American Journal of Mathematics and Statistics 2015, 5(5): 293-305
DOI: 10.5923/j.ajms.20150505.10

A Mixture Model for Longitudinal Trajectories with Covariates

Victor Mooto Nawa

Department of Mathematics and Statistics, University of Zambia, Lusaka, Zambia

Abstract An alternative method of estimating parameters in a mixture model for longitudinal trajectories with covariates using the expectation – maximization (EM) algorithm is proposed. Explicit expressions for the expectation and maximization steps required in the parameter estimation of group and covariate parameters are derived. Expressions for the variances of group and covariate parameters for the mixture model are also derived. Simulation results suggest that the proposed approach has good convergence properties especially when covariates are introduced in the model 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 in many different fields of study. Closely related to mixture models is the subject of cluster analysis which deals with the search of related observations in a data set. A general methodology for model-based clustering is given by Fraley and Raftery (2002). The application of finite mixture models to heterogeneous data is explained in details in McLachlan and Basford (1988) as well as McLachlan and Peel (2000). Most of the researchers who have studied finite mixture models have used the method of maximum likelihood estimation and the EM algorithm. The EM algorithm (Dempster et al. 1977; McLachlan and Krishnan, 2008) is a general approach for obtaining maximum likelihood estimates for problems in which data can be viewed as incomplete. The basic idea behind the EM algorithm is to frame a given incomplete-data problem into a complete-data problem for which maximum likelihood estimation is computationally tractable. The EM algorithm estimates the parameters of a model iteratively, starting from some initial guess. Each iteration consists of an expectation step (E-step) and a maximization step (M-step).

This paper deals with a mixture model for longitudinal trajectories. In particular the paper focusses on the semiparametric group based model proposed by Nagin (1999). This group based model assumes that the population consists of a mixture of distinct groups defined by different trajectory groups. Identifying the different trajectory groups that exist in the population is one of the primary objectives of the modelling strategy. The modelling strategy presumes that two types of variables have been measured: response variables and covariates. Out of the three different data types namely count, binary and psychometric scale data that this modelling approach can handle, we will only concentrate on binary data.

Roeder et al. (1999) considered the problem of estimating parameters for this model when response variables and covariates are in the model using the EM algorithm but restricted to count data only. Nawa (2014) considered the problem of estimating parameters in this model based on response variables only using the EM algorithm for longitudinal binary data. This article will consider parameter estimation for a model involving binary longitudinal data based on response variables and covariates using the EM algorithm.

Parameter estimates for this modelling approach can be obtained using a SAS procedure called PROC TRAJ written by Jones et al. (2001). This software is a customized SAS procedure that was developed with the SAS product SAS/TOOLKIT. In this SAS procedure, the parameters are obtained by maximum likelihood estimation using the Quasi-Newton method. The results obtained from this procedure are, however, highly dependent on the starting values used (Nawa, 2009). As such, this article presents an alternative approach using the EM algorithm.

The remainder of this paper is organised as follows. A brief discussion of mixture models in general and an application to the model under discussion is given in Section 2. This section begins with a discussion of the standard mixture model in Section 2.1. Thereafter a discussion of the longitudinal mixture model with covariates follows in Section 2.2. Section 2.2.1 gives the

* Corresponding author:
vnawa@yahoo.com (Victor Mooto Nawa)
Published online at http://journal.sapub.org/ajms

likelihood formulation of the longitudinal mixture model with covariates clearly outlining the E-steps and M-steps required to estimate the group and the covariate parameters in the model. This is followed by a discussion on estimation of standard errors for the group and covariate parameter values 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 mixture model with g groups (McLachlan and Peel, 2000) takes the form

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

where $\psi=(\theta_1,\theta_2,...,\theta_g,\pi_1,\pi_2,...,\pi_{g-1})$ is a vector of all unknown parameters with $f_i(y_j;\theta_i)$ representing the distribution of group C_i and $\pi_1,\pi_2,...,\pi_g$ are the unknown mixing proportions where $C_1,C_2,\ldots,C_g$ are the g groups in the mixture model. If $y=(y_1,y_2,...,y_n)^T$ is a random sample from the mixture model (1), then the likelihood function for ψ can be written in the form

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

The parameter vector ψ can be obtained by maximizing the likelihood in (2) or maximizing the log-likelihood given in (3) below.

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

Equivalently, the likelihood in (2) can be maximized using the EM (expectation - maximization) algorithm. This is done by maximizing the likelihood in (4) or the log-likelihood in (5) usually referred to as the complete-data log-likelihood

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

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

where z_{ij} is an indicator variable defined by

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

2.2. Longitudinal Mixture Model with Covariates

2.2.1. Likelihood Formulation and Parameter Estimation

One of the goals in longitudinal trajectory modelling is to study the effect of risk factors on longitudinal trajectories. The model under investigation has a provision for incorporating the effect of risk factors, time stable and time dependent covariates, on the longitudinal trajectories. Our focus in this paper will just be on the time stable covariates. It is assumed that risk factors affect the likelihood of a particular data trajectory, but nothing more can be learned about the response (Y) from the risk factors (X) given group membership (C) (Jones et al. 2001).

Time stable covariates are incorporated in the model by assuming that they influence the probability of belonging to a particular group. Given a sequence $y_j=(y_{j1},y_{j2},...,y_{jm})$ of longitudinal observations measured at *m* time points on subject *j* and a set of covariates $x_j=(1,x_{j1},x_{j2},...,x_{jp})$, the likelihood of the joint model based on a sample of *n* subjects and *g* groups takes the form

$$L(\psi)=\prod_{j=1}^{n}\sum_{i=1}^{g}\Pr(C_j=i\,|\,X_j=x_j)\Pr(Y_j=y_j\,|\,C_j=i) \tag{6}$$

where $\Pr(C_j=i\,|\,X_j=x_j)$ is the probability that the j^{th} subject belongs to group *i* given the set of covariates *x* and $\Pr(Y_j=y_j\,|\,C_j=i)$ is the probability distribution of the i^{th} group. A polychotomous logistic regression model is used to relate risk factors to group membership and therefore the probability that subject *j* belongs to group *i* given a vector of risk factors takes the form

$$\Pr(C_j=i\,|\,X_j=x_j)=\frac{\exp(\alpha_i^T x_j)}{\sum_{i=1}^{g}\exp(\alpha_i^T x_j)} \tag{7}$$

where $x_j=(1,x_{j1},x_{j2},...,x_{jp})$ is the j^{th} row of the *n* by(p+1) matrix of risk factors *x* and $\alpha_i=(\alpha_0^i,\alpha_1^i,...,\alpha_p^i)$ is a vector of length (p+1) consisting of covariate parameters associated with group *i*. Group one is taken as baseline with α_1 taken as zero and the log odds of membership in group *i* versus group one are given by

$$\log\left(\frac{\Pr(C_j=i\,|\,X_j=x_j)}{\Pr(C_j=1\,|\,X_j=x_j)}\right)=\alpha_i^T x_j=\alpha_0^i+\alpha_1^i x_{j1}+...+\alpha_p^i x_{jp}\,. \tag{8}$$

The likelihood in (6) can be equivalently written as

$$L(\psi)=\prod_{j=1}^{n}\sum_{i=1}^{g}\pi_i(x_j,\alpha_i)f_i(y_j,\beta_i) \tag{9}$$

where

$$\begin{aligned}\pi_i(x_j,\alpha_i) &= \frac{\exp(\alpha_i^T x_j)}{\sum_{i=1}^{g}\exp(\alpha_i^T x_j)}\\ &= \frac{\exp(\alpha_0^i+\alpha_1^i x_{j1}+...+\alpha_p^i x_{jp})}{1+\sum_{i=2}^{g}\exp(\alpha_0^i+\alpha_1^i x_{j1}+...+\alpha_p^i x_{jp})}\end{aligned}$$

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

$\beta_i=(\beta_0^i,\beta_1^i,\beta_2^i)$ is a vector of parameters which determines the shape of the trajectory for the i^{th} group and a_{jt} is the age of subject j at time t (Nagin, 1999). Since group one is taken as the baseline with $\alpha_1=0$, the set of all parameters ψ is given by $\psi=(\alpha_2^T,\alpha_3^T,...,\alpha_g^T,\beta_1^T,\beta_2^T,...,\beta_g^T)$.

The likelihood in (9) can be maximized directly using PROC TRAJ (Jones et al. 2001), however, this paper discusses an alternative way of maximization using the EM – algorithm. Instead of maximizing the likelihood in (9), the EM – algorithm maximizes the complete-data likelihood (or log-likelihood) obtained by introducing an indicator variable z_{ij} which takes the value one if the j^{th} observation belongs to group i and zero otherwise. The resulting complete-data likelihood takes the form

$$L_c(\psi)=\prod_{j=1}^{n}\prod_{i=1}^{g}\pi_i(x_j,\alpha_i)^{z_{ij}}\,f_i(y_j,\beta_i)^{z_{ij}} \tag{10}$$

while the complete-data log-likelihood is given by

$$l_c(\psi)=\sum_{j=1}^{n}\sum_{i=1}^{g}z_{ij}\log\pi_i(x_j,\alpha_i)+\sum_{j=1}^{n}\sum_{i=1}^{g}z_{ij}\log f_i(y_j,\beta_i). \tag{11}$$

Substituting for $\pi_i(x_j,\alpha_i)$ and $f_i(y_j,\beta_i)$, the complete-data log-likelihood (11) becomes

$$\begin{aligned}l_c(\psi) &= \sum_{j=1}^{n}\sum_{i=1}^{g}z_{ij}\left\{\alpha_i^T x_j-\log\left(1+\sum_{i=2}^{g}\exp(\alpha_i^T x_j)\right)\right\}+\\ &\quad \sum_{j=1}^{n}\sum_{i=1}^{g}z_{ij}\left\{\sum_{t=1}^{m}y_{jt}\left(\beta_0^i+\beta_1^i a_{jt}+\beta_2^i a_{jt}^2\right)-\log\left(1+\exp(\beta_0^i+\beta_1^i a_{jt}+\beta_2^i a_{jt}^2)\right)\right\}\\ &= \sum_{j=1}^{n}\left[z_{1j}\left\{-\log\left(1+\sum_{i=2}^{g}\exp(\alpha_i^T x_j)\right)\right\}+\sum_{i=2}^{g}z_{ij}\left\{\alpha_i^T x_j-\log\left(1+\sum_{i=2}^{g}\exp(\alpha_i^T x_j)\right)\right\}\right]+\\ &\quad \sum_{j=1}^{n}\sum_{i=1}^{g}z_{ij}\left\{\sum_{t=1}^{m}y_{jt}\left(\beta_0^i+\beta_1^i a_{jt}+\beta_2^i a_{jt}^2\right)-\log\left(1+\exp(\beta_0^i+\beta_1^i a_{jt}+\beta_2^i a_{jt}^2)\right)\right\}\end{aligned}$$

The E-step (expectation step) on the (k+1)th iteration involves evaluating $E\left(l_c(\psi \mid y;\psi^{(k)})\right)$, where $\psi^{(k)} = (\alpha_2^{(k)T}, \alpha_3^{(k)T}, ..., \alpha_g^{(k)T}, \beta_1^{(k)T}, \beta_2^{(k)T}, ..., \beta_g^{(k)T})$ are parameter estimates obtained on the kth iteration. The only random component in the complete-data log-likelihood is Z_{ij} whose expectation is given by

$$E\left(Z_{ij} \mid y_j;\psi^{(k)}\right) = \frac{\pi_i^{(k)}(x_j, \alpha_i^{(k)}) f_i(y_j, \beta_i^{(k)})}{\sum_{i=1}^{g} \pi_i^{(k)}(x_j, \alpha_i^{(k)}) f_i(y_j, \beta_i^{(k)})} = z_{ij}^{(k)} \tag{12}$$

Similarly, the M-step on the (k+1)th iteration constitutes finding a value of the parameter vector ψ that maximizes the expected log-likelihood, thus

$$\psi^{(k+1)} = \arg\max_{\psi} E\left(l_c(\psi \mid y;\psi^{(k)})\right) \tag{13}$$

This expected complete-data log-likelihood consists of two sums which can be maximized separately – the first component only depends on the parameter vector $\alpha = (\alpha_2^T, \alpha_3^T, ..., \alpha_g^T)$ while the second component only depends on the parameter vector $(\beta_1^T, \beta_2^T, ..., \beta_g^T)$. In fact, the second component of the expected complete-data log-likelihood can further be separately maximized to obtain $\beta_i^{(k+1)}$ for each group for $i = 1, 2, ..., g$. Thus we have

$$\alpha^{(k+1)} = \arg\max_{\alpha} \sum_{j=1}^{n} \sum_{i=1}^{g} z_{ij}^{(k)} \left\{ \alpha_i^{(k)T} x_j - \log\left(1 + \sum_{i=2}^{g} \exp(\alpha_i^{(k)T} x_j)\right)\right\} \tag{14}$$

and

$$\beta_i^{(k+1)} = \arg\max_{\beta_i} \sum_{j=1}^{n} z_{ij}^{(k)} \left\{ \sum_{t=1}^{m} y_{jt}(A_{jt}^T \beta_i^{(k)}) - \log\left(1 + \exp(A_{jt}^T \beta_i^{(k)})\right)\right\} \tag{15}$$

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 $i = 1, 2, ..., g$. 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 can be obtained from the inverse of the observed matrix. The procedure developed by Louis (1982) for extracting the observed information matrix from the complete data log-likelihood when the EM-algorithm is used to find maximum likelihood estimates is used.

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

$$I(\hat{\psi}; y) = J_c(\hat{\psi}; y) - J_m(\hat{\psi}; y) \tag{16}$$

where

$$J_c(\psi; y) = E\left[I_c(\psi \mid y)\right] \tag{17}$$

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

$$J_m(\psi; y) = \mathrm{cov}\left[S_c(\psi \mid y)\right]. \tag{18}$$

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

$$S_c(\psi) = \left(S_c(\alpha_2)^T, S_c(\alpha_3)^T, \ldots, S_c(\alpha_g)^T, S_c(\beta_1)^T, S_c(\beta_2)^T, \ldots, S_c(\beta_g)^T \right), \tag{19}$$

where

$$S_c(\alpha_i)^T = \left(\frac{\partial l}{\partial \alpha_0^i}, \frac{\partial l}{\partial \alpha_1^i}, \ldots, \frac{\partial l}{\partial \alpha_p^i} \right) \tag{20}$$

and

$$S_c(\beta_i)^T = \left(\frac{\partial l}{\partial \beta_0^i}, \frac{\partial l}{\partial \beta_1^i}, \frac{\partial l}{\partial \beta_2^i} \right) \tag{21}$$

An expression for $S_c(\beta_i)$ is given in Nawa (2014), thus we only need to find $S_c(\alpha_i)$. Differentiating the complete-data log-likelihood (11) with respect to α_i gives

$$\frac{\partial l}{\partial \alpha_i} = \left(\frac{\partial l}{\partial \alpha_0^i}, \frac{\partial l}{\partial \alpha_1^i}, \ldots, \frac{\partial l}{\partial \alpha_p^i} \right)^T \tag{22}$$

where,

$$\frac{\partial l}{\partial \alpha_0^i} = \sum_{j=1}^{n} \left(z_{ij} - \frac{\exp(\alpha_i^T x_j)}{1 + \sum_{i=2}^{g} \alpha_i^T x_j} \right)$$

$$\frac{\partial l}{\partial \alpha_1^i} = \sum_{j=1}^{n} x_{j1} \left(z_{ij} - \frac{\exp(\alpha_i^T x_j)}{1 + \sum_{i=2}^{g} \alpha_i^T x_j} \right)$$

$$\frac{\partial l}{\partial \alpha_2^i} = \sum_{j=1}^{n} x_{j2} \left(z_{ij} - \frac{\exp(\alpha_i^T x_j)}{1 + \sum_{i=2}^{g} \alpha_i^T x_j} \right)$$

$$\vdots$$

$$\frac{\partial l}{\partial \alpha_p^i} = \sum_{j=1}^{n} x_{jp} \left(z_{ij} - \frac{\exp(\alpha_i^T x_j)}{1 + \sum_{i=2}^{g} \alpha_i^T x_j} \right)$$

for $i = 2, 3, \ldots, g$. Thus,

$$\frac{\partial l}{\partial \alpha_k^i} = \sum_{j=1}^{n} x_{jk} \left(z_{ij} - \frac{\exp(\alpha_i^T x_j)}{1 + \sum_{i=2}^{g} \alpha_i^T x_j} \right),$$

for $i = 2, 3, \ldots, g$ and $k = 0, 1, 2, \ldots, p$ $(i.e.\ x_{j0} = 1)$.

The information matrix based on the complete-data log-likelihood (11) can be written as a block-diagonal matrix

$$I_c(\psi) = \begin{pmatrix} I_c(\alpha) & 0 & 0 & \ldots & 0 \\ 0 & I_c(\beta_1) & 0 & \ldots & 0 \\ 0 & 0 & I_c(\beta_2) & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & I_c(\beta_g) \end{pmatrix} \tag{23}$$

An expression for $I_c(\beta_i)$ is given in Nawa (2014), thus we only need to find $I_c(\alpha)$. The matrix $I_c(\alpha)$ can be written as

$$I_c(\alpha) = \begin{pmatrix} I_c(\alpha_2) & I_c(\alpha_2, \alpha_3) & I_c(\alpha_2, \alpha_4) & \cdots & I_c(\alpha_2, \alpha_g) \\ I_c(\alpha_3, \alpha_2) & I_c(\alpha_3) & I_c(\alpha_3, \alpha_4) & \cdots & I_c(\alpha_3, \alpha_g) \\ I_c(\alpha_4, \alpha_2) & I_c(\alpha_4, \alpha_3) & I_c(\alpha_4) & \cdots & I_c(\alpha_4, \alpha_g) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ I_c(\alpha_g, \alpha_2) & I_c(\alpha_g, \alpha_3) & I_c(\alpha_g, \alpha_4) & \cdots & I_c(\alpha_g) \end{pmatrix} \tag{24}$$

where $I_c(\alpha_i)$ and $I_c(\alpha_i, \alpha_k)$ (where $i \neq k$) are $(p+1)$ by $(p+1)$ matrices respectively given by

$$I_c(\alpha_i) = \begin{pmatrix} -\frac{\partial^2 l}{\partial \alpha_0^{i\,2}} & -\frac{\partial^2 l}{\partial \alpha_0^i \partial \alpha_1^i} & -\frac{\partial^2 l}{\partial \alpha_0^i \partial \alpha_2^i} & \cdots & -\frac{\partial^2 l}{\partial \alpha_0^i \partial \alpha_p^i} \\ -\frac{\partial^2 l}{\partial \alpha_1^i \partial \alpha_0^i} & -\frac{\partial^2 l}{\partial \alpha_1^{i\,2}} & -\frac{\partial^2 l}{\partial \alpha_1^i \partial \alpha_2^i} & \cdots & -\frac{\partial^2 l}{\partial \alpha_1^i \partial \alpha_p^i} \\ -\frac{\partial^2 l}{\partial \alpha_2^i \partial \alpha_0^i} & -\frac{\partial^2 l}{\partial \alpha_2^i \partial \alpha_1^i} & -\frac{\partial^2 l}{\partial \alpha_2^{i\,2}} & \cdots & -\frac{\partial^2 l}{\partial \alpha_2^i \partial \alpha_p^i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial^2 l}{\partial \alpha_p^i \partial \alpha_0^i} & -\frac{\partial^2 l}{\partial \alpha_p^i \partial \alpha_1^i} & -\frac{\partial^2 l}{\partial \alpha_p^i \partial \alpha_2^i} & \cdots & -\frac{\partial^2 l}{\partial \alpha_p^{i\,2}} \end{pmatrix} \tag{25}$$

and

$$I_c(\alpha_i, \alpha_k) = \begin{pmatrix} -\frac{\partial^2 l}{\partial \alpha_0^i \partial \alpha_0^k} & -\frac{\partial^2 l}{\partial \alpha_0^i \partial \alpha_1^k} & -\frac{\partial^2 l}{\partial \alpha_0^i \partial \alpha_2^k} & \cdots & -\frac{\partial^2 l}{\partial \alpha_0^i \partial \alpha_p^k} \\ -\frac{\partial^2 l}{\partial \alpha_0^i \partial \alpha_1^k} & -\frac{\partial^2 l}{\partial \alpha_1^i \partial \alpha_1^k} & -\frac{\partial^2 l}{\partial \alpha_1^i \partial \alpha_2^k} & \cdots & -\frac{\partial^2 l}{\partial \alpha_1^i \partial \alpha_p^k} \\ -\frac{\partial^2 l}{\partial \alpha_0^i \partial \alpha_2^k} & -\frac{\partial^2 l}{\partial \alpha_1^i \partial \alpha_2^k} & -\frac{\partial^2 l}{\partial \alpha_2^i \partial \alpha_2^k} & \cdots & -\frac{\partial^2 l}{\partial \alpha_2^i \partial \alpha_p^k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial^2 l}{\partial \alpha_0^i \partial \alpha_p^k} & -\frac{\partial^2 l}{\partial \alpha_1^i \partial \alpha_p^k} & -\frac{\partial^2 l}{\partial \alpha_2^i \partial \alpha_p^k} & \cdots & -\frac{\partial^2 l}{\partial \alpha_p^i \partial \alpha_p^k} \end{pmatrix}. \tag{26}$$

The $(kl)^{th}$ element of the matrix $I_c(\alpha_i)$ is given by

$$-\frac{\partial^2 l}{\partial\alpha_k^i\partial\alpha_l^i}=\sum_{j=1}^{n}\left(\frac{x_{jk}x_{jl}\exp(\alpha_i^T x_j)\left(1-\exp(\alpha_i^T x_j)+\sum_{i=2}^{g}\exp(\alpha_i^T x_j)\right)}{\left(1+\sum_{i=2}^{g}\exp(\alpha_i^T x_j)\right)^2}\right),\quad for\ k\neq l \tag{27}$$

and

$$-\frac{\partial^2 l}{\partial\alpha_k^{i\,2}}=\sum_{j=1}^{n}\left(\frac{x_{jk}^2\exp(\alpha_i^T x_j)\left(1-\exp(\alpha_i^T x_j)+\sum_{i=2}^{g}\exp(\alpha_i^T x_j)\right)}{\left(1+\sum_{i=2}^{g}\exp(\alpha_i^T x_j)\right)^2}\right),\quad for\ k=l \tag{28}$$

for $k,l=0,1,2,\ldots,p$. Similarly, the $(fh)^{th}$ element of the matrix $I_c(\alpha_i,\alpha_k)$ is given by

$$-\frac{\partial^2 l}{\partial\alpha_f^i\partial\alpha_h^k}=-\sum_{j=1}^{n}\left(\frac{x_{jf}x_{jh}\exp(\alpha_i^T x_j)\exp(\alpha_k^T x_j)}{\left(1+\sum_{i=2}^{g}\exp(\alpha_i^T x_j)\right)^2}\right),\quad for\ f\neq h \tag{29}$$

and

$$-\frac{\partial^2 l}{\partial\alpha_h^i\partial\alpha_h^k}=-\sum_{j=1}^{n}\left(\frac{x_{jh}^2\exp(\alpha_i^T x_j)\exp(\alpha_k^T x_j)}{\left(1+\sum_{i=2}^{g}\exp(\alpha_i^T x_j)\right)^2}\right),\quad for\ f=h \tag{30}$$

for $f,h=0,1,2,\ldots,p$.

We now find $\mathrm{cov}\left[S_c(\psi\mid y)\right]$, the conditional covariance of the score vector (19). This covariance matrix can be written as

$$J_m(\psi)=\begin{pmatrix}\mathrm{cov}\left(S_c(\alpha)\right) & \mathrm{cov}\left(S_c(\alpha),S_c(\beta_1)\right) & \mathrm{cov}\left(S_c(\alpha),S_c(\beta_2)\right) & \cdots & \mathrm{cov}\left(S_c(\alpha),S_c(\beta_g)\right)\\ \mathrm{cov}\left(S_c(\beta_1),S_c(\alpha)\right) & \mathrm{cov}\left(S_c(\beta_1)\right) & \mathrm{cov}\left(S_c(\beta_1),S_c(\beta_2)\right) & \cdots & \mathrm{cov}\left(S_c(\beta_1),S_c(\beta_g)\right)\\ \mathrm{cov}\left(S_c(\beta_2),S_c(\alpha)\right) & \mathrm{cov}\left(S_c(\beta_2),S_c(\beta_1)\right) & \mathrm{cov}\left(S_c(\beta_2)\right) & \cdots & \mathrm{cov}\left(S_c(\beta_2),S_c(\beta_g)\right)\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathrm{cov}\left(S_c(\beta_g),S_c(\alpha)\right) & \mathrm{cov}\left(S_c(\beta_g),S_c(\beta_1)\right) & \mathrm{cov}\left(S_c(\beta_g),S_c(\beta_2)\right) & \cdots & \mathrm{cov}\left(S_c(\beta_g)\right)\end{pmatrix} \tag{31}$$

Expressions for $\mathrm{cov}\left(S_c(\beta_i)\right)$ (for $i=1,2,\ldots,g$) and $\mathrm{cov}\left(S_c(\beta_i),S_c(\beta_k)\right)$ (for $i\neq k$) can be found in Nawa (2014). Therefore we only need to find the other components of the matrix. The covariance matrix $\mathrm{cov}\left(S_c(\alpha)\right)$ can be written as

$$\operatorname{cov}\left(S_c(\alpha)\right)=\begin{pmatrix} \operatorname{cov}\left(S_c(\alpha_2)\right) & \operatorname{cov}\left(S_c(\alpha_2),S_c(\alpha_3)\right) & \cdots & \operatorname{cov}\left(S_c(\alpha_2),S_c(\alpha_g)\right) \\ \operatorname{cov}\left(S_c(\alpha_3),S_c(\alpha_2)\right) & \operatorname{cov}\left(S_c(\alpha_3)\right) & \cdots & \operatorname{cov}\left(S_c(\alpha_3),S_c(\alpha_g)\right) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}\left(S_c(\alpha_g),S_c(\alpha_2)\right) & \operatorname{cov}\left(S_c(\alpha_g),S_c(\alpha_3)\right) & \cdots & \operatorname{cov}\left(S_c(\alpha_g)\right) \end{pmatrix} \tag{32}$$

while $\operatorname{cov}\left(S_c(\alpha),S_c(\beta_i)\right)$, for $i=1,2,\ldots,g$, can be written as

$$\operatorname{cov}\left(S_c(\alpha),S_c(\beta_i)\right)=\begin{pmatrix} \operatorname{cov}\left(S_c(\alpha_2),S_c(\beta_i)\right) \\ \operatorname{cov}\left(S_c(\alpha_3),S_c(\beta_i)\right) \\ \vdots \\ \operatorname{cov}\left(S_c(\alpha_g),S_c(\beta_i)\right) \end{pmatrix}. \tag{33}$$

Let

$$E\left(Z_{ij}\mid y_j;\psi\right)=\frac{\pi_i(x_j,\alpha_i)f_i(y_j,\beta_i)}{\sum_{i=1}^{g}\pi_i(x_j,\alpha_i)f_i(y_j,\beta_i)}=\tau_{ij} \tag{34}$$

then

$$Var(Z_{ij})=\tau_{ij}(1-\tau_{ij})=v_{ij} \tag{35}$$

and

$$Cov(Z_{ij},Z_{kj})=-\tau_{ij}\tau_{kj}=\rho_{ijk}\quad,\ i\neq k \tag{36}$$

Using (34), (35) and (36), we can show that the (p+1) by (p+1) dimensional matrices $\operatorname{cov}\left(S_c(\alpha_i)\right)$ and $\operatorname{cov}\left(S_c(\alpha_i),S_c(\alpha_k)\right)$ are respectively given by

$$\operatorname{cov}\left(S_c(\alpha_i)\right)=\begin{pmatrix} \sum_{j=1}^{n}v_{ij} & \sum_{j=1}^{n}x_{j1}v_{ij} & \sum_{j=1}^{n}x_{j2}v_{ij} & \cdots & \sum_{j=1}^{n}x_{jp}v_{ij} \\ \sum_{j=1}^{n}x_{j1}v_{ij} & \sum_{j=1}^{n}x_{j1}^2v_{ij} & \sum_{j=1}^{n}x_{j1}x_{j2}v_{ij} & \cdots & \sum_{j=1}^{n}x_{j1}x_{jp}v_{ij} \\ \sum_{j=1}^{n}x_{j2}v_{ij} & \sum_{j=1}^{n}x_{j2}x_{j1}v_{ij} & \sum_{j=1}^{n}x_{j2}^2v_{ij} & \cdots & \sum_{j=1}^{n}x_{j2}x_{jp}v_{ij} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{n}x_{jp}v_{ij} & \sum_{j=1}^{n}x_{jp}x_{j1}v_{ij} & \sum_{j=1}^{n}x_{jp}x_{j2}v_{ij} & \cdots & \sum_{j=1}^{n}x_{jp}^2v_{ij} \end{pmatrix} \tag{37}$$

and

$$\operatorname{cov}\left(S_c(\alpha_i), S_c(\alpha_k)\right) = \begin{pmatrix} \sum_{j=1}^{n} \rho_{ijk} & \sum_{j=1}^{n} x_{j1}\rho_{ijk} & \sum_{j=1}^{n} x_{j2}\rho_{ijk} & \cdots & \sum_{j=1}^{n} x_{jp}\rho_{ijk} \\ \sum_{j=1}^{n} x_{j1}\rho_{ijk} & \sum_{j=1}^{n} x_{j1}^2\rho_{ijk} & \sum_{j=1}^{n} x_{j1}x_{j2}\rho_{ijk} & \cdots & \sum_{j=1}^{n} x_{j1}x_{jp}\rho_{ijk} \\ \sum_{j=1}^{n} x_{j2}\rho_{ijk} & \sum_{j=1}^{n} x_{j2}x_{j1}\rho_{ijk} & \sum_{j=1}^{n} x_{j2}^2\rho_{ijk} & \cdots & \sum_{j=1}^{n} x_{j2}x_{jp}\rho_{ijk} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^{n} x_{jp}\rho_{ijk} & \sum_{j=1}^{n} x_{jp}x_{j1}\rho_{ijk} & \sum_{j=1}^{n} x_{jp}x_{j2}\rho_{ijk} & \cdots & \sum_{j=1}^{n} x_{jp}^2\rho_{ijk} \end{pmatrix} \tag{38}$$

for $i = 2, 3, \ldots, g$ and $i \neq k$. Similarly, we can show that the (p+1) by 3 dimensional covariance matrix $\operatorname{cov}\left(S_c(\alpha_i), S_c(\beta_k)\right)$ is given by

$$\operatorname{cov}\left(S_c(\alpha_i), S_c(\beta_k)\right) = \begin{pmatrix} \sum_{j=1}^{n} A_{0j}^{i} v_{ij} & \sum_{j=1}^{n} A_{1j}^{i} v_{ij} & \sum_{j=1}^{n} A_{2j}^{i} v_{ij} \\ \sum_{j=1}^{n} A_{0j}^{i} x_{j1} v_{ij} & \sum_{j=1}^{n} A_{1j}^{i} x_{j1} v_{ij} & \sum_{j=1}^{n} A_{2j}^{i} x_{j1} v_{ij} \\ \sum_{j=1}^{n} A_{0j}^{i} x_{j2} v_{ij} & \sum_{j=1}^{n} A_{1j}^{i} x_{j2} v_{ij} & \sum_{j=1}^{n} A_{1j}^{i} x_{j2} v_{ij} \\ \vdots & \vdots & \vdots \\ \sum_{j=1}^{n} A_{0j}^{i} x_{jp} v_{ij} & \sum_{j=1}^{n} A_{1j}^{i} x_{jp} v_{ij} & \sum_{j=1}^{n} A_{2j}^{i} x_{jp} v_{ij} \end{pmatrix}, \quad for \ \ i = k \tag{39}$$

and

$$\operatorname{cov}\left(S_c(\alpha_i), S_c(\beta_k)\right) = \begin{pmatrix} \sum_{j=1}^{n} A_{0j}^{k} \rho_{ijk} & \sum_{j=1}^{n} A_{1j}^{k} \rho_{ijk} & \sum_{j=1}^{n} A_{2j}^{k} \rho_{ijk} \\ \sum_{j=1}^{n} A_{0j}^{k} x_{j1} \rho_{ijk} & \sum_{j=1}^{n} A_{1j}^{k} x_{j1} \rho_{ijk} & \sum_{j=1}^{n} A_{2j}^{k} x_{j1} \rho_{ijk} \\ \sum_{j=1}^{n} A_{0j}^{k} x_{j2} \rho_{ijk} & \sum_{j=1}^{n} A_{1j}^{k} x_{j2} \rho_{ijk} & \sum_{j=1}^{n} A_{2j}^{k} x_{j2} \rho_{ijk} \\ \vdots & \vdots & \vdots \\ \sum_{j=1}^{n} A_{0j}^{k} x_{jp} \rho_{ijk} & \sum_{j=1}^{n} A_{1j}^{k} x_{jp} \rho_{ijk} & \sum_{j=1}^{n} A_{2j}^{k} x_{jp} \rho_{ijk} \end{pmatrix}, \quad for \ \ i \neq k \tag{40}$$

for $i = 2, 3, \ldots, g$ and $k = 1, 2, \ldots, g$, where

$$A_{0j}^{i} = \sum_{t=1}^{m}\left(y_{jt} - \frac{\exp(A_{jt}^{T}\beta_i)}{1+\exp(A_{jt}^{T})} \right)$$

$$A_{1j}^{i} = \sum_{t=1}^{m}\left(y_{jt}a_{jt} - \frac{a_{jt}\exp(A_{jt}^{T}\beta_i)}{1+\exp(A_{jt}^{T}\beta_i)} \right)$$

$$A_{2j}^{i} = \sum_{t=1}^{m}\left(y_{jt}a_{jt}^{2} - \frac{a_{jt}^{2}\exp(A_{jt}^{T}\beta_i)}{1+\exp(A_{jt}^{T}\beta_i)} \right)$$

Putting all the above results together, we can find the expectation of the information matrix based on the complete-data log-likelihood given by $J_c(\psi;y) = E\left[I_c(\psi \mid y)\right]$ by finding the expectation of each component of the matrix $I_c(\psi \mid y)$. The matrix $I_c(\alpha)$ does not involve Z_{ij} and is therefore constant with respect to the expectation. The variance estimate for $\hat{\psi}$ is then obtained from the inverse of the matrix $J_c(\hat{\psi};y) - J_m(\hat{\psi};y)$ where the τ_{ij} 's in (34) are replaced by the estimated posterior probabilities $\hat{\tau}_{ij}$.

3. Simulation Results

We consider simulation results for mixtures of two and three groups of longitudinal trajectories. For a two group model we consider group parameters $\beta_1 = (6.170, -5.781, 0.997)$, $\beta_2 = (-7.690, 6.592, -1.099)$ and a covariate parameter $\alpha_2 = (2, -3)$. For a three group model we consider group parameters $\beta_1 = (6.170, -5.781, 0.997)$, $\beta_2 = (-7.690, 6.592, -1.099)$, $\beta_3 = (-2.237, -0.172, 0.212)$ and covariate parameters $\alpha_2 = (-5.4, -3.5)$ and $\alpha_3 = (-5.8, 3.2)$. Consider a situation involving five time points where measurements are taken at times 1 to 5 (i.e. $a_{j1} = 1$, $a_{j2} = 2$, $a_{j3} = 3$, $a_{j4} = 4$, $a_{j5} = 5$).

The group parameter values β_1, β_2 and β_3 are calculated based on the trajectory shapes given in Figure 1. Using these group parameter values, a sequence of binary responses are independently generated for the time points 1 to 5 with the response from the j[th] subject of group i at time t being generated with success probability

$$p = \frac{\exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)}{1+\exp(\beta_0^i + \beta_1^i a_{jt} + \beta_2^i a_{jt}^2)}. \qquad (41)$$

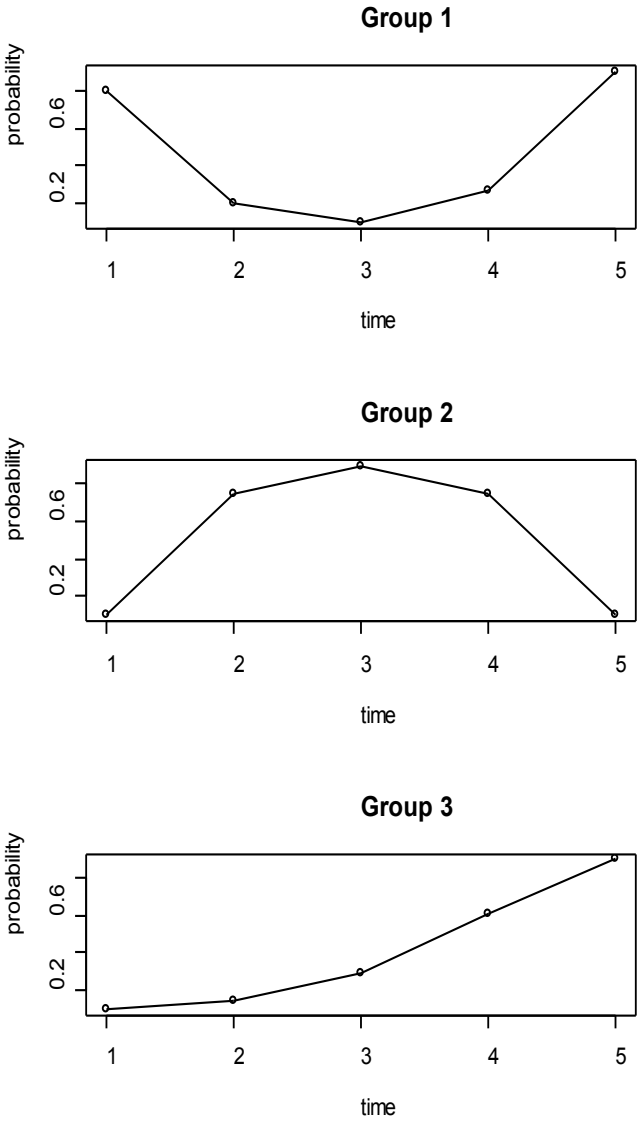

Figure 1. Longitudinal trajectories for three groups

In a two group model the covariate parameter is obtained by writing the probability of membership in group 2 as function of *x* as

$$\Pr(Y=1 \mid x) = \frac{\exp(\alpha_0^2 + \alpha_1^2 x)}{1+\exp(\alpha_0^2 + \alpha_1^2 x)} \qquad (42)$$

where $Y=1$ for observations in group 2 and $Y=0$ for observations in group 1. The values of *x* and $\alpha_2 = (\alpha_0^2, \alpha_1^2)$ are chosen based on a particular value of (42), which is the probability of a subject belonging to group 2. If $\Pr(Y=1 \mid x) = p_2$, then *x* and $\alpha_2 = (\alpha_0^2, \alpha_1^2)$ are chosen in such a way that the proportion of ones generated from (42) is equal to p_2. Observations for a three group model can be generated in a similar way. Taking group 1 as the baseline $(Y=0)$, we need the probability of belonging

to group 2 and the probability of belonging to group 3 as functions of x. These probabilities are respectively given by

$$\Pr(Y=1\mid x)=\frac{\exp(\alpha_0^2+\alpha_1^2 x)}{1+\exp(\alpha_0^2+\alpha_1^2 x)+\exp(\alpha_0^3+\alpha_1^3 x)} \quad (43)$$

and

$$\Pr(Y=2\mid x)=\frac{\exp(\alpha_0^3+\alpha_1^3 x)}{1+\exp(\alpha_0^2+\alpha_1^2 x)+\exp(\alpha_0^3+\alpha_1^3 x)} \quad (44)$$

Taking $\Pr(Y=1\mid x)=p_2$ and $\Pr(Y=2\mid x)=p_3$, we choose x, $\alpha_2=(\alpha_0^2,\alpha_1^2)$ and $\alpha_3=(\alpha_0^3,\alpha_1^3)$ such that (43) and (44) hold.

3.1. Mixture of Two Trajectory Groups

Consider 150 simulated data sets from a mixture of two trajectory groups with group parameters $\beta_1=(6.170,-5.781,0.997)$, $\beta_2=(-7.690,6.592,-1.099)$ and a covariate parameter $\alpha_2=(2,-3)$. For all the 150 simulations, the two methods converged to the same log-likelihood value correct to five decimal places. The mean number of EM steps until convergence was 16.92 while the mean number of iterations for PROC TRAJ was 38.95.

Table 1. Group parameter estimates, covariate parameter estimates and standard errors based on 150 simulations

	Group							
	1			2			2	
Parameter	β_0	β_1	β_2	β_0	β_1	β_2	α_0	α_1
Theoretical	6.170	-5.781	0.997	-7.690	6.592	-1.099	2	-3
Estimate (EM)	6.254	-5.850	1.008	-7.717	6.613	-1.102	2.044	-3.073
Estimate (PROC TRAJ)	6.256	-5.851	1.008	-7.717	6.613	-1.102	2.044	-3.073
SE (EM)	0.496	0.406	0.070	0.395	0.298	0.049	0.330	0.346
Empirical SE (EM)	0.490	0.393	0.066	0.396	0.300	0.049	0.337	0.367
SE (PROC TRAJ)	0.496	0.406	0.070	0.395	0.298	0.049	0.330	0.346
Empirical SE (PROC TRAJ)	0.490	0.393	0.066	0.396	0.300	0.049	0.337	0.367

Table 2. Group parameter estimates, covariate parameter estimates and standard errors based on 200 simulations

	Group											
	1			2			3			2		3	
Parameter	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	α_0	α_1	α_0	α_1
Theoretical	6.170	-5.781	0.997	-7.690	6.592	-1.099	-2.237	-0.172	0.212	-5.4	-3.5	-5.8	3.2
Estimate (EM)	6.220	-5.816	1.003	-7.708	6.609	-1.101	-2.228	-0.188	0.215	-5.669	-3.670	-6.066	3.348
Estimate (PROC TRAJ)	6.220	-5.816	1.003	-7.709	6.609	-1.101	-2.225	-0.189	0.215	-5.675	-3.673	-6.063	3.346
SE (EM)	0.453	0.368	0.063	0.391	0.299	0.050	0.414	0.302	0.051	0.789	0.507	0.871	0.464
Empirical SE (EM)	0.486	0.390	0.066	0.372	0.294	0.050	0.427	0.313	0.051	0.763	0.483	0.875	0.479
SE (PROC TRAJ)	0.453	0.368	0.063	0.391	0.299	0.050	0.414	0.302	0.051	0.790	0.508	0.870	0.463
Empirical SE (PROC TRAJ)	0.487	0.391	0.067	0.373	0.294	0.050	0.426	0.313	0.051	0.762	0.481	0.877	0.480

Table 1 shows theoretical group parameter values, theoretical covariate parameter values, group parameter estimates, covariate 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. Group and covariate parameter estimates obtained from the two methods are very similar and close to the theoretical values. Standard error estimates of the group and covariate parameter estimates obtained from the two methods are also very similar and close to the empirical standard errors. This suggests that the standard error estimates given by the two methods of estimation are a true representation of the variability in the parameter estimates.

3.2. Mixture of Three Trajectory Groups

Consider 200 simulations from a three group model with group parameters $\beta_1 = (6.170, -5.781, 0.997)$, $\beta_2 = (-7.690, 6.592, -1.099)$, $\beta_3 = (-2.237, -0.172, 0.212)$ and covariate parameters $\alpha_2 = (-5.4, -3.5)$ and $\alpha_3 = (-5.8, 3.2)$. For all the 200 simulations, the two methods converged to the same log-likelihood value correct to five decimal places. The mean number of EM steps until convergence was 25.11 while the mean number of iterations for PROC TRAJ was 62.16.

Table 2 gives the theoretical parameter values, group and covariate parameter estimates obtained from the two methods along with the corresponding standard error and empirical standard error estimates. The table shows that the parameter estimates given by the two methods are almost identical and also very close to the theoretical values. This is true for both the group parameter and covariate parameter estimates. The empirical standard errors are also very close to the estimated standard errors.

4. Conclusions

This paper is an extension of the work in Nawa (2014), which considered an application of the EM algorithm in estimating group parameters and mixing proportions and the corresponding standard errors in a mixture model of longitudinal trajectories. The paper also compared the results obtained from the EM algorithm to those obtained from the Quasi-Newton method through a SAS procedure called PROC TRAJ proposed by Jones et al. (2001). The extension looks at how the two methods of estimation compare when covariates are introduced in the model. When covariates are introduced in the model, the parameters to be estimated are group parameters and covariates.

The paper describes how estimation of various parameters is done using the EM algorithm. Explicit expressions for the expectation steps (E-steps) and maximization steps (M-steps) for each of the parameters are derived. Expressions for computing variances for the parameter estimates are also derived. Most of the expressions are a direct extension of the expressions for the model without covariates.

Simulations results indicate that the group parameter estimates, covariate parameter estimates and standard error estimates obtained from the two methods are practically the same. Compared to the model without covariates, parameter estimation in the model with covariates appears to have fewer challenges probably because we have more information separating out the different groups. The simulation results also show that parameter estimates from the model with covariates are also closer to theoretical values compared to the model without covariates. While the EM algorithm seems to have some convergence challenges, especially as the number of groups in the model increases, which is indicated by the number of additional EM steps until convergence to five or more decimal places, this is not the case in the model with covariates. In fact, the number of EM steps until convergence reduces significantly in the model with covariates.

The proposed application of the EM algorithm to a model with covariates offers a good alternative to the current method used in the parameter estimation. This is based on the fact that it is known that the current method of estimation is very sensitive to starting values while the EM algorithm is not and from the simulation results that suggest that the convergence seems to improve significantly when the proposed approach is used on a model with covariates.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] 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.

[2] Fraley, C. and Raftery, A. E. (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation." Journal of the American Statistical Association, 97, 611 – 631.

[3] 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.

[4] Louis, T. A. (1982). "Finding the Observed Information Matrix when using the EM Algorithm." Journal of the Royal Statistical Society B, 44, 226 – 233.

[5] McLachlan, G. J. and Basford, K. E. (1988). "Mixture Models: Inference and Applications to Clustering." New York: Marcel Dekker.

[6] McLachlan, G. J. and Krishnan, T. (2008). "The EM Algorithms and Extensions, Second Edition." New York: Wiley.

[7] McLachlan, G. J. and Peel, D. (2000). "Finite Mixture Models." New York: Wiley.

[8] Nagin, D. S. (1999). "Analyzing Developmental Trajectories: A Semiparametric Group-Based Approach." Psychological Methods, 4, 139 – 157.

[9] Nawa, V. M. (2009). "Comparison of the EM algorithm and the Quasi-Newton Method: An Application to Developmental Trajectories." University of Zambia Journal of Science and Technology, 13 (2), 41 – 54.

[10] Nawa, V. M. (2014). "A Mixture Model for Longitudinal Trajectories". International Journal of Statistics and Applications, 4 (4), 181 – 191.

[11] Roeder, K, Lynch, K. G. and Nagin, D. S. (1999). "Modelling Uncertainty in Latent Class Membership: A Case Study in Criminology." Journal of the American Statistical Association, 94, 766 – 776.