Estimation in the Mixture of Markov Chains Moving with Different Speeds

Halina Frydman
New York University

January 9, 2003

Abstract

This paper considers a new mixture of time homogeneous finite Markov chains where the mixing is on the rate of movement and develops the EM algorithm for the maximum likelihood estimation of the parameters of the mixture. A continuous and discrete time versions of the mixture are defined and their estimation is considered separately. The simulation study is carried out for the continuous time mixture. To simplify the exposition the results are derived for a mixture of two Markov chains, but can be easily extended to a mixture of any finite number of Markov chains. The class of mixture models proposed in this paper provides a framework for modeling population heterogeneity with respect to the rate of movement. The proposed mixture generalizes the mover-stayer model, which has been widely employed in applications.

KEY WORDS: Mixtures of Markov Chains; Mover-stayer model; EM algorithm.

1 Introduction

In this paper we consider a new mixture of time homogeneous finite Markov chains and develop the EM algorithm for the maximum likelihood estimation of its parameters. The proposed mixture generalizes the mover-stayer model. A mover-stayer model postulates a simple form of population heterogeneity: the population consists of two types of individuals: “movers” and “stayers”.

“Movers” evolve according to a Markov chain, whereas “stayers” stay in their initial states. Thus, a continuous time mover-stayer model is a mixture of two Markov chains, one which evolves according to some intensity matrix Q, and the other whose transition probability matrix is an identity matrix. The transition probability matrix, $P(t)$, of a continuous time mover-stayer model on state space $D = \{1, 2, .., w\}$ is

$$P(t) = SI + (I - S)\exp(tQ), t \geq 0,$$

where $S =$diag$(s_1, s_2, ..., s_w)$, with

$$s_r = \text{proportion of individuals among those who are initially in state } r \text{ who are "stayers"},$$

and Q is a matrix with entries q_{ij} satisfying

$$q_{ii} \leq 0, q_{ij} \geq 0, \sum_{j \neq i} q_{ij} = -q_{ii} \equiv q_i, i \in D.$$

In particular the diagonal entries in a Q matrix have the following interpretation

$$\frac{1}{-q_{ii}} = \text{expected length of time for an individual in state } i \text{ to remain in that state.} \tag{1}$$

In discrete time, the n-step transition matrix of the mover-stayer model, $P^{(n)}$, is

$$P^{(n)} = SI + (I - S)M^n, n \geq 0,$$

where M is the one-step transition probability matrix of a discrete time Markov chain.

The discrete time mover-stayer model was introduced by Blumen, Kogan and McCarthy (1955) to account for the special discrepancy between observed and predicted by a Markov chain transition matrices in the context of modeling inter-industry labor mobility (see Singer and Spilerman (1977) for the discussion of some theoretical issues involved in modeling with the mover-stayer model.) The discrete time version of the mover-stayer model has since been popular in various application contexts as providing a more realistic and better description of an empirical process than a Markov chain. The substantive contexts varied over occupational mobility (Mahoney and Milkovich (1971), Sampson (1990))) income dynamics (McCall (1973)), consumer brand preferences (Chaterjee and Ramaswamy (1996), Colombo and Morrison (1989)), bond ratings migrations (Altman and Kao (1991))), credit behavior (Frydman, Kallberg and Kao (1985)), and tumor progression (Tabar et al (1996), Chen et al (1997)) to list only a few topics. However, in at least some of these substantive contexts, the assumption that part of the population never leaves their original states seems to be too restrictive. For example, even though Altman and Kao (1991) demonstrated that the mover-stayer model fits the bond ratings data better than a Markov chain, the mover stayer model is unlikely to become an established model in this area because it entails financially implausible assumption that some bonds will never change their ratings.

To provide a richer framework than the mover-stayer model for describing population heterogeneity with respect to the rate of movement we propose the following generalization of the continuous time mover-stayer model. Let Q be an intensity matrix and consider a set of intensity matrices Ω obtained from Q :

$$\Omega = \{\Gamma Q : \Gamma = \text{diag}(\gamma_1, \gamma_2, ...\gamma_w), \gamma_r \geq 0, 1 \leq r \leq w\}$$

Now consider a mixture of N independent Markov chains such that each Markov chain has an intensity matrix in Ω, and which includes a Markov chain generated by Q. The N intensity matrices are: $A_m \equiv \Gamma_m Q, 1 \leq m \leq N$, where Γ_m =diag$(\gamma_{1,m}, \gamma_{2,m}, ...\gamma_{w,m})$ with $\Gamma_N = I$, the identity matrix, so that $\Gamma_N Q = Q$. The discrete mixing distribution on these Markov chains is obtained by specifying for $i \in D$ and $1 \leq m \leq N$

$$s_{i,m} \;=\; \text{proportion of realizations initially in state } i \text{ generated by } A_m,$$

where, for every i, $s_{i,m} \geq 0$ and $\sum_{m=1}^{N} s_{i,m} = 1$. The transition probability matrix of the mixture process is

$$P(t) = \sum_{m=1}^{N} S_m P_m(t), t \geq 0, \tag{2}$$

where $P_m(t) = \exp(tA_m)$, and S_m =diag$(s_{1,m}, s_{2,m}, ..., s_{w,m})$. Note that mixtures of Markov chains do not have a Markov property.

The mover-stayer model is obtained from the mixture in (2) by setting $N = 2$ and $\gamma_{1,1} = \gamma_{2,1} = ... = \gamma_{w,1} = 0$. By (1) when, $0 < \gamma_{i,1} < 1, i \in D$, the realizations generated by A_1 move, on average, more slowly than realizations generated by Q. Thus, in particular, the proposed mixture can easily accommodate slowly moving individuals without requiring them to be "stayers".

A discrete time analog of the mixture process in (2) is also defined. For both continuous and discrete time mixtures we consider the maximum likelihood estimation of their parameters from a sample of independent continuously observed realizations of the mixture process. For each mixture we first obtain the maximum likelihood estimators (mles) of their parameters under the complete information, that is, when we also know which Markov chain generated each realization. These are then used in the expectation step of the EM algorithms for the estimation of the mixtures under incomplete information. For a continuous time mixture we present the details of the EM algorithm. We omit the statement of the EM algorithm for a discrete time mixture since it is similar to the statement for a continuous time mixture. To simplify the exposition and without loss of generality the results are developed for mixtures of two Markov chains. They can be easily extended, as indicated below, to a mixture of an arbitrary number of Markov Chains.

A small simulation study involving a mixture of two continuous time Markov chains was conducted to investigate the convergence properties of the EM algorithm and the dependence of the accuracy of the estimates on the magnitudes of true values of γ's.

Since the mover-stayer model is a special case of the proposed mixture, our work extends the results in Frydman (1984) and in Fuchs and Greenhouse (1988). Fuchs and Greenhouse (1988) discussed the EM algorithm for obtaining the ml estimates in the discrete time mover-stayer model. The EM algorithm developed here reduces, in the case of the mover-stayer model, to the simpler algorithm than the one in Fuchs and Greenhouse (1988).

Similar mixtures of Markov chains to the ones defined above, but not their estimation, were considered in Singer and Spilerman (1976). Aalen (1988) discussed some probabilistic aspects of mixtures of time homogeneous Markov chains.

The paper is organized as follows. In Sections 2 and 4 we obtain the mles of the parameters of a continuous and discrete time mixture, respectively, under complete information. In Section 3 we present the EM algorithm for the continuous time mixture and indicate how this algorithm can be extended to the case of a mixture of more than two Markov chains. Section 5 describes the simulation study and its results.

2 Estimation in the continuous-time mixture with complete information

Let $X = (X_t, t \geq 0)$, be a mixture of two Markov chains with the transition probability function defined in (2). To simplify notation, let $\Gamma \equiv \Gamma_1$, $A \equiv \Gamma Q$, where $\Gamma =$diag$(\gamma_1, \gamma_2, ...\gamma_w)$. Similarly, let $S \equiv S_1$, where $S =$diag$(s_1, s_2, ..., s_w)$. In this notation the transition probability function of X is

$$P(t) = S\exp(tA) + (I - S)\exp(tQ), t \geq 0. \tag{3}$$

Assume that we observe n independent realizations of X on time interval $[0, T]$. More precisely assume that the k'th realization, X^k, is observed continuously on time interval $[0, T^k]$ with $T^k \leq T$. Thus, the individual realizations may be observed over time intervals of different lengths. This may be the case when right censoring is present or when the mixture process has an absorbing state. The right censoring is assumed to be independent. In case of complete information, we also know whether the realization was generated by Q or by A. We will refer to a realization generated by Q (A) as a Q $(A)-$ realization. For X^k, we summarize the observations as $(n^k_{ij}, \tau^k_i, Y_k, i, j \in D, j \neq i)$ where

$$\begin{aligned} n^k_{ij} &= \text{number of times } X^k \text{ makes an } i \rightarrow j \text{ transition, } i \neq j, \\ \tau^k_i &= \text{total time } X^k \text{ spends in state } i, \end{aligned}$$

and $Y_k = 1$ if X^k is generated by A, and $Y_k = 0$, otherwise.

Let L^Q_k be the likelihood of observing X^k under Q, conditional on knowing the initial state. By the result in Albert (1962)

$$L^Q_k = \left(\prod_{i \neq j} (q_{ij})^{n^k_{ij}} \prod_i \exp(-q_i \tau^k_i) \right).$$

Similarly define L^A_k to be the likelihood of observing X^k under A. If $(Y_k, 1 \leq k \leq n)$ are not observed, the likelihood of X^k, L^*_k, is

$$L^*_k = \prod_{i=1}^{w} (s_i)^{I^k_i} L^A_k + \prod_{i=1}^{w} [(1 - s_i)]^{I^k_i} L^Q_k,$$

where

$$\begin{aligned} I_i^k &= 1 \text{ if } X_0^k = i \\ &= 0, \text{ otherwise.} \end{aligned}$$

The likelihood function of n independent realizations, $L^* \equiv \prod_{k=1}^n L_k^*$, is difficult to maximize directly. Instead we develop the EM algorithm for obtaining the mles of Q, Γ and S. The maximization step of the EM algorithm requires the maximum likelihood estimators of the parameters based on complete information. We now obtain these estimators which we denote by $\tilde{Q}, \tilde{\Gamma}$ and $\tilde{S}$. The likelihood function, L_k, of X^k with complete information is

$$\begin{aligned} L_k &= \left\{ \prod_i (s_i)^{I_i^k} \, L_k^A \right\}^{Y_k} \left\{ \prod_i [(1-s_i)]^{I_i^k} \, L_k^Q \right\}^{(1-Y_k)} \\ &= \prod_i (s_i)^{I_i^k Y_k} (1-s_i)^{I_r^k (1-Y_k)} \prod_{i \neq j} (q_{ij})^{n_{ij}^k} \times \\ &\quad \left(\prod_i (\gamma_i)^{n_i^k} \exp(-q_i \gamma_i \tau_i^k) \right)^{Y_k} \left(\prod_i \exp(-q_i \tau_i^k) \right)^{1-Y_k} \\ &= \prod_i [(1-s_i)]^{I_i^k} \left[s_i/(1-s_i) \right]^{I_i^k Y_k} \prod_{i \neq j} (q_{ij})^{n_{ij}^k} \times \\ &\quad \prod_i (\gamma_i)^{n_i^k Y_k} \exp(-q_i \tau_i^k) \exp\left[q_i \tau_i^k (1-\gamma_i) Y_k \right] \end{aligned}$$

where

$$n_i^k = \sum_{j \neq i} n_{ij}^k = \text{ the total number of transitions of } X^k \text{ from state } i.$$

The loglikelihood of X^k is

$$\begin{aligned} \log L_k &= \sum_i I_i^k \log(1-s_i) + Y_k \sum_i I_r^k \log\left[s_i/(1-s_i) \right] \\ &\quad + \sum_{i \neq j} n_{ij}^k \log(q_{ij}) + Y_k \sum_i n_i^k \log(\gamma_i) - \sum_i q_i \tau_i^k \\ &\quad + Y_k \sum_i \left[q_i \tau_i^k (1-\gamma_i) \right], \end{aligned}$$

and for all realizations becomes

$$\begin{aligned}\log L &= \sum_i \sum_{k=1}^n I_i^k \log(1-s_i) \\ &\quad + \sum_i \sum_{k=1}^n I_i^k Y_k \log\left[s_i/(1-s_i)\right] \\ &\quad + \sum_{i\neq j} \sum_{k=1}^n n_{ij}^k \log(q_{ij}) + \sum_i \sum_{k=1}^n n_i^k Y_k \log(\gamma_i) \\ &\quad - \sum_i \sum_{k=1}^n q_i \tau_i^k + \sum_i \sum_{k=1}^n Y_k \left[q_i \tau_i^k (1-\gamma_i)\right]\end{aligned}$$

or

$$\begin{aligned}\log L &= \sum_i b_i \log(1-s_i) \\ &\quad + \sum_i b_i^A \log\left[s_i/(1-s_i)\right] \\ &\quad + \sum_{i\neq j} n_{ij} \log(q_{ij}) + \sum_i n_i^A \log(\gamma_i) - \sum_i q_i \tau_i \\ &\quad + \sum_i q_i (1-\gamma_i) \tau_i^A \end{aligned} \qquad (4)$$

where

$$\begin{aligned} b_i &= \sum_{k=1}^n I_i^k = \text{total number of realizations that begin in state } i, \\ b_i^A &= \sum_{k=1}^n I_i^k Y_k = \text{total number of } A-\text{realizations that begin in state } i, \end{aligned}$$

$$
\begin{aligned}
n_{ij} &= \sum_{k=1}^{n} n_{ij}^{k} = \text{total number of } i \to j \text{ transitions in the sample,} \\
n_{i} &= \sum_{k=1}^{n} n_{i}^{k} = \text{total number of transitions out of state } i \text{ in the sample,} \\
n_{i}^{A} &= \sum_{k=1}^{n} n_{i}^{k} Y_k = \text{total number of transitions out of state } i \\
& \quad \text{for all } A - \text{realizations,} \\
\tau_{i} &= \sum_{k=1}^{n} \tau_{i}^{k} = \text{total time in state } i \text{ for all realizations in the sample,} \\
\tau_{i}^{A} &= \sum_{k=1}^{n} Y_k \tau_{i}^{k} = \text{total time in state } i \text{ for all } A - \text{realizations.}
\end{aligned}
$$

Setting

$$\frac{\partial \log L}{\partial s_i} = 0,$$

we obtain

$$-\frac{b_i}{1-s_i} + b_i^A\left(\frac{1}{s_i} + \frac{1}{1-s_i}\right) = 0$$

or for $1 \leq i \leq w$,

$$\tilde{s}_i = \frac{b_i^A}{b_i}. \tag{5}$$

Now setting

$$\frac{\partial \log L}{\partial \gamma_i} = \frac{n_i^A}{\gamma_i} - q_i \tau_i^A = 0, \tag{6}$$

gives

$$\gamma_i = \frac{n_i^A}{q_i \tau_i^A}. \tag{7}$$

Substituting (7) into the loglikelihood function we obtain

$$
\begin{aligned}
\log L \sim & \sum_{i \neq j} n_{ij} \log(q_{ij}) + \sum_i n_i^A \log n_i^A - \sum_i n_i^A \log\left(q_i \tau_i^A\right) - \sum_i q_i \tau_i \\
& + \sum_i q_i \tau_i^A - \sum_i n_i^A.
\end{aligned}
$$

Now setting

$$\frac{\partial \log L}{\partial q_{ij}} = 0,$$

gives

$$n_{ij} q_i - n_i^A q_{ij} - \tau_i^Q q_i q_{ij} = 0$$

so that

$$q_{ij} = \frac{n_{ij} q_i}{n_i^A + \tau_i^Q q_i}.$$

Using $\sum_{j \neq i}^{w} q_{ij} = q_i$, we obtain

$$\sum_{j \neq i}^{w} \frac{n_{ij}}{n_i^A + \tau_i^Q q_i} = 1,$$

or

$$\frac{n_i}{n_i^A + \tau_i^Q q_i} = 1.$$

Thus, the maximum likelihood estimators with complete information are given by (5) and for $1 \leq i \leq w$, and $j \neq i$, by

$$\tilde{q}_i = \frac{n_i - n_i^A}{\tau_i^Q} = \frac{n_i^Q}{\tau_i^Q}, \tag{8}$$

$$\tilde{q}_{ij} = \frac{n_{ij} \tilde{q}_i}{n_i^A + \tau_i^Q \tilde{q}_i} = \frac{n_{ij}}{n_i} \tilde{q}_i, \tag{9}$$

and

$$\tilde{\gamma}_i = \frac{n_i^A}{\tilde{q}_i \tau_i^A} = \frac{n_i^A \tau_i^Q}{n_i^Q \tau_i^A}, \tag{10}$$

where $n_i^Q = n_i - n_i^A$ and $\tau_i^Q = \tau_i - \tau_i^A$.

The extension of the described estimation procedure to the mixture of N Markov Chains with $N > 2$ is straightforward and thus we only state the result. The mle of Q is again given by (8) and (9). The mles of Γ_m and S_m are

$$\tilde{\gamma}_{i,m} = \frac{n_i^{A_m}}{\tilde{q}_i \tau_i^{A_m}}, \tilde{s}_i^m = \frac{b_i^{A_m}}{b_i}, 2 \leq m \leq N, i \in D$$

where quantities $n_i^{A_m}, \tau_i^{A_m}, b_i^{A_m}$ are defined in the same way as n_i^A, τ_i^A, b_i^A, but with respect to generator A_m.

3 The EM algorithm for the continuous-time mixture.

I describe the steps of the EM algorithm. As a consequence of the expression in (9), the value of q_{ij} does not have to updated at each iteration of the algorithm. Thus, the parameters estimated by the algorithm are $(s_i, q_i, \gamma_i, i \in D)$. After the algorithm converges at, say, $(\widehat{s}_i, \hat{q}_i, \hat{\gamma}_i, i \in D)$, $\hat{q}_{ij}$ is computed using (9):

$$\hat{q}_{ij} = \frac{n_{ij}}{n_i}\hat{q}_i, j \neq i. \tag{11}$$

Step 1 Choose initial values $(s_i^0, q_i^0, \gamma_i^0, i \in D)$ and define Q^0 to be the intensity matrix with the entries given by $q_{ij}^0 = (n_{ij}/n_i)q_i^0, i \neq j$, and $q_{ii}^0 = -q_i^0$. Similarly define A^0 to be the intensity matrix with $a_{ij}^0 = \gamma_i^0 q_{ij}^0, i \neq j$, and $a_{ii}^0 = -\gamma_i^0 q_i^0$.

Step 2 (Expectation step) Let

$$\begin{aligned}
L_k^{A^0} &= \prod_{i \neq j}(\gamma_i^0 q_{ij}^0)^{n_{ij}^k} \prod_i \exp(-\gamma_i^0 q_i^0 \tau_i^k) \\
&= \prod_{i \neq j}(\gamma_i^0 \frac{n_{ij}}{n_i} q_i^0)^{n_{ij}^k} \prod_i \exp(-\gamma_i^0 q_i^0 \tau_i^k), \\
L_k^{Q^0} &= \prod_{i \neq j}(\frac{n_{ij}}{n_i} q_i^0)^{n_{ij}^k} \prod_i \exp(-q_i^0 \tau_i^k).
\end{aligned}$$

For the k'th history, which starts in state r, $1 \leq k \leq n, r \in D$, compute the probability that it is generated by A^0 :

$$\begin{aligned}
E^0(Y_k) &= \frac{s_r^0 L_k^{A^0}}{s_r^0 L_k^{A^0} + (1 - s_r^0) L_k^{Q^0}} \\
&= \frac{s_r^0 \prod_i (\gamma_i^0)^{n_i^k}}{s_r^0 \prod_i (\gamma_i^0)^{n_i^k} + (1 - s_r^0) \prod_i \exp\left[q_i^0 \tau_i^k (\gamma_i^0 - 1)\right]}.
\end{aligned}$$

Then for $i \in D$, compute the following quantities

$$\begin{aligned}
E^0(n_i^A) &\equiv \sum_{k=1}^{n} n_i^k E^0(Y_k), E^0(n_i^Q) \equiv n_i - E^0(n_i^A), \\
E^0(\tau_i^A) &\equiv \sum_{k=1}^{n} \tau_i^k E^0(Y_k), E^0(\tau_i^Q) \equiv \tau_i - E^0(\tau_i^A), \\
E^0(b_i^A) &\equiv \sum_{k=1}^{n} I_i^k E^0(Y_k).
\end{aligned}$$

Clearly $E^0(n_i^A)$ is the expected total number of transitions out of state i for all realizations generated by A, conditional on the available data and computed under $(s_i^0, q_i^0, \gamma_i^0, i \in D)$.

Step 3 (Maximization step). Compute new values $(s_i^1, q_i^1, \gamma_i^1, i \in D)$ using (6),(9) and (10) by

$$s_i^1 = \frac{E^0(b_i^A)}{b_i}.$$

$$q_i^1 = \frac{E^0(n_i^Q)}{E^0(\tau_i^Q)},$$

$$\gamma_i^1 = \frac{E^0(n_i^A)}{q_i^1 E^0(\tau_i^A)}.$$

Step 4 Stop if a solution is found with required accuracy. Otherwise, return to Step 2 and replace $(s_i^0, q_i^0, \gamma_i^0, i \in D)$ with $(s_i^1, q_i^1, \gamma_i^1, i \in D)$.

In any particular application the standard errors of the estimators $(\widehat{s}_i, \hat{q}_i, \hat{\gamma}_i, \hat{q}_{ij}, i \neq j \in D)$, resulting from the EM algorithm, can be computed using the method introduced in Louis (1982). The method requires computation of the score vector and the second derivative matrix corresponding to the complete information loglikelihood. In our case the complete information loglikehood, $logL$ in (4), is easily twice differentiable with respect to the parameters and thus the method in Louis (1982) can be applied in a straightforward manner to obtain the standard errors of $(\widehat{s}_i, \hat{q}_i, \hat{\gamma}_i, \hat{q}_{ij}, i \neq j \in D)$.

4 Estimation of the discrete-time mixture with complete information.

Let $M = (m_{ij})$ be a one-step transition probability matrix of a discrete time Markov chain on D, and define a new one step transition matrix B by

$$\begin{aligned} b_{ii} &= 1 - \lambda_i + \lambda_i m_{ii}, \\ b_{ij} &= \lambda_i m_{ij}, i, j \in D, \end{aligned}$$

so that

$$B = I - \Lambda + \Lambda M,$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_w)$ with

$$0 \leq \lambda_i \leq \frac{1}{1 - m_{ii}}, 1 \leq i \leq w. \tag{12}$$

By construction B is a stochastic matrix. Note that the expected length of stay in state i of a Markov chain governed by M is $1/(1 - m_{ii})$ and of that governed by B is $1/\left[\lambda_i(1 - m_{ii})\right]$. The condition (12) is not restrictive; it states that

$$1/\left[\lambda_i(1 - m_{ii})\right] \geq 1.$$

The discrete-time mixture has the n-step transition probability matrix given by

$$P^{(n)} = SB^n + (I - S)M^n, n \geq 1,$$

where $S =$diag$(s_1, s_2, ..., s_w)$, and s_i is the proportion of individuals initially in state i who move according to a Markov chain with the transition probability matrix B. The mover-stayer model is obtained from the mixture process by setting $\lambda_i = 0, i \in D$.

As in continuous time case the data consist of n independent realizations, $\{X^k, 1 \leq k \leq n\}$, of a discrete time mixture. In addition to the notation introduced before, let

$$n_{ii}^k = \text{total number of times } X^k \text{ makes an } i \rightarrow i \text{ transition.}$$

Let $Y_k = 1$ if X^k is generated by B, and let $Y_k = 0$, otherwise. The loglikelihood of X^k, L_k, based on observing $(n_{ij}^k, Y_k, i, j \in D)$ is

$$\log L_k = Y_k \log \left\{ \prod_{i=1}^{w} (s_i)^{I_i^k} L_k^B \right\} + (1 - Y_k) \log \left\{ \prod_{i=1}^{w} \left[(1 - s_i)\right]^{I_i^k} L_k^M \right\} \tag{13}$$

where, see Anderson and Goodman (1957),

$$L_k^M = \prod_{i,j} (m_{ij})^{n_{ij}^k} = \prod_i (m_{ii})^{n_{ii}^k} \prod_{i \neq j} (m_{ij})^{n_{ij}^k}\,, \tag{14}$$

is the likelihood of the k'th realization generated by M, and similarly

$$\begin{aligned} L_k^B &= \prod_{i,j} (b_{ij})^{n_{ij}^k} = \prod_i (b_{ii})^{n_{ii}^k} \prod_{i \neq j} (b_{ij})^{n_{ij}^k} = \prod_i (b_{ii})^{n_{ii}^k} \prod_{i \neq j} (\lambda_i m_{ij})^{n_{ij}^k} \\ &= \prod_i (b_{ii})^{n_{ii}^k} \prod_{i \neq j} (m_{ij})^{n_{ij}^k} \prod_i (\lambda_i)^{n_i^k}\,, \end{aligned} \tag{15}$$

is the likelihood of the k'th realization generated by B, and the rest of notation is as in Section 2. Substituting (14) and (15) into (13) gives

$$\begin{aligned} \log L_k &= \sum_i I_i^k \log(1-s_i) + Y_k \sum_i I_i^k \log\left[s_i/(1-s_i)\right] \\ &\quad + Y_k \sum_i n_{ii}^k \log b_{ii} + Y_k \sum_{j \neq i} n_{ij}^k \log m_{ij} + Y_k \sum_i n_i^k \log \lambda_i \\ &\quad + (1-Y_k) \sum_i n_{ii}^k \log m_{ii} + (1-Y_k) \sum_{j \neq i} n_{ij}^k \log m_{ij} \\ &= \sum_i I_i^k \log(1-s_i) + Y_k \sum_i I_i^k \log\left[s_i/(1-s_i)\right] \\ &\quad + \sum_{j \neq i} n_{ij}^k \log m_{ij} + Y_k \sum_i n_{ii}^k \log b_{ii} + Y_k \sum_i n_i^k \log \lambda_i \\ &\quad + (1-Y_k) \sum_i n_{ii}^k \log m_{ii} \end{aligned}$$

The loglikelihood for all realizations is

$$\begin{aligned} \log L &= \sum_i m_i \log(1-s_i) + \sum_i m_i^B \log\left[s_i/(1-s_i)\right] \\ &\quad + \sum_{j \neq i} n_{ij} \log m_{ij} + \sum_i n_{ii}^B \log b_{ii} + \sum_i n_i^B \log \lambda_i \\ &\quad + \sum_i n_{ii}^M \log m_{ii}. \end{aligned} \tag{16}$$

In what follows the same notation is used as in Section 2 with an intensity matrix A replaced by a transition probability matrix B. As in continuous time case solving $\partial \log L/\partial s_i = 0$ gives

$$\tilde{s}_i = b_i^B/b_i.$$

Then solving $\partial \log L/\partial \lambda_i = 0$ gives

$$\lambda_i = \frac{n_i^B}{(1-m_{ii})(n_i^B + n_{ii}^B)}, \tag{17}$$

and thus

$$\tilde{b}_{ii} = 1 - \lambda_i + \lambda_i m_{ii} = \frac{n_{ii}^B}{v_i^B}, \tag{18}$$

where $v_i^B = (n_i^B + n_{ii}^B)$. Substituting (17) and (18) into (16) we get, up to the terms not depending on the parameters

$$\begin{aligned} \log L \;\; \sim \;\; & \sum_{i \neq j} n_{ij} \log(m_{ij}) - \sum_i n_i^B \log(1 - m_{ii}) \\ & + \sum_i n_{ii}^M \log m_{ii}. \end{aligned}$$

Next by setting

$$\frac{\partial \log L}{\partial m_{ij}} = \frac{n_{ij}}{m_{ij}} - \frac{n_i^B}{1-m_{ii}} - \frac{n_{ii}^M}{m_{ii}} = 0,$$

we obtain

$$m_{ij} = \frac{n_{ij}(1-m_{ii})m_{ii}}{n_i^B m_{ii} + n_{ii}^M(1-m_{ii})}, i \neq j. \tag{19}$$

Now solving the following equation

$$\sum_{i \neq j} m_{ij} = \sum_{j \neq i} \frac{n_{ij}(1-m_{ii})m_{ii}}{n_i^B m_{ii} + n_{ii}^M(1-m_{ii})} = 1 - m_{ii}$$

for m_{ii}, we obtain the mle of m_{ii}

$$\tilde{m}_{ii} = \frac{n_{ii}^M}{n_i^M + n_{ii}^M} \equiv \frac{n_{ii}^M}{v_i^M}, \tag{20}$$

and substituting (20) into (17) gives the mle of λ_i

$$\tilde{\lambda}_i = \frac{n_i^B}{v_i^B}\frac{v_i^M}{n_i^M} = \frac{n_i^B}{v_i^B}(1-\tilde{m}_{ii})^{-1}.$$

Finally substituting (20) into (19) gives the mle of m_{ij}.

$$\tilde{m}_{ij} = \frac{n_{ij}}{n_i} \frac{n_i^M}{v_i^M} = \frac{n_{ij}}{n_i}(1 - \tilde{m}_{ii}). \tag{21}$$

The EM algorithm can now be developed for the estimation of the parameters of a discrete time mixture when $(Y_k, 1 \leq k \leq n)$ are unknown along the same lines as in the continuous time case. The parameters estimated by the algorithm are $(s_i, m_{ii}, \lambda_i, i \in D)$. The estimates of the transition probabilities $m_{ij}, i \neq j$, are then obtained from (21). In particular for the estimation in the discrete time mover-stayer model (that is, when $\lambda_i = 0, i \in D$) the outlined EM algorithm estimates $(s_i, m_{ii}, i \in D)$ and $m_{ij}, i \neq j$, are obtained from (21). Thus, this version of the EM algorithm seems to be more efficient for estimation in the discrete time mover-stayer model than the one discussed by Fuchs and Greenhouse (1982), which requires that all parameters are updated at each iteration.

5 Simulation study

The purpose of this simulation is to study the convergence properties of the EM algorithm and to assess how the accuracy of estimation of various parameters depends on the magnitude of the γ parameters. The realizations are simulated from the mixture of two continuous time Markov chains with the transition probability function given in (3). The two Markov chains have four states: $\{1, 2, 3, 4\}$, with 4 being an absorbing state. For all simulations the true values were

$$s_1 = 0.7, s_2 = 0.5, s_3 = 0.3,$$

and

$$Q = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{pmatrix} -2 & 1 & 0.95 & 0.05 \\ 1.2 & -3 & 1.65 & 0.15 \\ 1.8 & 2 & -4 & 0.2 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Four different specifications of $(\gamma_1, \gamma_2, \gamma_3)$ were used:

$$\begin{aligned} (1) &: \quad \gamma_1 = \gamma_2 = \gamma_3 = 0, \\ (2) &: \quad \gamma_1 = \gamma_2 = \gamma_3 = 0.05 \\ (3) &: \quad \gamma_1 = \gamma_2 = \gamma_3 = 0.25, \\ (4) &: \quad \gamma_1 = \gamma_2 = \gamma_3 = 0.5. \end{aligned} \tag{22}$$

Note that the first specification corresponds to the mover-stayer model. Twenty simulations were conducted for each specification of gammas. Each simulation consisted of simulating 400 realizations from the mixture process on the interval $[0, 5]$. If absorption took place before time 5, the observed history was recorded until the absorption time. The initial distribution in all simulations was $\eta = (1/3, 1/3, 1/3, 0)$. In addition, for the specification with $\gamma_1 = \gamma_2 = \gamma_3 = 0.05$, ten simulations were conducted each comprising 800 realizations. A single realization was simulated as follows:

(i) An initial state, X_0, was selected according to the initial distribution η.

(ii) If $X_0 = i$, a Markov chain was selected according to the distribution $(s_i, 1 - s_i), 1 \leq i \leq 3$, that is, a Markov chain with the generator A was selected with probability s_i and the one with generator Q was selected with probability $1 - s_i$.

(iii) A realization of the selected Markov chain was simulated on the time interval $[0, 5]$. If a Markov chain with generator Q was selected then the realization was simulated as follows.

Given $X_0 = i$, obtain a realization T_1 from an exponential distribution with parameter q_i. T_1 is the time that the process stays in state i. Then use the $i'th$ row of the embedded transition probability matrix:

$$P_Q = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{pmatrix} 0 & 0.5 & 0.475 & 0.025 \\ 0.4 & 0 & 0.55 & 0.05 \\ 0.45 & 0.5 & 0 & 0.05 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

to draw the state to which the process makes a transition from state i. If this is state j, $j \neq 4$, then obtain, independently of T_1, a realization T_2 from an exponential distribution with parameter q_j. Then use the $j'th$ row of matrix

P_Q to draw the state to which the process makes a transition from state j. Continue in this way until the first time that $\sum_i T_i \geq 5$, or a sample realization hits the absorbing state 4.

The convergence of the EM algorithm was investigated with different initial values for the simulated data sets. In each case the algorithm converged to the same final values independently of the initial values. In the simulations reported below the following values were subsequently used as initial values for the parameters:

$$s_i^0 = \frac{b_i^s}{b_i}, q_i^0 = \frac{n_i}{\tau_i}, \gamma_i^0 = 0.5, 1 \leq i \leq 3,$$

where, for a given simulation, b_i^s is the number of realizations that start in state i and never leave state i, and b_i is as defined above. If $b_i^s = 0$, then an arbitrary small number was used for s_i^0.

The algorithm was assumed to converge when the difference between the updated value of each parameter and its previous value was no more than 0.001. The convergence of the algorithm was very fast in all simulations. The average number of iterations of the EM algorithm was 3.9, 4.75, 12.65 and 42.85, respectively, for the four sets of gamma parameters in (22).

For each set of true values of $(\gamma_1, \gamma_2, \gamma_3)$ the results of the simulations are summarized in Tables $1. - 4.$ by reporting the average values $(\bar{s}, \bar{Q}, \bar{\gamma})$ of the estimates over 20 simulations, their root mean squared errors (rmses), and for the estimates of $(\gamma_1, \gamma_2, \gamma_3)$, also their relative rmses obtained by dividing the root mean squared error by the true value of the parameter.

In all four simulations the average values of the estimates, $(\bar{s}, \bar{Q}, \bar{\gamma})$, are very close to the true values. Interestingly, for the simulation with $\gamma_1 = \gamma_2 = \gamma_3 = 0$, the average values of estimated gammas are zero. In fact in each of the twenty simulations with $\gamma_1 = \gamma_2 = \gamma_3 = 0$, the gammas are estimated by the EM algorithm to be zero (up to the third decimal point) resulting in the rmses of zero for all gammas. This shows that when the true model is the mover-stayer model the EM algorithm will identify the mixture process as the mover-stayer model.

The rmses for s_1, s_2, s_3 and for q_1, q_2, q_3 are comparable across all four simulations. The rmses of $q_{ij}, j \neq i$, tend to decrease as true gammas increase. This can be explained by observing that the effective sample size for the estimation of $q_{ij}, j \neq i$, increases as true gammas increase. Note that by (11) the effective sample size for the estimation of $q_{ij}, j \neq i$, is $n_i, 1 \leq i \leq 3$. As true gammas increase the realizations generated by A will tend to move faster through the state space resulting in larger total number of transitions out of state $i : n_i, 1 \leq i \leq 3$.

The relative rmses for gammas in simulations $(2)-(4)$ also tend to decrease as true values of gammas increase. The relative rmses for gammas are quite large when $\gamma_1 = \gamma_2 = \gamma_3 = 0.05$. They were also large in other simulations, not reported here, with small true values of gammas. This prompted an additional simulation with $\gamma_1 = \gamma_2 = \gamma_3 = 0.05$ consisting of a series of ten simulations comprising 800 realizations each. The results of this simulation, reported in Table 5, show that the relative rmses are now much smaller. Thus, larger sample sizes may be needed to accurately estimate gammas and $q_{ij}, j \neq i$, when the true gammas are small.

Table 1. True values: $\gamma_1 = \gamma_2 = \gamma_3 = 0$. Averages of the estimates and their rmses (in parentheses) based on 20 simulations with each simulation consisting of 400 realizations.

$$\begin{aligned}
\bar{s} &= \begin{pmatrix} 0.688 & 0.506 & 0.318 \\ (.046) & (.047) & (.047) \end{pmatrix}, \\
\bar{\gamma} &= \begin{pmatrix} 0 & 0 & 0 \\ (0) & (0) & (0) \end{pmatrix}
\end{aligned}$$

$$\bar{Q} = \begin{array}{c} 1 \\ \\ 2 \\ \\ 3 \\ \\ 4 \end{array} \left(\begin{array}{cccc}
-1.975 & 0.981 & 0.943 & 0.051 \\
(.076) & (.064) & (.053) & (.013) \\
1.209 & -3.046 & 1.679 & 0.158 \\
(.027) & (.139) & (.099) & (.027) \\
1.778 & 1.986 & -3.969 & 0.204 \\
(.112) & (.093) & (.153) & (.021) \\
0 & 0 & 0 & 0
\end{array} \right).$$

Table 2. True values: $\gamma_1 = \gamma_2 = \gamma_3 = 0.05$. Averages of the estimates and their rmses (in parentheses) based on 20 simulations with each simulation consisting of 400 realizations. For estimates of gammas, the second number in parentheses is the relative rmse.

$$\begin{aligned}
\bar{s} &= \begin{pmatrix} 0.681 & 0.484 & 0.313 \\ (.046) & (.044) & (.038) \end{pmatrix}, \\
\bar{\gamma} &= \begin{pmatrix} 0.049 & 0.052 & 0.051 \\ (.008.17.141\%) & (.009, 16.941\%) & (.006, 11.507\%) \end{pmatrix}
\end{aligned}$$

$$\bar{Q} = \begin{array}{c} 1 \\ \\ 2 \\ \\ 3 \\ \\ 4 \end{array} \left(\begin{array}{cccc}
-2.032 & 1.012 & 0.972 & 0.048 \\
(.082) & (.072) & (.054) & (.011) \\
1.181 & -2.980 & 1.652 & 0.147 \\
(.026) & (.095) & (.071) & (.026) \\
1.755 & 2.052 & -3.999 & 0.192 \\
(.087) & (.101) & (.118) & (.03) \\
0 & 0 & 0 & 0
\end{array} \right).$$

Table 3. True values: $\gamma_1 = \gamma_2 = \gamma_3 = 0.25$. Averages of the estimates and their rmses (in parentheses) based on 20 simulations with each simulation consisting of 400 realizations. For estimates of gammas, the second number in parentheses is the relative rmse.

$$\begin{aligned}\bar{s} &= \begin{pmatrix} 0.702 & 0.495 & 0.293 \\ (.035) & (.04) & (.042) \end{pmatrix}, \\ \bar{\gamma} &= \begin{pmatrix} 0.25 & 0.25 & 0.254 \\ (.012, 4.721\%) & (.023, 9.385\%) & (.019, 7.439\%) \end{pmatrix}\end{aligned}$$

$$\bar{Q} = \begin{array}{c} 1 \\ \\ 2 \\ \\ 3 \\ \\ 4 \end{array} \begin{pmatrix} -1.985 & 0.996 & 0.944 & 0.045 \\ (.053) & (.039) & (.043) & (.01) \\ 1.200 & -2.994 & 1.641 & 0.152 \\ (.025) & (.107) & (.08) & (.025) \\ 1.761 & 2.02 & -3.967 & 0.186 \\ (.097) & (.125) & (.176) & (.033) \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Table 4. True values: $\gamma_1 = \gamma_2 = \gamma_3 = 0.5$. Averages of the estimates and their rmses (in parentheses) based on 20 simulations with each simulation consisting of 400 realizations. For estimates of gammas, the second number in parentheses is the relative rmse.

$$\begin{aligned}\bar{s} &= \begin{pmatrix} 0.702 & 0.507 & 0.281 \\ (.046) & (.046) & (.045) \end{pmatrix}, \\ \bar{\gamma} &= \begin{pmatrix} 0.505 & 0.505 & 0.501 \\ (.034, 6.682\%) & (.028, 5.481\%) & (.027, 5.439\%) \end{pmatrix},\end{aligned}$$

$$\bar{Q} = \begin{array}{c} 1 \\ \\ 2 \\ \\ 3 \\ \\ 4 \end{array} \begin{pmatrix} -1.997 & 1.003 & 0.944 & 0.049 \\ (.095) & (.042) & (.061) & (.007) \\ 1.203 & -3.005 & 1.646 & 0.156 \\ (.017) & (.127) & (.069) & (.017) \\ 1.786 & 1.998 & -3.986 & 0.202 \\ (.057) & (.078) & (.099) & (.015) \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Table 5. True values: $\gamma_1 = \gamma_2 = \gamma_3 = 0.05$. Averages of the estimates and their rmses (in parentheses) based on 10 simulations with each simulation consisting of 800 realizations. For estimates of gammas, the second number in parentheses is the relative rmse. It took on average 4.6 iterations for the EM algorithm to converge.

$$\begin{aligned}
\bar{s} &= \begin{pmatrix} 0.682 & 0.484 & 0.312 \\ (.023) & (.022) & (.022) \end{pmatrix}, \\
\bar{\gamma} &= \begin{pmatrix} 0.048 & 0.052 & 0.051 \\ (.004, 7.988\%) & (.004, 8.104\%) & (.003, 5.356\%) \end{pmatrix},
\end{aligned}$$

$$\bar{Q} = \begin{array}{c} 1 \\ \\ 2 \\ \\ 3 \\ \\ 4 \end{array} \begin{pmatrix} -2.032 & 1.012 & 0.972 & 0.048 \\ (.048) & (.038) & (.034) & (.007) \\ 1.18 & -2.979 & 1.652 & 0.147 \\ (.014) & (.043) & (.028) & (.014) \\ 1.754 & 2.051 & -3.997 & 0.192 \\ (.052) & (.054) & (.058) & (.019) \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Aknowledgement A substantial part of this work was completed during my visit to the Department of Biostatistics at the University of Copenhagen. I am grateful to the Department for its hospitality.

REFERENCES

Aalen, O. O. (1988) Dynamic Description of a Markov Chain with Random time scale. *Mathematical Scientist* 13, 90-103.

Albert, A. (1962) Estimating the Infinitesimal Generator of a Continuous Time Finite State Markov Process. *Annals of Mathematical Statistics,* 38, 727-753.

Altman, E. I. and Kao, D. L. (1991) Corporate Bond Rating Drift: An Examination of Rating Agency Credit Quality Changes. Charlottesville, VA: Association for Investment Management and Research.

Anderson, T. W. and Goodman, L. A. (1957) Statistical Inference about Markov Chains. *Annals of Mathematical Statistics,* 28, 89-100.

Blumen, I., Kogan, M. and McCarthy, P. J. (1955) *The Industrial Mobility of Labor as a Probability Process.* Cornell University Press, Ithaca, N. Y.

Chatterjee, R. and Ramaswamy, V. (1996), An Extended Mover-Stayer Model for Diagnosing the Dynamics of Trial and Repeat for a New Brand. *Applied Stochastic Processes and Data Analysis,* Vol. 12, 165-178.

Chen, H. H., Duffy, S. W. and Tabar, L. (1997) A Mover-Stayer Mixture of Markov Chain Models for the Assessment of Dedifferentiation and Tumor Progresion in Breast Cancer. *Journal of Applied Statistics* 24 (3), 265-278.

Colombo, R. A. and Morrison, D. G. (1989)) A Brand Switching Model with Implications for Marketing Strategies. *Marketing Science,* 8 *(*Winter*),* 89-99.

Frydman, H. (1984) Maximum likelihood Estimation in the Mover-Stayer Model. *Journal of the American Statistical Association,* 79, 632-638.

Frydman, H., Kallberg J. G. and Kao. D. L. (1985) Testing the Adequacy of Markov Chain and Mover-Stayer Models as Representations of Credit Behavior. *Operations Research* 33 (6)*,* 1203-1214.

Fuchs, C. and Greenhouse, J. B. (1988) The EM Algorithm for Maximum-Likelihood Estimation in the Mover-Stayer Model. *Biometrics* 44 (2), 605-613.

Louis, T. A. (1982) Finding the Observed Information Matrix when using the EM algorithm. *Journal of Royal Statistical Society, Series B,* 44 (2), 226-233.

Mahoney, T. A. and Milkovich, G. T. (1971) The Internal Labor Market as a Stochastic Process. In *Manpower and Management Science,* ed. D. J. Bartholomew and A. R. Smith, D. C. Heath, Lexington, Mass.

Mc Call, J. J. (1973) *Income Mobility, Racial Discrimination, and Economic Growth.* D. C. Heath, Lexington, Mass.

Sampson, M. (1990) A Markov Chain Model for Unskilled Workers and the Highly Mobile. *Journal of the American Statistical Association,* 85, 177-180.

Singer, B. and Spilerman, S. (1977) Trace Inequalities for Mixtures of Markov Chains. *Advances in Applied Probability* 9, 747-764.

Singer, B. and Spilerman, S. (1976) Some Methodological Issues in the Analysis of Longitudinal Surveys. *Annals of Economic and Social Measurement* 5/4, 447-474.