta_standard_error;
    print trajectory;
  end;

quit;
%mend traj_model;
```

Bibliography

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle, in *Second International Symposium on Information Theory*, eds. B.N. Petrox and F. Caski. Budapest: Akademiai Kiado, p.267.

Aitkin, M. and Aitkin, I. (1996). A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing*, **6**, 127-130.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report R-97-021, University of California at Berkeley.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Cambridge: Springer.

Bollen, K. A. and Curran, P. J. (2006). *Latent Curve Models: a structural equation perspective.* New Jersey: John Wiley and Sons.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345-370.

Brown, K.S., Cameron, R., Madill, C., Payne, M.E., Filsinger, S., Manske, S.R. and Best, J.A. (2002). Outcome evaluation of a high school smoking reduction intervention based on extracurricular activities. *Preventive Medicine*, **35**, 506-510.

Casella, G. and Berger, R. L. (1990). *Statistical Inference.* California: Wadsworth.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**, 195-212.

Davenport, J. W., Pierce, M. A. and Hathaway, R. J. (1988). A numerical comparison of EM and Quasi-Newton type algorithms for computing MLE's for a mixture of normal distributions. *Proceedings of the 20th Symposium on the Interface: Computationally Intensive Methods in Statistics*, ed. Wegman, E. J., Gantz, D. T. and Miller, J. J., pp. 410-415.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463-474.

Demidenko, E. (2004). *Mixed Models: Theory and Applications.* New Jersey: John Wiley and Sons.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood for Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society B.* **39**, 1-38.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* New York: Oxford University Press.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research.* London: Arnold.

Driezen, P. (2001). The Development of Youth Smoking: Clusters of Initiation and Regular Smoking Trajectories. Unpublished M.Sc Thesis, Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada.

Everitt, B. S. (1981). A Monte Carlo investigation of the likelihood ratio test for number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, **16**, 171-180.

Everitt, B. S. (1988). A Monte Carlo investigation of the likelihood ratio test for number of classes in latent classes analysis. *Multivariate Behavioral Research*, **23**, 531-538.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions.* London: Chapman and Hall.

Finch, S. J., Mendell, N. R. and Thode, H. C. (1989). Probabilistic measures of adequacy of a numerical search for global maximum. *Journal of the American Statistical Association*, **84**, 1020-1023.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* New York: Springer.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics*, **8**, 431-444.

Hathaway, R. J. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statistics and Probability Letters*, **4**, 53-56.

Heckman, J. J., Robb, R., and Walker, J. R. (1990). Testing the mixture of exponentials hypothesis and estimating the mixture distribution by the method of moments. *Journal of the American Statistical Association*, **85**, 582-589.

Jones, B. L. and Nagin, D. S. (2007). Advances in group-based trajectory modeling and an SAS procedure for estimating them. *Sociological Methods and Research*, **35**, 542-571.

Jones, B. L., Nagin, D. S. and Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods and Research*, **29**, 374-393.

Jorgensen, M. A. (2004). Using multinomial mixture models to cluster Internet traffic. *Australian and New Zealand Journal of Statistics*, **46**, 205-218.

Karlis, D. and Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review*, **73**, 35-58.

Karp, I., O'Loughlin, J., Paradis, G., Hanley, J. and DiFranza, J. (2005). Smoking trajectories of adolescent novice smokers in a longitudinal study of tobacco use. *Annals of Epidemiology*, **15**, 445-452.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795.

Kelley, C. T. (2003). *Solving Nonlinear Equations with Newton's Method.* Philadelphia: Society for Industrial and Applied Mathematics.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya: The Indian Journal of Statistics A*, **62**, 49-66.

Lacourse, E., Nagin, D., Tremblay, R.E., Vitaro, F. and Claes, M. (2003). Developmental trajectories of boys' delinquent group membership and facilitation of violent behaviours during adolescence. *Development and Psychopathology*, **15**, 183-197.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, **20**, 1250-1360.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society B*, **44**, 226-233.

Maggi, S., Hertzman, C. and Vaillancourt, T. (2007). Changes in smoking behaviors from late childhood to adolescence: Insights from the Canadian National Longitudinal Survey of Children and Youth. *Health Psychology*, **26**(2), 232-240.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second Edition. Cambridge: Chapman and Hall.

McLachlan, G. J. and Krishnan T. (2008). *The EM Algorithm and Extensions*, Second Edition. New Jersey: John Wiley and Sons.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* Wiley Series in Probability and Statistics. New York: Wiley.

McNicholas, P. D., Murphy, T. B., McDaid, A. F. and Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*, **54**, 711-723.

Meng, X. L. (1997). The EM algorithm and medical studies: A historical link. *Statistical Methods in Medical Research*, **6**, 3-23.

Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, **86**, 899-909.

Muthén, B. and Muthén, L. K. (2007). *Mplus: User's guide*, Fifth Edition. Los Angeles: Muthén and Muthén.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, **55**, 463-469.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric group-based approach. *Psychological Methods*, **4**, 139-157.

Nagin, D. S. (2005). *Group-Based Modeling of Development.* Cambridge: Harvard University Press.

Nash, J. C. (1990). *Compact Numerical Methods for Computers: Linear algebra and function minimisation*, Second Edition. New York, NY: Adam Hilger.

Nawa, V. M. (2004). Analysis of developmental trajectories and binary longitudinal data. Unpublished Ph.D. Thesis, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada.

Nurmi, P. *Mixture Models.* Available: http://www.cs.helsinki.fi/u/salmenki/lda-seminaari04/mixturemodels.pdf. Last accessed March 2010.

Nylund, K. L., Asparouhov, T. and Muthn, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, **14**(4), 535-569.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A*, **185**, 71-110.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, **52**, 501-525.

Redner, R. A. and Walker, H. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195-239.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society A*, **158**, 73-89.

Roeder, K., Lynch, K. G., and Nagin, D. S. (1999). Modeling uncertainty in latent class membership: A case study in criminology. *Journal of the American Statistical Association*, 94, 766-776.

SAS Institute Inc. (2008). *SAS/IML 9.2 User's Guide.* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2009). *SAS/STAT 9.2 User's Guide*, Second Edition. Cary, NC: SAS Institute Inc.

Schwarz, G. (1978). Estimating the dimension of a Model. *Annals of Statistics*, **6**, 461-464.

Wang, P. and Puterman, M. L. (1998). Mixed Logistic Regression Models. *Journal of Agricultural, Biological, and Environmental Sciences*, **3**(2), 175-200.

White, H. R., Pandina, R. J. and Chen, P. H. (2002). Developmental trajectories of cigarette use from early adolescence into young adulthood. *Drug and Alcohol Dependence*, **65**, 167-178.

Wolfe, J. H. (1970). Pattern clustering of multivariate mixture analysis. *Multivariate Behavioral Research*, **5**, 329-350.

Yang, C. C. (2006). Evaluating latent class analyses in qualitative phenotype identification. *Computational Statistics and Data Analysis*, **50**, 1090-1104.