

A Mixture Model for Longitudinal Trajectories

Victor Mooto Nawa

Department of Mathematics and Statistics, University of Zambia, P.O. Box 32379, Lusaka, Zambia

Abstract An alternative method of estimating parameters in a mixture model for longitudinal trajectories using the expectation – maximization (EM) algorithm is proposed. Explicit expressions for the expectation and maximization steps required in the parameter estimation of group parameters and mixing proportions are derived. Expressions for the variances of group parameters and mixing proportions for the mixture model are also derived. Simulation results suggest that results from the proposed approach are less dependent on starting values and therefore a good alternative to the current approach which is based on the Quasi-Newton method.

Keywords Mixture model, Longitudinal trajectory, PROC TRAJ, Quasi-Newton, EM Algorithm

1. Introduction

Mixture models have been used extensively in a variety of important practical applications. In a mixture model, data can be viewed as arising from two or more populations mixed in varying proportions. There is a close link between clustering and finite mixture models. Cluster analysis is the search of groups of related observations in a data set. For example, in taxonomy clustering is used to identify subclasses of species. Fraley and Raftery (2002), outline a general methodology for model-based clustering. A number of researchers have studied finite mixture models in the context of clustering. In finite mixture models, each component probability distribution corresponds to a cluster. McLachlan and Basford (1988), highlighted the role of finite mixture distributions in modelling heterogeneous data, with the focus on applications in the field of cluster analysis. Their main focus was on mixtures of normal distributions. The problem of estimating parameters in a normal mixture distribution was first considered by Pearson (1894). Other researchers who have studied mixtures of normal distributions include Day (1969), Wolfe (1970), Marriot (1975) and Symons (1981). Mixtures of other distributions that have been considered by other researchers include exponential (Rider, 1961), beta (Bremmer, 1978), Weibull (Kao, 1959) and binomial (Blischke, 1962, 1964; Rider, 1962).

Most researchers have concentrated on the fitting of mixture models by a likelihood based approach using maximum likelihood estimation with the EM algorithm providing a convenient way for the iterative computation of solutions of the likelihood equation (Redner, et al, 1984;

Biernacki, et al, 2003; O'Hagan, et. al, 2012). However, other methods that have been used include method of moments, minimum chi-square, least squares and Bayesian methods. Day (1969), showed that the method of moments, minimum chi-square and Bayesian methods are inferior to the maximum likelihood method.

Under the mixture likelihood approach to clustering, it is assumed that the observations to be clustered are from a mixture of an initially specified number of populations or groups mixed in various proportions. This approach to clustering is model based in the sense that the form of the density of an observation in each of the underlying populations is specified. One advantage of model based clustering is that it provides a specific framework for assessing the resulting partitions of the data and especially for choosing a relevant number of clusters (Biernacki, et al, 2000). By assuming some parametric form for the density function in each of the underlying groups, a likelihood is formed in terms of the mixture density and unknown parameters are estimated by maximum likelihood estimation. A probabilistic clustering of the observations is obtained in terms of their estimated probabilities of group membership. The assignment of the observations to the groups is done by allocating each observation to the group to which it has the highest probability of belonging.

In this paper we present a mixture of developmental trajectories. Our focus will be on the semiparametric group based model proposed by Nagin (1999). The model assumes that the population is composed of a mixture of distinct groups defined by their developmental trajectories. This model is designed to identify rather than assume distinctive groups of trajectories, estimate the proportion of the population following each trajectory, relate group membership to individual characteristics and create profiles of group members using group membership probabilities (Nagin, 1999). The model presumes that two types of variables have been measured: response variable and

* Corresponding author:

vnawa@yahoo.com (Victor Mooto Nawa)

Published online at <http://journal.sapub.org/statistics>

Copyright © 2014 Scientific & Academic Publishing. All Rights Reserved

covariates or risk factors.

Jones, Nagin and Roeder (2001), wrote a SAS based procedure that can be used to estimate the parameters of this semiparametric model. The parameters are obtained by maximum likelihood estimation using the general Quasi-Newton procedure. Standard errors are obtained by inverting the observed information matrix. This software is a customized SAS procedure that was developed with the SAS product SAS/TOOLKIT. The SAS procedure is called is called PROC TRAJ.

Whereas, the semiparametric model can be used to model three data types – count, binary and psychometric scale data, we will concentrate on binary data. Roeder et al. (1999) concentrated on count data using the EM algorithm. This paper will concentrate on binary longitudinal data using the EM algorithm. While Roeder et al. (1999) looked at a model that incorporated covariates, this paper will concentrate on a model without covariates.

The remainder of this paper is organised as follows. The model under discussion is presented in Section 2. This section begins with a discussion of the standard mixture model in Section 2.1. The longitudinal mixture model, which is the main subject of this paper, is presented in Section 2.2. Section 2.2.1 gives the likelihood formulation of the longitudinal mixture model and the E-steps and M-steps required to estimate the group parameters as well as the mixing proportions in the model. This is followed by a discussion on estimation of standard errors for the group parameters and mixing proportions in Section 2.2.2. Simulation results are presented in Section 3 and conclusions are presented in Section 4.

2. The Model

2.1. Standard Mixture Model

A standard g -component mixture model (McLachlan and Basford, 1988) takes the form

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

where y_j is an observation on a random variable Y_j and $\psi = (\theta_1, \theta_2, \dots, \theta_g, \pi_1, \pi_2, \dots, \pi_{g-1})$ is a vector of all unknown parameters. The underlying population is modelled as consisting of g distinct groups C_1, C_2, \dots, C_g in some unknown proportions $\pi_1, \pi_2, \dots, \pi_g$ and where the conditional function of Y_j given membership of the i^{th} group C_i is $f_i(y_j; \theta_i)$.

Let $y = (y_1^T, y_2^T, \dots, y_n^T)^T$ be an observed random sample from the mixture density in (1), then the likelihood function for ψ based on the observed data can be written as

$$L(\psi) = \prod_{j=1}^n \sum_{i=1}^g \pi_i f_i(y_j; \theta_i) \quad (2)$$

To estimate the parameter vector ψ we either maximize the likelihood in (2) or maximize the loglikelihood given in (3) below directly.

$$\log L(\psi) = \sum_{j=1}^n \log \left\{ \sum_{i=1}^g \pi_i f_i(y_j; \theta_i) \right\} \quad (3)$$

The likelihood in (2) can also be maximized using the EM (expectation maximization) algorithm. This is done by first introducing an observable or missing data vector $z = (z_1^T, z_2^T, \dots, z_n^T)^T$. The vector of indicator variables $z_j = (z_{1j}, z_{2j}, \dots, z_{gj})^T$ is defined by

$$z_{ij} = \begin{cases} 1 & , \text{ if } y_j \in C_i \\ 0 & , \text{ otherwise} \end{cases}$$

Instead of maximizing the likelihood in (2) above, the EM – algorithm can be used to maximize the likelihood in (4) or the loglikelihood in (5) which is often referred to as the complete-data loglikelihood.

$$L_c(\psi) = \prod_{j=1}^n \prod_{i=1}^g \pi_i^{z_{ij}} f_i(y_j; \theta_i)^{z_{ij}} \quad (4)$$

$$\log L_c(\psi) = \sum_{i=1}^g \sum_{j=1}^n z_{ij} \log \pi_i + \sum_{i=1}^g \sum_{j=1}^n \log f_i(y_j; \theta_i) \quad (5)$$

2.2. Longitudinal Mixture Model

2.2.1. Likelihood Formulation and Parameter Estimation

Nagin (1999) uses a polynomial relationship model to relate age to the response. Let y_{jt} be a binary response for subject j at time t . It is assumed that conditional on membership in group i

$$\Pr(Y_{jt} = 1) = \frac{\exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)}{1 + \exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)} \quad (6)$$

where $\Pr(Y_{jt} = 1)$ is the probability of the outcome of interest and a_{jt} is the age of subject j at time t . The

coefficients β_0^i , β_1^i and β_2^i determine the shape of the trajectory and are superscripted by i to indicate that the coefficients are allowed to vary across the different groups.

Let $y_j = (y_{j1}, y_{j2}, \dots, y_{jm})$ denote the longitudinal sequence of observations for subject j . Conditional on being in group i , a subject's longitudinal observations are assumed independent, thus

$$f_i(y_j; \beta_i) = \prod_{t=1}^m \left(\frac{\exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)}{1 + \exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)} \right)^{y_{jt}} \left(\frac{1}{1 + \exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)} \right)^{1-y_{jt}} \quad (7)$$

where a_{jt} is the age of subject j at time t and y_{jt} is the response of subject j at time t (which is either zero or one). As described by Nagin (1999), the form of the likelihood for each subject j is given by

$$f(y_j; \psi) = \sum_{i=1}^g \pi_i f_i(y_j; \beta_i) \quad (8)$$

where $f(y_j; \psi)$ is the unconditional probability of observing subject j 's longitudinal measurements, $f_i(y_j; \beta_i)$ is the probability of y_j given membership in group i . The likelihood for the entire sample of n subjects is given by

$$L(\psi) = \prod_{j=1}^n \sum_{i=1}^g \pi_i f_i(y_j; \beta_i) \quad (9)$$

where g is the number of groups and π_i is the probability of membership in group i . Each $\beta_i = (\beta_0^i, \beta_1^i, \beta_2^i)$ is a set of parameters describing the i^{th} group and $\psi = (\theta_1, \theta_2, \dots, \theta_g, \pi_1, \pi_2, \dots, \pi_g)$ is the set of all parameters. Being probabilities ($\pi_1, \pi_2, \dots, \pi_g$) must satisfy the constraints

$$\pi_i \geq 0, \quad i = 1, 2, \dots, g \quad \text{and} \quad \sum_{i=1}^g \pi_i = 1$$

The likelihood in (9) can be maximized directly using PROC TRAJ. Maximization of this likelihood using the EM – algorithm requires introducing the missing data indicator z and obtaining the complete-data likelihood in (10).

$$L_c(\psi) = \prod_{j=1}^n \prod_{i=1}^g \pi_i^{z_{ij}} f_i(y_j; \beta_i)^{z_{ij}} \quad (10)$$

Substituting (7) into (10) and finding the log of the result gives the complete-data loglikelihood in (11).

$$l_c(\psi) = \sum_{j=1}^n \sum_{i=1}^g z_{ij} \log \pi_i + \sum_{j=1}^n \sum_{i=1}^g z_{ij} \left\{ \sum_{t=1}^m y_{jt} (\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2) - \log (1 + \exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)) \right\} \quad (11)$$

The E-step (expectation step) on the $(k+1)^{\text{th}}$ involves evaluating $E(l_c(\psi | y; \psi^{(k)}))$, where $\psi^{(k)} = (\beta_1^{(k)T}, \beta_2^{(k)T}, \dots, \beta_g^{(k)T}, \pi_1^{(k)}, \pi_2^{(k)}, \dots, \pi_{g-1}^{(k)T})$. This expectation only requires the evaluation of $E(Z_{ij} | y_j; \psi^{(k)})$ which is given by

$$\begin{aligned} E(Z_{ij} | y_j; \psi^{(k)}) &= \frac{\pi_i^{(k)} f_i(y_j; \beta_i^{(k)})}{f(y_j; \psi^{(k)})} \\ &= z_{ij}^{(k)} \end{aligned} \quad (12)$$

On the $(k+1)^{\text{th}}$ iteration, the M-step constitutes finding a value of ψ , that maximizes the expected loglikelihood, thus

$$\psi^{(k+1)} = \arg \max_{\psi} E(l_c(\psi | y; \psi^{(k)})) \quad (13)$$

This is given by

$$\begin{aligned}\pi_i^{(k+1)} &= \frac{1}{n} \sum_{j=1}^n z_{ij}^{(k)} \\ \beta_i^{(k+1)} &= \arg \max_{\beta_i} \sum_{j=1}^n z_{ij}^{(k)} \left\{ \sum_{t=1}^m y_{jt} (A_{jt}^T \beta_i^{(k)}) - \log(1 + \exp(A_{jt}^T \beta_i^{(k)})) \right\}\end{aligned}\quad (14)$$

where $A_{jt}^T = (1 \ a_{jt} \ a_{jt}^2)$ so that $A_{jt}^T \beta_i^{(k)} = \beta_0^{i(k)} + \beta_1^{i(k)} a_{jt} + \beta_2^{i(k)} a_{jt}^2$ for $i = 1, 2, \dots, g$. Since there is no closed form solution for $\beta_i^{(k+1)}$, the maximization requires iteration. Starting from some initial parameter value $\psi^{(0)}$, the E- and M-steps are repeated until convergence.

2.2.2. Estimation of Standard Errors

Standard errors of the parameter estimates are obtained from the inverse of the observed matrix. Unlike Newton-type methods, the EM-algorithm does not automatically provide an estimate of the covariance matrix of the maximum likelihood estimates. Louis (1982) derived a procedure for extracting the observed information matrix from the complete data loglikelihood when the EM-algorithm is used to find maximum likelihood estimates. McLachlan and Krishnan (1997) demonstrated the use of this procedure on a number of examples including its use on a mixture of standard normal distributions.

According to the procedure, the observed information matrix $I(\hat{\psi})$ is computed as

$$I(\hat{\psi}; y) = J_c(\hat{\psi}; y) - J_m(\hat{\psi}; y) \quad (15)$$

where

$$J_c(\psi; y) = E[I_c(\psi | y)] \quad (16)$$

is the conditional expectation of the complete-data information matrix $I_c(\psi)$ given y and

$$J_m(\psi; y) = \text{cov}[S_c(\psi | y)] \quad (17)$$

where $S_c(\psi)$ is the score vector based on the complete-data loglikelihood. The complete-data loglikelihood for a g -component mixture in (11) can be written in the form

$$l_c(\psi) = \sum_{j=1}^n \left\{ \sum_{i=1}^{g-1} z_{ij} \log \pi_i + z_{gj} \log \left(1 - \sum_{i=1}^{g-1} \pi_i \right) \right\} + \sum_{j=1}^n \sum_{i=1}^g z_{ij} \log f_i(y_j; \beta_i) \quad (18)$$

From (18) the complete-data information $I_c(\psi)$ is obtained as a $4g - 1$ by $4g - 1$ block-diagonal matrix given by

$$I_c(\psi) = \begin{pmatrix} I_c(\pi) & 0 & 0 & \dots & 0 \\ 0 & I_c(\beta_1) & 0 & \dots & 0 \\ 0 & 0 & I_c(\beta_2) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & I_c(\beta_g) \end{pmatrix} \quad (19)$$

where $I_c(\pi)$ is a $g - 1$ by $g - 1$ matrix with diagonal elements

$$-\frac{\partial^2 l}{\partial \pi_i^2} = \sum_{j=1}^n \left(\frac{z_{ij}}{\pi_i^2} + \frac{z_{gj}}{\pi_g^2} \right), \quad i = 1, 2, \dots, g-1, \quad (20)$$

and off-diagonal elements

$$-\frac{\partial^2 l}{\partial \pi_i \partial \pi_k} = \sum_{j=1}^n \frac{z_{gj}}{\pi_g^2}, \quad i \neq k. \quad (21)$$

Each $I_c(\beta_i)$ is a 3 by 3 matrix given by

$$I_c(\beta_i) = \begin{pmatrix} -\frac{\partial^2 l}{\partial \beta_0^{i2}} & -\frac{\partial^2 l}{\partial \beta_0^i \partial \beta_1^i} & -\frac{\partial^2 l}{\partial \beta_0^i \partial \beta_2^i} \\ -\frac{\partial^2 l}{\partial \beta_1^i \partial \beta_0^i} & -\frac{\partial^2 l}{\partial \beta_1^{i2}} & -\frac{\partial^2 l}{\partial \beta_1^i \partial \beta_2^i} \\ -\frac{\partial^2 l}{\partial \beta_2^i \partial \beta_0^i} & -\frac{\partial^2 l}{\partial \beta_2^i \partial \beta_1^i} & -\frac{\partial^2 l}{\partial \beta_2^{i2}} \end{pmatrix} \quad (22)$$

where

$$\begin{aligned} -\frac{\partial^2 l}{\partial \beta_0^{i2}} &= \sum_{j=1}^n z_{ij} \left\{ \sum_{t=1}^m \left(\frac{\exp(A_{jt}^T \beta_i)}{(1 + \exp(A_{jt}^T \beta_i))^2} \right) \right\} \\ -\frac{\partial^2 l}{\partial \beta_0^i \partial \beta_1^i} &= \sum_{j=1}^n z_{ij} \left\{ \sum_{t=1}^m \left(\frac{a_{jt} \exp(A_{jt}^T \beta_i)}{(1 + \exp(A_{jt}^T \beta_i))^2} \right) \right\} \\ -\frac{\partial^2 l}{\partial \beta_0^i \partial \beta_2^i} &= -\frac{\partial^2 l}{\partial \beta_1^{i2}} = \sum_{j=1}^n z_{ij} \left\{ \sum_{t=1}^m \left(\frac{a_{jt}^2 \exp(A_{jt}^T \beta_i)}{(1 + \exp(A_{jt}^T \beta_i))^2} \right) \right\} \\ -\frac{\partial^2 l}{\partial \beta_1^i \partial \beta_2^i} &= \sum_{j=1}^n z_{ij} \left\{ \sum_{t=1}^m \left(\frac{a_{jt}^3 \exp(A_{jt}^T \beta_i)}{(1 + \exp(A_{jt}^T \beta_i))^2} \right) \right\} \\ -\frac{\partial^2 l}{\partial \beta_2^{i2}} &= \sum_{j=1}^n z_{ij} \left\{ \sum_{t=1}^m \left(\frac{a_{jt}^4 \exp(A_{jt}^T \beta_i)}{(1 + \exp(A_{jt}^T \beta_i))^2} \right) \right\} \end{aligned}$$

The score vector $S_c(\psi)$ based on the complete-data loglikelihood is given by

$$S_c(\psi) = (S_c(\pi)^T, S_c(\beta_1)^T, S_c(\beta_2)^T, \dots, S_c(\beta_g)^T)^T \quad (23)$$

where

$$S_c(\pi) = \left(\frac{\partial l}{\partial \pi_1}, \frac{\partial l}{\partial \pi_2}, \dots, \frac{\partial l}{\partial \pi_{g-1}} \right) \quad (24)$$

and

$$\frac{\partial l}{\partial \pi_i} = \sum_{j=1}^n \left(\frac{z_{ij}}{\pi_i} - \frac{z_{gj}}{\pi_g} \right), \quad i = 1, 2, \dots, g-1 \quad (25)$$

The score vector for each group $S_c(\beta_i)$ is given by

$$\frac{\partial l}{\partial \beta_i} = \begin{pmatrix} \frac{\partial l}{\partial \beta_0^i} \\ \frac{\partial l}{\partial \beta_1^i} \\ \frac{\partial l}{\partial \beta_2^i} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^m z_{ij} \left(y_{jt} - \frac{\exp(A_{jt}^T \beta_i)}{1 + \exp(A_{jt}^T \beta_i)} \right) \\ \sum_{t=1}^m z_{ij} \left(y_{jt} a_{jt} - \frac{a_{jt} \exp(A_{jt}^T \beta_i)}{1 + \exp(A_{jt}^T \beta_i)} \right) \\ \sum_{t=1}^m z_{ij} \left(y_{jt} a_{jt}^2 - \frac{a_{jt}^2 \exp(A_{jt}^T \beta_i)}{1 + \exp(A_{jt}^T \beta_i)} \right) \end{pmatrix} \quad (26)$$

The conditional covariance of the score vector $S_c(\psi)$ is given by

$$J_m(\psi) = \begin{pmatrix} \text{cov}(S_c(\pi)) & \text{cov}(S_c(\pi), S_c(\beta_1)) & \dots & \text{cov}(S_c(\pi), S_c(\beta_g)) \\ \text{cov}(S_c(\beta_1), S_c(\pi)) & \text{cov}(S_c(\beta_1)) & \dots & \text{cov}(S_c(\beta_1), S_c(\beta_g)) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(S_c(\beta_g), S_c(\pi)) & \text{cov}(S_c(\beta_g), S_c(\beta_1)) & \dots & \text{cov}(S_c(\beta_g)) \end{pmatrix} \quad (27)$$

where $\text{cov}(S_c(\pi))$ a $g-1$ by $g-1$ matrix, $\text{cov}(S_c(\pi), S_c(\beta_i))$ is a $g-1$ by 3 matrix and $\text{cov}(S_c(\beta_i))$ is a 3 by 3 matrix for $i = 1, 2, \dots, g$. Since we are taking the conditional covariance of the score vector given the vector y , we only need to evaluate $E(Z_{ij})$, $\text{Var}(Z_{ij})$ and $\text{Cov}(Z_{ij}, Z_{kj})$ for $i \neq k$, where Z_{ij} is a random variable corresponding to z_{ij} . Let

$$E(Z_{ij} | y_j; \psi) = \frac{\pi_i f_i(y_j; \beta_i)}{f(y_j; \psi)} = \tau_{ij} \quad (28)$$

then

$$\text{Var}(Z_{ij}) = \tau_{ij}(1 - \tau_{ij}) \quad (29)$$

and

$$\text{Cov}(Z_{ij}, Z_{kj}) = -\tau_{ij}\tau_{kj}, \quad i \neq k \quad (30)$$

Using the expressions (28) to (30), the diagonal elements of the matrix $\text{cov}(S_c(\pi))$ can be expressed as

$$\text{cov}(S_c(\pi))_{ii} = \sum_{j=1}^n \left(\frac{\tau_{ij}(1 - \tau_{ij})}{\pi_i^2} + \frac{\tau_{gj}(1 - \tau_{gj})}{\pi_g^2} + \frac{2\tau_{ij}\tau_{gj}}{\pi_i\pi_g} \right) \quad (31)$$

while the off diagonal elements can be expressed as

$$\text{cov}(S_c(\pi))_{ik} = \sum_{j=1}^n \left(-\frac{\tau_{ij}\tau_{kj}}{\pi_i\pi_k} + \frac{\tau_{ij}\tau_{gj}}{\pi_i\pi_g} + \frac{\tau_{kj}\tau_{gj}}{\pi_k\pi_g} + \frac{\tau_{gj}(1 - \tau_{gj})}{\pi_g^2} \right), \quad i \neq k \quad (32)$$

Similarly, the k^{th} row of the matrix $\text{cov}(S_c(\pi), S_c(\beta_i))$ is given by

$$\begin{aligned}
\text{cov}(S_c(\pi), S_c(\beta_i)) &= \sum_{j=1}^n B_{pj}^i \tau_{ij} \left(\frac{1 - \tau_{kj}}{\pi_k} + \frac{\tau_{gj}}{\pi_g} \right), \quad p = 0, 1, 2; \quad k = 1, 2, \dots, g-1 \text{ (for } k = i) \\
&= \sum_{j=1}^n B_{pj}^i \tau_{ij} \left(-\frac{\tau_{kj}}{\pi_k} + \frac{\tau_{gj}}{\pi_g} \right), \quad p = 0, 1, 2; \quad k = 1, 2, \dots, g-1 \text{ (for } k \neq i) \\
&= -\sum_{j=1}^n B_{pj}^i \tau_{ij} \left(\frac{\tau_{kj}}{\pi_k} + \frac{1 - \tau_{gj}}{\pi_g} \right), \quad p = 0, 1, 2; \quad k = 1, 2, \dots, g-1 \text{ (for } i = g)
\end{aligned} \tag{33}$$

where

$$\begin{aligned}
B_{0j}^i &= \sum_{t=1}^m \left(y_{jt} - \frac{\exp(A_{jt}^T \beta_i)}{1 + \exp(A_{jt}^T \beta_i)} \right) \\
B_{1j}^i &= \sum_{t=1}^m \left(y_{jt} a_{jt} - \frac{a_{jt} \exp(A_{jt}^T \beta_i)}{1 + \exp(A_{jt}^T \beta_i)} \right) \\
B_{2j}^i &= \sum_{t=1}^m \left(y_{jt} a_{jt}^2 - \frac{a_{jt}^2 \exp(A_{jt}^T \beta_i)}{1 + \exp(A_{jt}^T \beta_i)} \right)
\end{aligned} \tag{34}$$

The covariance matrices $\text{cov}(S_c(\beta_i))$ and $\text{cov}(S_c(\beta_i), S_c(\beta_k))$ (for $i = 1, 2, \dots, g$ and $i \neq k$) are respectively given by

$$\text{cov}(S_c(\beta_i)) = \begin{pmatrix} \sum_{j=1}^n (B_{0j}^i)^2 \tau_{ij} (1 - \tau_{ij}) & \sum_{j=1}^n B_{0j}^i B_{1j}^i \tau_{ij} (1 - \tau_{ij}) & \sum_{j=1}^n B_{0j}^i B_{2j}^i \tau_{ij} (1 - \tau_{ij}) \\ \sum_{j=1}^n B_{1j}^i B_{0j}^i \tau_{ij} (1 - \tau_{ij}) & \sum_{j=1}^n (B_{1j}^i)^2 \tau_{ij} (1 - \tau_{ij}) & \sum_{j=1}^n B_{1j}^i B_{2j}^i \tau_{ij} (1 - \tau_{ij}) \\ \sum_{j=1}^n B_{2j}^i B_{0j}^i \tau_{ij} (1 - \tau_{ij}) & \sum_{j=1}^n B_{2j}^i B_{1j}^i \tau_{ij} (1 - \tau_{ij}) & \sum_{j=1}^n (B_{2j}^i)^2 \tau_{ij} (1 - \tau_{ij}) \end{pmatrix} \tag{35}$$

and

$$\text{cov}(S_c(\beta_i), S_c(\beta_k)) = - \begin{pmatrix} \sum_{j=1}^n B_{0j}^i B_{0j}^k \tau_{ij} \tau_{kj} & \sum_{j=1}^n B_{0j}^i B_{1j}^k \tau_{ij} \tau_{kj} & \sum_{j=1}^n B_{0j}^i B_{2j}^k \tau_{ij} \tau_{kj} \\ \sum_{j=1}^n B_{1j}^i B_{0j}^k \tau_{ij} \tau_{kj} & \sum_{j=1}^n B_{1j}^i B_{1j}^k \tau_{ij} \tau_{kj} & \sum_{j=1}^n B_{1j}^i B_{2j}^k \tau_{ij} \tau_{kj} \\ \sum_{j=1}^n B_{2j}^i B_{0j}^k \tau_{ij} \tau_{kj} & \sum_{j=1}^n B_{2j}^i B_{1j}^k \tau_{ij} \tau_{kj} & \sum_{j=1}^n B_{2j}^i B_{2j}^k \tau_{ij} \tau_{kj} \end{pmatrix} \tag{36}$$

After running the EM algorithm until convergence, the parameter estimate $\hat{\psi}$ and estimated posterior probabilities $\hat{\tau}_{ij}$ are obtained. The variance estimate for $\hat{\psi}$ is obtained from the inverse of the matrix $J_c(\hat{\psi}; y) - J_m(\hat{\psi}; y)$ with $E(Z_{ij})$ in $J_c(\hat{\psi}; y)$ and τ_{ij} in $J_m(\hat{\psi}; y)$ replaced by $\hat{\tau}_{ij}$.

3. Simulation Results

We consider simulation results for mixtures of two and three groups of longitudinal trajectories. Consider three groups of longitudinal trajectories with group parameters $\beta_1 = (6.17, -5.78, 0.997)$, $\beta_2 = (-7.69, 6.59, -1.099)$ and $\beta_3 = (-2.24, -0.17, 0.21)$. These parameter values are calculated based on the trajectory shapes given in Figure 1 below. We consider a situation involving five time points where measurements are taken at times 1 to 5 (i.e. $a_{j1} = 1$, $a_{j2} = 2$, $a_{j3} = 3$, $a_{j4} = 4$, $a_{j5} = 5$). Using the above parameter estimates, a sequence of binary responses are generated for time points 1 to 5. Since the responses are assumed to be independent at each time point, the binary responses are generated independently at each time point. The binary response for the j th subject from group i at time t is generated with success probability

$$p = \frac{\exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)}{1 + \exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)} \quad (37)$$

For example, observations from group one at time point three are generated with success probability

$$p = \frac{\exp(6.17 - 5.78 \times 3 + 0.997 \times 3^2)}{1 + \exp(6.17 - 5.78 \times 3 + 0.997 \times 3^2)} \quad (38)$$

In the theoretical longitudinal trajectory groups given in Figure 1 below, time is plotted against the probability of having the characteristic of interest. If smoking was the characteristic of interest, for example, group one would represent a group of smokers who temporarily stop smoking for some time and then continue smoking again while group three would represent a group of subjects who are initially non-smokers who start smoking and continue smoking throughout the course of the study.

In the simulation results that follow, results obtained using the EM algorithm explained in section 2 and PROC TRAJ are presented. For all the results obtained from the EM algorithm we present, the convergence criterion is based on changes in the maximized loglikelihood values. We stop the steps for the EM algorithm when the difference in the

loglikelihoods is less than 10^{-8} for the model involving a mixture of two trajectory groups and 10^{-3} for the model involving a mixture of three trajectory groups.

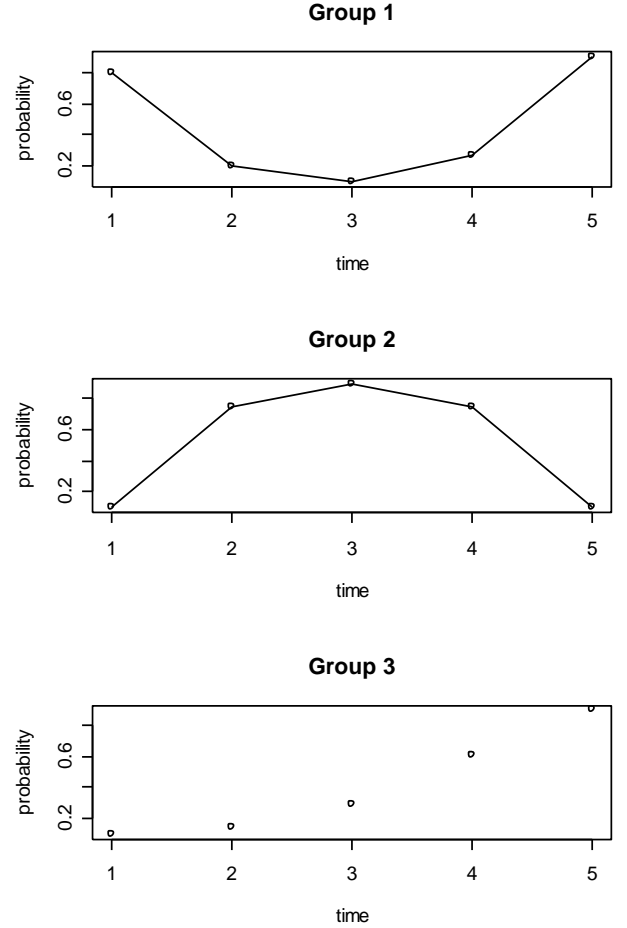


Figure 1. Longitudinal trajectories for three groups

3.1. Mixture of Two Trajectory Groups

Consider 100 simulations of data sets of 500 observations from a mixture of two trajectory groups with group parameters $\beta_1 = (6.17, -5.78, 0.997)$ and $\beta_2 = (-7.69, 6.59, -1.099)$. For each simulation a total of 500 observations are generated – observations are generated from group one with success probability 0.32. The results are given in Table 1 and Table 2 below. Table 1 shows group parameter estimates, standard error estimates and empirical standard error estimates obtained from the EM algorithm and the PROC TRAJ procedure. The empirical standard errors reported in the table are sample standard deviations of the actual parameter estimates. Table 2 shows the estimates for the mixing proportions.

As can be seen from the tables, the estimated group parameter estimates and mixing proportion parameter estimates are very close to the theoretical values. It can also be seen that the standard errors obtained are comparable to the empirical standard errors.

Table 1. Group parameter estimates and standard errors based on 100 simulations

| | Group | | | | | |
|--------------------------|-----------|-----------|-----------|-----------|-----------|-----------|
| | 1 | | | 2 | | |
| Parameter | β_0 | β_1 | β_2 | β_0 | β_1 | β_2 |
| Theoretical | 6.170 | -5.781 | 0.997 | -7.690 | 6.590 | -1.099 |
| Estimate (EM) | 6.230 | -5.831 | 1.007 | -7.785 | 6.658 | -1.110 |
| Estimate (PROC TRAJ) | 6.225 | -5.828 | 1.006 | -7.786 | 6.660 | -1.109 |
| SE (EM) | 0.576 | 0.485 | 0.084 | 0.447 | 0.344 | 0.057 |
| Empirical SE (EM) | 0.565 | 0.474 | 0.088 | 0.494 | 0.380 | 0.063 |
| SE (PROC TRAJ) | 0.576 | 0.485 | 0.084 | 0.447 | 0.344 | 0.057 |
| Empirical SE (PROC TRAJ) | 0.565 | 0.475 | 0.082 | 0.497 | 0.382 | 0.063 |

Table 2. Average proportion parameter estimates based on 100 simulations

| | π_1 | π_2 |
|-------------------------|---------|---------|
| Theoretical | 0.32 | 0.68 |
| Estimate (EM algorithm) | 0.322 | 0.678 |
| Estimate (PROC TRAJ) | 0.322 | 0.678 |

Table 3. Average number of steps/iterations required to get the loglikelihood correct to 0 to 5 decimal places for the EM Algorithm compared to PROC TRAJ based on 100 simulations

| | Decimal places | | | | | |
|--------------|----------------|-------|-------|-------|-------|-------|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| EM algorithm | 8.13 | 10.21 | 12.4 | 16.89 | 28.37 | 38.4 |
| PROC TRAJ | 23.69 | 24.87 | 25.98 | 26.97 | 27.8 | 28.63 |

Table 3 shows the average number of EM steps and PROC TRAJ iterations required to obtain the maximized loglikelihood value correct to between 0 and 5 decimal places. From the table we see that the EM algorithm takes an average of about 8 steps to get to the integer part of the maximized loglikelihood value. In fact if we were only interested in obtaining the maximized loglikelihood value correct to two decimal places, the table shows that on average the EM algorithm requires only two additional steps after getting the integer part of the maximized loglikelihood value. On the other hand, the table shows that on average the EM algorithm requires 30 more steps to get from within zero to five decimal places of the maximized loglikelihood value. The table further shows that PROC TRAJ requires 23 iterations on average to get to the integer part of the maximized loglikelihood value and only an additional five iterations on average to get from within zero to five decimal places of the maximized loglikelihood value. Thus PROC TRAJ has a faster convergence rate once the procedure is within one or two decimal places of the maximized loglikelihood value.

3.2. Mixture of Three Trajectory Groups

Unlike a mixture of two trajectory groups, a model with three groups is a little more complex as the results obtained tend to depend on the starting values used. This is especially more pronounced when the PROC TRAJ procedure is used. The EM algorithm tends to be a little less sensitive to starting values. As such starting values for three or more trajectory

groups have to be carefully chosen.

A procedure of coming up with starting values based on the logistic regression model is proposed. We propose dividing the data into groups based on the criterion described below and fitting a logistic regression model in each group and using the parameter values as starting values. For example, in a three group model, the data are initially divided into three groups and a logistic regression model is used to estimate the parameters in each of the three groups.

The initial division of the data is based on the response at each of the five time points. One possible way of dividing the data is by looking at the response at two distinct time points. Since we are dealing with binary responses, there are four different possible patterns that naturally arise. The patterns are given in Table 4 below:

Table 4. Possible patterns for binary data at two distinct time points

| Pattern | Time1 | Time 2 |
|---------|-------|--------|
| 1 | 0 | 0 |
| 2 | 0 | 1 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |

If we were interested in fitting a model with four groups, we would make the initial division of the data based on the four patterns in the above table. For a three group model, one possible way of doing the division is by dividing the data into four groups and then merging the two smallest groups. One possible way of carrying out the initial division can be based on the first and last response values. To fit a model with more than four groups, we can look at three or more time points and then divide the data appropriately. After the data have been divided into groups, the proportions of the number of observations in each group can be used as the starting values for the proportions. For each group i , we fit a logistic regression model

$$\Pr(Y_{jt} = 1) = \frac{\exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)}{1 + \exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)} \quad (38)$$

where a_{jt} are the times at which observations are made on subject j . The parameter estimates $\hat{\beta}_0^i$, $\hat{\beta}_1^i$ and $\hat{\beta}_2^i$ are then used as starting values for group i .

Table 5. Group parameter estimates and standard errors based on 150 simulations

| Parameter | β_0 | β_1 | β_2 | β_0 | β_1 | β_2 | β_0 | β_1 | β_2 |
|--------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| Theoretical | 6.170 | -5.781 | 0.997 | -7.690 | 6.590 | -1.099 | -2.237 | -0.172 | 0.211 |
| Estimate (EM) | 6.668 | -6.127 | 1.050 | -7.786 | 6.677 | -1.113 | -2.660 | 0.069 | 0.182 |
| Estimate (PROC TRAJ) | 6.996 | -6.352 | 1.082 | -7.782 | 6.671 | -1.112 | -2.777 | 0.133 | 0.176 |
| SE (EM) | 2.762 | 1.877 | 0.274 | 0.510 | 0.435 | 0.077 | 1.881 | 1.196 | 0.174 |
| Empirical SE (EM) | 2.301 | 1.543 | 0.226 | 0.510 | 0.416 | 0.073 | 1.740 | 1.082 | 0.157 |
| SE (PROC TRAJ) | 2.908 | 2.000 | 0.292 | 0.510 | 0.434 | 0.077 | 1.848 | 1.168 | 0.171 |
| Empirical SE (PROC TRAJ) | 2.973 | 1.982 | 0.286 | 0.513 | 0.419 | 0.074 | 2.171 | 1.330 | 0.189 |

A total of 150 data sets of 800 observations from the three groups with parameters $\beta_1 = (6.17, -5.78, 0.997)$, $\beta_2 = (-7.69, 6.59, -1.099)$ and $\beta_3 = (-2.24, -0.17, 0.21)$ with mixing proportions $\pi_1 = 0.2$, $\pi_2 = 0.425$ and $\pi_3 = 0.375$ respectively, were considered. A closer look at the simulation results showed that the EM algorithm attains convergence to the first few decimal places of the maximised loglikelihood values quickly and then takes many steps to get the loglikelihood value correct to third and fourth decimal places. For all practical purposes we do not need to have a maximised loglikelihood value correct to five decimal places. As such the convergence criteria for the EM algorithm used for the three group model was 10^{-3} . This means iterations in the procedure continues until the increase in the loglikelihood is less than 10^{-3} . The simulation results are given in the following Table 5 and Table 6.

Table 6. Proportion parameter estimates obtained using the EM algorithm and PROC TRAJ compared to true parameter values

| | π_1 | π_2 | π_3 |
|-------------------------|---------|---------|---------|
| Theoretical | 0.2 | 0.425 | 0.375 |
| Estimate (EM algorithm) | 0.208 | 0.416 | 0.376 |
| Estimate (PROC TRAJ) | 0.208 | 0.416 | 0.376 |

The results given in Table 5 and Table 6 show that the group parameter estimates and mixing proportion estimates obtained are close to the theoretical values though not as close as those obtained for the two parameter model given in Table 1 and Table 2. This may be as a result of the complexity of the model involved; the three group model is more complex than a two group model. The estimated standard errors obtained are comparable to the empirical standard errors.

4. Conclusions

In this article we have proposed the use of the EM algorithm to estimate parameters for a model involving a mixture of longitudinal trajectories. While the original mixture model of longitudinal trajectories (Nagin, 1999) can be applied to count, binary and psychometric data, this article only considered binary longitudinal data. Parameter

estimation in this model involves estimating the mixing proportions, group parameter estimates and the corresponding standard errors. This is currently done using the Quasi-Newton method through a SAS procedure called PROC TRAJ as proposed by Jones et al. (2001). This paper therefore proposes an alternative way of carrying out the parameter estimation.

The paper describes how to estimate the various parameters in the longitudinal trajectory mixture model using the EM algorithm. This is done by deriving expressions for the expectation steps (E-steps) and maximization steps (M-steps) for each of the parameters. Expressions for computing the variances for the parameter estimates are also derived. Simulation results comparing parameter estimates and standard errors obtained using the EM algorithm and the Quasi-Newton methods are presented.

The simulation results show that results obtained by the two methods are basically the same especially for the two group model. Parameter estimation, however, tends to become difficult as the model becomes complex such as increasing the number of groups from two to three. There were no problems in parameter estimation for a two group model, however, there were some challenges fitting a three group model especially using the Quasi-Newton method. Simulation results suggest that the results obtained using the Quasi-Newton method are highly dependent on starting values and hence the proposal of logistic starting values. The EM algorithm appears to be less sensitive to starting values.

Simulation results also suggest that the convergence rate of the EM algorithm is initially faster than that of the Quasi-Newton method as it requires a few EM-steps to get to the integer part of the maximized loglikelihood value and then requires a substantial number of EM steps to get the maximized loglikelihood value correct to a number of decimal places. As earlier stated under simulation results for the three group model, it may not matter a lot to get the maximized loglikelihood value correct to five decimal places, for example, as the parameter estimates do not reflect substantial changes. The simulation results also suggest that the Quasi-Newton method requires very few iterations to obtain the value of the maximized loglikelihood once the integer part of the maximized loglikelihood value is obtained.

The proposed parameter estimation approach is therefore

useful in the sense it is an alternative to the current method of parameter estimation. It can also be used with the current method to increase chances of getting the correct maximized loglikelihood value correct to a number of decimal places without a lot of iterations. The proposed approach may have problems with slow convergence rate beyond a certain value of the maximized loglikelihood especially for more complex models, however, it is less sensitive to starting values. As observed from the simulation results the current approach has a fast convergence rate but very sensitive to starting values. To solve the problems with the two approaches, the two approaches can be used together. Parameter estimation can begin with the proposed method since it is less sensitive to starting values and then the parameter estimates obtained can be used as starting values for the current method.

ACKNOWLEDGEMENTS

I would like to thank Prof K.S. Brown for his invaluable contribution to the work.

REFERENCES

- [1] Biernacki, C., Celeux, G. and Govaert, G. (2003). "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models." *Computational Statistics and Data Analysis*, 41, 561 – 575.
- [2] Blischke, W. R. (1962), "Moment estimators for the parameters of a mixture of two binomial distributions." *Annals of Mathematical Statistics*, 33, 444 – 454.
- [3] Blischke, W. R. (1964), "Estimating the parameters of mixtures of binomial distributions." *Journal of the American Statistical Association*, 59, 510 – 528.
- [4] Bremner, J. M. (1978). "Algorithm AS 123: Mixtures of Beta distributions." *Applied Statistics*, 27, 104 – 109.
- [5] Day, N. E. (1969). "Estimating the components of a mixture of normal distributions." *Biometrika*, 56, 468 – 474.
- [6] Fraley, C. and Raftery, A. E. (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation." *Journal of the American Statistical Association*, 97, 611–631.
- [7] Jones, B. L., Nagin, D. S. and Roeder, K. (2001). "A SAS procedure based mixture models for estimating developmental trajectories." *Sociological Methods and Research*, 29, 374 – 393.
- [8] Kao, J. H. K. (1959). "A graphical estimation of mixed Weibull parameters in life-testing electron tubes." *Technometrics*, 1, 389 – 407.
- [9] Louis, T. A. (1982). "Finding the Observed Information Matrix when using the EM Algorithm." *Journal of the Royal Statistical Society B*, 44, 226 – 233.
- [10] Marriot, F. H. C. (1975). "Separating mixtures of normal distributions." *Biometrics*, 31, 767 – 769.
- [11] McLachlan, G. J. and Basford, K. E. (1988). "Mixture Models: Inference and Applications to Clustering." New York: Marcel Dekker.
- [12] Nagin, D. S. (1999). "Analyzing developmental trajectories: A semiparametric group-based approach." *Psychological Methods*, 4, 139 – 157.
- [13] O'Hagan, A., Murphy, T. B. and Gormley, I. C. (2012). "Computational aspects of fitting mixture models via the expectation-maximization algorithm." *Computational Statistics and Data Analysis*, 56, 3843 – 3864.
- [14] Pearson, K. (1894). "Contributions to the theory of mathematical evolution." *Philosophical Transactions of the Royal Society of London A*, 185, 71 – 1196.
- [15] Rider, P. R. (1961). "The method of moments applied to a mixture of two exponential distributions." *Annals of Mathematical Statistics*, 32, 143 – 147.
- [16] Rider, P. R. (1962). "Estimating the parameters of mixed Poisson, binomial, and Weibull distributions by the method of moments." *Bulletin of the International Statistical Institute*, 39, 225 – 232.
- [17] Symons, M. J. (1981). "Clustering criteria and multivariate normal distributions." *Biometrics*, 37, 35 – 43.
- [18] Wolfe, J. H. (1970). "Pattern clustering of multivariate mixture analysis." *Multivariate Behavioral Research*, 5, 329 – 350.