ELSEVIER

Journal of Econometrics 75 (1996) 79–97

JOURNAL OF Econometrics

# Calculating posterior distributions and modal estimates in Markov mixture models

Siddhartha Chib

*John M. Olin School of Business, Washington University, St. Louis, MO 63130, USA*

**Abstract**

This paper is concerned with finite mixture models in which the populations from one observation to the next are selected according to an unobserved Markov process. A new, full Bayesian approach based on the method of Gibbs sampling is developed. Calculations are simplified by data augmentation, achieved by introducing a population index variable into the list of unknown parameters. It is shown that the latent variables, one for each observation, can be simulated from their joint distribution given the data and the remaining parameters. This result serves to accelerate the convergence of the Gibbs sample. Modal estimates are also computed by stochastic versions of the EM algorithm. These provide an alternative to a full Bayesian approach and to existing methods of locating the maximum likelihood estimate. The ideas are applied in detail to Poisson data, mixtures of multivariate normal distributions, and autoregressive time series.

*Key words*: Autoregressive time series; Finite mixture distributions; Gibbs sampling; Hidden Markov models; Markov chain Monte Carlo; Markov switching models; Multivariate normal mixtures; Poisson distribution; Stochastic EM algorithm
*JEL classification*: C11; C12; C22

## 1. Introduction

### 1.1. Model

This paper is concerned with the Bayesian analysis of a class of finite mixture distributions in which the component populations, from one observation to the next, are selected according to an unobserved Markov process. This model

This paper has benefited from comments made by Herman van Dijk, two anonymous referees, and participants of the Rhine Riverboat Conference and the Midwest Econometric Group Meetings at Urbana-Champaign.


*SSDI* 0304-4076(95)01770-E

generalizes the situation in which the populations are selected *independently* according to a discrete probability mass function. A complete discussion of the latter class of models is provided by Everitt and Hand (1981) and Titterington, Smith, and Makov (1985).

In contrast to the classical mixture model, the Markov mixture model (MMM) is especially useful for modeling persistence, i.e., serial correlation in time series data. The general model can be described in terms of a sequence of *unobservable* finite state random variables, $s_t \in \{1, \ldots, m\}$, which evolve according to a Markov process:

$$s_t | s_{t-1} \sim \text{Markov}(P, \pi_1), \tag{1}$$

where $P = \{p_{ij}\}$ is the one-step transition probability matrix of the chain, i.e., $p_{ij} = \Pr(s_t = j \mid s_{t-1} = i)$, and $\pi_1$ is the probability distribution at $t = 1$. For identifiability reasons, assume that this chain is time-homogeneous, irreducible, and aperiodic. At each observation point $t$, a realization of the state occurs. Then, given that $s_t = k$, the observation $y_t$ is drawn from the population given by the conditional density

$$y_t \mid Y_{t-1}, \theta_k \sim f(y_t \mid Y_{t-1}, \theta_k), \qquad k = 1, \ldots, m, \tag{2}$$

where $Y_{t-1} = (y_1, \ldots, y_{t-1})$, $f$ is a density (or mass) function with respect to a $\sigma$ finite measure, and $\theta_k$ is the parameter vector of the $k$th population. Thus, the observation at $t$ is drawn from the finite mixture distribution

$$f(y_t | Y_{t-1}, s_{t-1}, \theta) = \begin{cases} \sum_{k=1}^{m} f(y_t | Y_{t-1}, \theta_k)\, \pi_1(s_t = k), & t = 1, \\ \sum_{k=1}^{m} f(y_t | Y_{t-1}, \theta_k)\, p(s_t = k | s_{t-1}), & t \geqslant 2, \end{cases} \tag{3}$$

where

$$\theta = \left\{ \bigcup_{k}^{m} \theta_k \right\} \cup \{ p_{ij}, 1 \leqslant i,\ j \leqslant m - 1 \}.$$

It is possible to obtain a number of important models as special cases of this structure. For example, the classical finite mixture model is obtained if $s_t$ is distributed independently and identically across time. Single and multiple change-point models can be also be obtained if suitable restrictions are placed on the off-diagonal elements of $P$. The Markov switching regression model of Goldfeld and Quandt (1973) and Markov switching autoregressive time series models are also particular cases of (3). Even the autoregressive models considered by Hamilton (1989) and Albert and Chib (1993), in which the conditional density in (2) depends on lagged values of $s_t$, can be put in this family by redefining the states (Hamilton, 1994, p. 691). However, in the latter case the support of the distribution of the states can become quite large.

In recent years, such models have attracted considerable attention in econometrics, biometrics, and engineering. These models are referred to as hidden Markov

models although the terminology Markov mixture models is more appropriate. A major problem with MMM's is that the likelihood function of the parameters is not available in simple form. Much earlier, Baum, Petrie, Soules, and Weiss (1970) addressed this problem and proposed a recursive algorithm, now called the forward–backward algorithm, to compute the likelihood function. Even so, Leroux and Puterman (1992) note that the algorithm is often not stable to small perturbations of the data. Leroux (1992) in an important paper has established the asymptotic properties of the maximum likelihood estimator.

A quite different approach to the estimation of mixture models is possible from the Bayesian Markov chain simulation perspective. Basically, the point is that the computation of the likelihood function can be avoided if the population index variable $\{s_t\}$ is treated as an unknown parameter and simulated along side the other parameters of the model by Gibbs sampling methods. Such an approach is used by Diebolt and Robert (1993) to estimate the classical mixture model. For the Markov mixture model, Albert and Chib (1993) and McCulloch and Tsay (1994), both in the context of Gaussian time series models, exploit this idea to simulate the posterior distribution.

The first main contribution of this paper is to show that it is possible to simulate the latent data $S_n = (s_1, s_2, \ldots, s_n)$ from the joint distribution

$$s_1, s_2, \ldots, s_n \mid Y_n, \theta, \qquad S_n \in \mathcal{S} = \{1, 2, \ldots, m\}^n, \tag{4}$$

rather than the sequence of full conditional distributions $s_t \mid Y_n, s_j,\ j \neq t$. This new result is extremely significant. Instead of $n$ additional blocks in the Gibbs sampler (the number required if each state is sampled from its full conditional distribution), only one additional block is required. This dramatically improves the convergence of the Markov chain induced by the Gibbs sampling algorithm.

Second, it is shown how the Markov chain Monte Carlo approach can be modified slightly to obtain modal estimates, or alternatively maximum likelihood estimates if diffuse priors are adopted. These modal estimates are obtained using stochastic versions of the EM algorithm such as the stochastic EM (SEM), and the Monte Carlo EM. The performance of these procedures is contrasted with the full Bayesian approach.

Third, the ideas are applied to both Gaussian and non-Gaussian discrete data, and more than two component problems. The examples involve the Poisson distribution, multivariate Gaussian distributions, and autoregressive time series.

### *1.2. Gibbs sampling*

The approach taken in this paper is motivated by the Gibbs sampling algorithm. The idea in Gibbs sampling is to simulate, in turn, the distribution of each parameter vector conditioned on the data and the remaining parameters (the so-called full conditional distribution). This process generates a Markov chain, which under mild conditions converges under the $L^1$ norm to the desired posterior

distribution. The output of the Markov chain, once it has passed its transient stage, is taken as a sample from the posterior distribution for purposes of computing moments and marginal densities. Briefly, the systematic form of the Gibbs sampler for a parameter vector $\psi$ (which may include the missing data), with blocking $(\psi_1,\ldots,\psi_d)$ and full conditional distributions $\{\psi_j \mid (Y_n,\psi_{-j}), 1 \leqslant j \leqslant d\}$, is given by the following steps:

Step 1: Specify starting values $\psi^0 = (\psi_1^0,\ldots,\psi_d^0)$ and set $i = 1$.

Step 2: Simulate

$$
\begin{array}{lll}
\psi_1^{i+1} & \text{from} & \psi_1 \mid Y_n, \psi_2^i, \psi_3^i, \ldots, \psi_d^i, \\
\psi_2^{i+1} & \text{from} & \psi_2 \mid Y_n, \psi_1^{i+1}, \psi_3^i, \ldots, \psi_d^i, \\
\psi_3^{i+1} & \text{from} & \psi_3 \mid Y_n, \psi_1^{i+1}, \psi_2^{i+1}, \psi_4^i \ldots, \psi_d^i, \\
\vdots & & \\
\psi_d^{i+1} & \text{from} & \psi_d \mid Y_n, \psi_1^{i+1}, \ldots, \psi_{d-1}^{i+1}.
\end{array}
$$

Step 3: Set $i = i + 1$, and go to 2.

The above cycle is repeated a large number of times and the simulated values $\{\psi^i, i \geqslant T\}$, where $T$ is a number sufficiently large so that the sampler has converged, is used as a sample from the joint distribution $\psi|Y_n$. Full details are provided in Gelfand and Smith (1990). If the full conditional distributions are readily sampled, this method is quite easy to implement. Note, that the sampler is defined by the choice of $\psi$ and the choice of blocking (i.e., the choice of $\psi_j$). Due to the fact that we include $\{s_t\}$ in $\psi$, there is a considerable proliferation in the number of parameters if each $s_t$ is treated individually. A technique to avoid this problem, by treating all the states as one block and sampling the states from their joint distribution, is developed next.

## 2. Full conditional distributions

### 2.1. Simulation of the states

The key feature of the new Bayesian Markov chain Monte Carlo approach is the simulation of the states (the population index) from the distribution $p(S_n|Y_n, \theta)$, which is the joint posterior mass function of all the states given $\theta$. This simulation might seem to be intractable because the range space is $\mathscr{S}$, the $n$-fold product of the set $\{1,2,\ldots,m\}$. However, it is possible to develop a quite simple expression for the joint distribution that leads to a recursive simulation procedure. At each step, starting with the terminal state, $s_n$, only a single state has to be drawn. To simplify the notation and the discussion it is convenient to adopt the

following conventions:

$$S_t = (s_1, \ldots, s_t), \qquad S^{t+1} = (s_{t+1}, \ldots, s_n),$$

with a similar convention applying to $Y_t$ and $Y^{t+1}$. In words, $S_t$ is the history of the states up to time $t$, and $S^{t+1}$ is the future from $t+1$ to $n$. Now write the joint density (4) in the following manner:

$$p(S_n|Y_n, \theta) = p(s_n|Y_n, \theta) \times \cdots \times p(s_t|Y_n, S^{t+1}, \theta) \times \cdots \times p(s_1|Y_n, S^2, \theta), \tag{5}$$

in which the typical term, excluding the terminal point, is given by

$$p(s_t \mid Y_n, S^{t+1}, \theta). \tag{6}$$

By Bayes theorem,

$$\begin{aligned} p(s_t \mid Y_n, S^{t+1}, \theta) &\propto p(s_t|Y_t, \theta) \times f(Y^{t+1}, S^{t+1} \mid Y_t, s_t, \theta) \\ &\propto p(s_t|Y_t, \theta) \times p(s_{t+1}|s_t, \theta) \times f(Y^{t+1}, S^{t+2} \mid Y_t, s_t, s_{t+1}, \theta) \\ &\propto p(s_t \mid Y_t, \theta) \times p(s_{t+1} \mid s_t, \theta), \end{aligned} \tag{7}$$

since the term $f(Y^{t+1}, S^{t+2} \mid Y_t, s_t, s_{t+1}, \theta)$ is independent of $s_t$. Thus, the required mass function in (6) is the product of two terms, one of which is the mass function of $s_t$, given $(Y_t, \theta)$, and the other is the transition probability of going from $s_t$ to $s_{t+1}$, given $\theta$. The normalizing constant of this mass function is the sum of the numbers obtained in (7) as $s_t$ runs from 1 to $m$.

The rest of the calculation is concerned with determining the first mass function in (7). It can be determined recursively for all $t$ starting with period 1. The objective is to find $p(s_t \mid Y_t, \theta)$ and this is obtained as follows. Assume that the function $p(s_{t-1}|Y_{t-1}, \theta)$ is available. Then, repeat the following steps.

*Prediction step*: Determination of $p(s_t \mid Y_{t-1}, \theta)$. By the law of total probability,

$$p(s_t \mid Y_{t-1}, \theta) = \sum_{k=1}^{m} p(s_t \mid s_{t-1} = k, \theta) \times p(s_{t-1} = k \mid Y_{t-1}, \theta),$$

where the fact that $p(s_t \mid Y_{t-1}, s_{t-1}, \theta) = p(s_t \mid s_{t-1}, \theta)$ has been utilized.

*Update step*: Determination of $p(s_t \mid Y_t, \theta)$. By Bayes theorem, the mass function of the state given information up to time $t$ is now

$$p(s_t \mid Y_t, \theta) \propto p(s_t \mid Y_{t-1}, \theta) \times f(y_t \mid Y_{t-1}, \theta_{s_t}), \tag{8}$$

where the normalizing constant is the sum of all the terms over $s_t$ from 1 to $m$.

These steps can be initialized at $t=1$ by setting $p(s_1 \mid Y_0, \theta)$ to be the stationary distribution of the chain (the left eigenvector corresponding to the eigenvalue of 1); the prediction step is thus not required at the start of these recursions.

Based on these results, the states can be simulated from their joint distribution (5) in a very simple manner. (Note that if the prior on all the parameters is proper, it is not necessary to reject a particular $S_n$ that does not ascribe at least one observation to each population).

First run the *prediction* and *update* steps recursively to compute the mass functions $p(s_t \mid Y_t, \theta)$. [These mass functions are obtained by defining a $n \times m$ storage matrix, say $F$. Given the $t-1$ row $F_{t-1}$, the next row is $F_t$ which is proportional to $(F_{t-1}' P) \odot d_t$, where $d_t$ is a row vector consisting of $f(y_t \mid Y_{t-1}, \theta_{s_t})$ and $\odot$ is the element-by-element multiplication operator.] The last row of $F$ is then used to simulate $s_n$. After $s_n$ is simulated, the remaining states (beginning with $s_{n-1}$) are simulated using the probability mass function that emerges from (7). Note that the calculation of the latter distribution requires the numbers in the $t$th row of $F$, and those in the column of $P$ corresponding to the simulated value of $s_{t+1}$.

## 2.2. Simulation of P

Given the states, it is rather straightforward to determine the full conditional distribution of the unique element of the transition matrix $P$. This is because $P$ becomes independent of $(Y_n, \bigcup_{k=1}^{m} \theta_k)$, given $S_n$. Thus, the full conditional distribution of the transition matrix can be derived without regard to the sampling model.

Suppose the $i$th row of $P$ is denoted by $p_i = (p_{i1}, \ldots, p_{im})'$, and let the prior distribution of $p_i$, independently of $p_j$, $j \neq i$, be a Dirichlet on the $m$-dimensional simplex, i.e.,

$$p_i \sim \mathcal{D}(\alpha_{i1}, \ldots, \alpha_{im}). \tag{9}$$

Then, multiplying the prior by the likelihood function of $P|S_n$ immediately gives the result that the updated distribution is also Dirichlet. In particular,

$$p_i \mid S_n \sim \mathcal{D}(\alpha_{i1} + n_{i1}, \ldots, \alpha_{i1} + n_{im}), \qquad i = 1, \ldots, m, \tag{10}$$

where $n_{ik}$ is the total number of *one-step* transitions from state $i$ to state $k$ in the vector $S_n$. The vector $p_i$ $(1 \leqslant i \leqslant m)$ can now be simulated from (10) by letting

$$p_{i1} = \frac{x_1}{\sum_{j=1}^{m} x_j}, \ldots, p_{im} = \frac{x_m}{\sum_{j=1}^{m} x_j}, \qquad x_j \sim \text{Gamma}(\alpha_{ij} + n_{ij}, 1).$$

## 3. Modal estimates

An important implication of the above result on the simulation of the states is that it can be directly used to compute the maximizer of the likelihood function, or the maximizer of the posterior, through the Monte Carlo EM (MCEM) algorithm proposed by Wei and Tanner (1990). The latter algorithm is a stochastic modification of the original Dempster, Laird, and Rubin (1977) EM algorithm.

Suppose that, given the current guess of the maximizer, it is of interest to evaluate the E-step of the EM algorithm. In the Bayesian formulation that amounts to an evaluation of the integral

$$Q(\theta, \theta^i) = \int_{S_n} \log(\pi(\theta \mid Y_n, S_n))\, \mathrm{d}[S_n \mid Y_n, \theta^i], \tag{11}$$

where the integral is a sum with integrating measure given by the mass function in (5). As this is an intractable calculation, consider the evaluation of the $Q$ function by Monte Carlo. Given the current parameter value $\theta^i$, one can take a large number of draws of $S_n$ as per the approach described above. Suppose the draws are denoted by $S_{n,j}$, $j=1,\ldots,N$. Then the $Q$ function can be approximated via the average

$$\hat{Q}(\theta, \theta^i) = N^{-1} \sum_{j=1}^{N} \log(\pi(\theta \mid Y_n, S_{n,j})). \tag{12}$$

In the $M$-step, the $\hat{Q}$ function can be maximized over $\theta$ to obtain the new parameter $\theta^{i+1}$. The algorithm can be terminated once the difference $\|\theta^{i+1} - \theta^i\|$ is negligible. In producing the iterate sequence $\{\theta^1, \theta^2, \ldots, \theta^i \ldots\}$ via the above strategy, it is best to begin with a small value of $N$ and then let the number of replications of $S_n$ increase as one moves closer to the maximizer.

This procedure provides a straightforward device to locate the modal estimates due to the fact that the $\hat{Q}$ function is additive in the respective parameters. For example, to obtain the updated estimate of $\{p_{kl}\}$ under the Dirichlet prior (9), each row can be treated separately of all the other rows and the $\theta_k$'s. From $\sum_j \log(\pi(p_k \mid Y_n, S_{n,j}))$, which is proportional to

$$\sum_{j=1}^{N} \left\{ \sum_{l=1}^{m-1} (n_{kl,j} + \alpha_{kl} - 1) \log(p_{kl}) + (n_{km,j} + \alpha_{km} - 1) \log(1 - p_{k1} - \cdots - p_{km-1}) \right\}.$$

where $n_{kl,j}$ is the number of transitions from state $k$ to state $l$ in the simulation $S_{n,j}$, the next iterate is given by

$$\hat{p}_{kl} = \frac{\sum_{j=1}^{N} (\alpha_{kl} + n_{kl,j})}{\sum_{j=1}^{N} \left( \sum_{l=1}^{m} (\alpha_{kl} + n_{kl,j}) \right)}. \tag{13}$$

A modification of the MCEM algorithm leads to another version of the EM algorithm (Celeux and Diebolt, 1985). Suppose that instead of taking $N$ draws of $S_n$ for each value of $\theta$, only one draw is made. As before, the updated or new value of $\theta$ is found by maximizing the posterior density of $\theta$ given $(Y_n, S_n)$. However, unlike the MCEM which generates a deterministic sequence of parameter updates, the iterates in this algorithm follows a aperiodic, irreducible Markov chain.

## 4. Examples

### 4.1. Poisson fetal data

We begin by considering the fetal movement data analyzed in Leroux and Puterman (1992). The data consists of number of movements by a fetal lamb (observed by ultra sound) in 240 consecutive five-second intervals. The number of counts is modeled as a Poisson process in which the unknown rate parameter, $\lambda$, can vary from one interval to the next according to a Markov chain described by (1). In particular, given the state at time $t$, the count (the number of movements) during interval $t$ is given by

$$f(y_t|\lambda_k) = \frac{\lambda_k^{y_t} e^{-\lambda_k}}{y_t!}, \qquad t = 1, 2, \ldots, 240, \quad k = 1, \ldots, m. \tag{14}$$

The data used in the study is given in Fig. 1.

Two models are fit to this data set, one with two components and the second with three components. Note that it is convenient to take independent Gamma priors on $\lambda_k$ due to the fact that such a distribution is conjugate to the Poisson likelihood. Then, under the assumption that $\lambda_k \sim \mathcal{G}(a_k, b_k)$, the full conditional distribution of $\lambda_k$ is

$$\lambda_k | Y_n, S_n, P \sim \mathcal{G}\left( a_k + \sum_{t=1}^{n} y_t I[s_t = k], b_k + N_k \right), \qquad k = 1, \ldots, m, \tag{15}$$

where $I[s_t = k]$ is the indicator function that takes the value 1 if $s_t = k$ and 0 otherwise, and $N_k$ is the total number of observations from the $k$th population. Thus, given $S_n$, all the $\lambda_k$'s are simulated from Gamma distributions. The MCMC

algorithm is then based on iterating between the simulation of $S_n$ from (7), $P$ from (10), and $\lambda_k$ from (15).

For the MCEM algorithm, the updated iterate of $\lambda_k$ is also obtained quite easily. Due to the fact $\sum_{j=1}^{N} \log(\pi(\lambda_k, Y_n, S_{n,j})$ is proportional to

$$\sum_{j=1}^{N} (U_{k,j} + a_k - 1) \log(\lambda_k) - \sum_{j=1}^{N} (b_k + N_{k,j})\lambda_k, \tag{16}$$

where $U_{k,j} = \sum_{t=1}^{n} y_t I[s_{t,j} = k]$ is the sum of the $y$ values in state $k$ in the $j$th draw of $S_n$, the next iterate is obtained as

$$\hat{\lambda}_k = \sum_{j=1}^{N} (U_{k,j} + a_k - 1) \Big/ \sum_{j=1}^{N} (b_k + N_{k,j}). \tag{17}$$

The update value of $P$ is obtained from (13). Finally, the estimates for the SEM algorithm are obtained by dropping the summation over $j$ in (16) and (17).

Consider the case of two populations. Suppose the prior parameters of $\lambda_k$ are given by $(a_1, b_1) = (1, 2)$ and $(a_2, b_2) = (2, 1)$, which specifies the belief that the first population has the lower mean. Also suppose that in the Dirichlet prior on $P$, $(\alpha_{11}, \alpha_{12}) = (3, 1)$ and $(\alpha_{21}, \alpha_{22}) = (0.5, 0.5)$. The implied prior moments are given in Table 1. After initializing the iterations with values chosen from the prior distribution, we run the MCMC algorithm in one long stream till we have approximate convergence. Then, the first 200 sampled values are discarded and the next 6,000 are used to summarize the posterior distributions. The results are reported in Table 1, Fig. 1 (the posterior probability that $s_t = k$) and Fig. 2 (posterior densities of parameters).

Note that here and in examples later, the design of the Gibbs sampler algorithm (the number of iterations discarded and the choice of Gibbs sample size) is governed by inspecting the autocorrelation function of the sampled draws and the numerical standard errors of the estimates. In most cases, the autocorrelations in the sampled values decayed to zero by about the 10th lag. The numerical standard errors being small are not reported. In addition, the box plots reproduce the minimum and maximum values, and the 25th, 50th and 75th percentiles.

Table 1
ML and MCMC estimates for Poisson two-population Markov model

| | | Prior | | Posterior | | | |
|---|---|---|---|---|---|---|---|
| | MLE | Mean | Std. dev. | Mean | Std. dev. | Lower | Upper |
| $\lambda_1$ | 0.256 | 0.500 | 0.500 | 0.219 | 0.050 | 0.115 | 0.319 |
| $\lambda_2$ | 3.101 | 2.000 | 1.414 | 2.291 | 0.776 | 1.085 | 4.000 |
| $p_{11}$ | 0.984 | 0.750 | 0.194 | 0.967 | 0.025 | 0.898 | 0.995 |
| $p_{22}$ | 0.692 | 0.500 | 0.354 | 0.664 | 0.158 | 0.322 | 0.924 |

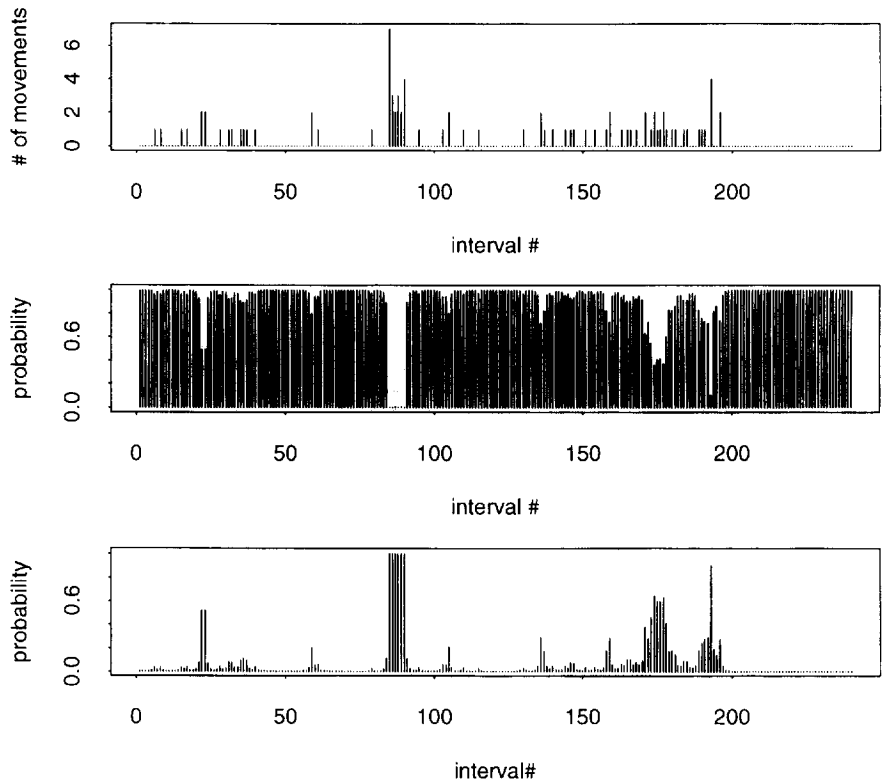


Fig. 1. Two-population Poisson mixture – Top: data $Y_n$, middle: $\Pr(s_t = 1 \mid Y_n)$, bottom: $\Pr(s_t = 2 \mid Y_n)$.

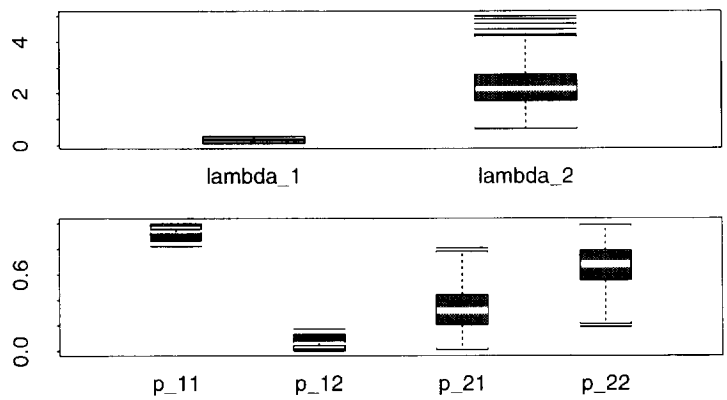


Fig. 2. Posterior box plots in two-population Poisson mixture – Top $\lambda$, bottom: $P$.

Table 2
Final five iterates in combined SEM–MCEM algorithm for Poisson two-population Markov model; in computing (12), $N = 1000$

| | $i = 101$ | $i = 102$ | $i = 103$ | $i = 104$ | $i = 105$ |
|---|---|---|---|---|---|
| $\lambda_1$ | 0.258 | 0.259 | 0.258 | 0.259 | 0.259 |
| $\lambda_2$ | 2.933 | 2.948 | 2.960 | 2.953 | 2.955 |
| $p_{11}$ | 0.989 | 0.990 | 0.990 | 0.990 | 0.990 |
| $p_{22}$ | 0.715 | 0.720 | 0.720 | 0.721 | 0.720 |

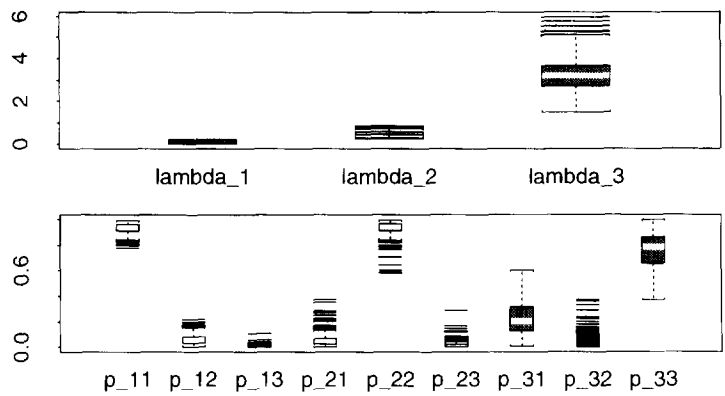


Fig. 3. Posterior box plots in three-population Poisson mixture – Top: $\lambda$, bottom: $P$.

The results indicate that observations in excess of 2 are classified as belonging to the high mean population. For two observations that are exactly 2 ($y_{23}, y_{24}$), either population is about as likely. Parameter estimates are precise and the model appears to be a good fit to the data. Note that the maximum likelihood estimates are taken from Leroux and Puterman (1992). They are similar to the Bayes point estimates and seem to differ mainly in the case of $\lambda_2$.

Point estimates are also obtained by the Monte Carlo EM. We decided to combine the SEM and MCEM algorithms in the following manner. First, during the burn-in period, the SEM algorithm was employed, and then after the values appeared to settle down, a switch was made to the MCEM algorithm. Specifically, the SEM algorithm was used for the first 100 iterations, then the MCEM for the last five iterations. The $Q$ function in the MCEM steps was approximated using 1000 draws. The evolution of the iterate sequence in those last five iterations is contained in Table 2. It appears that the algorithm has converged to the posterior mode. An average of the estimates, or the final iterate values, can be used as the output of the Monte Carlo algorithm.

Results are also obtained for a three-population mixture. The resulting posterior densities are summarized in Fig. 3 while the posterior probabilities of the

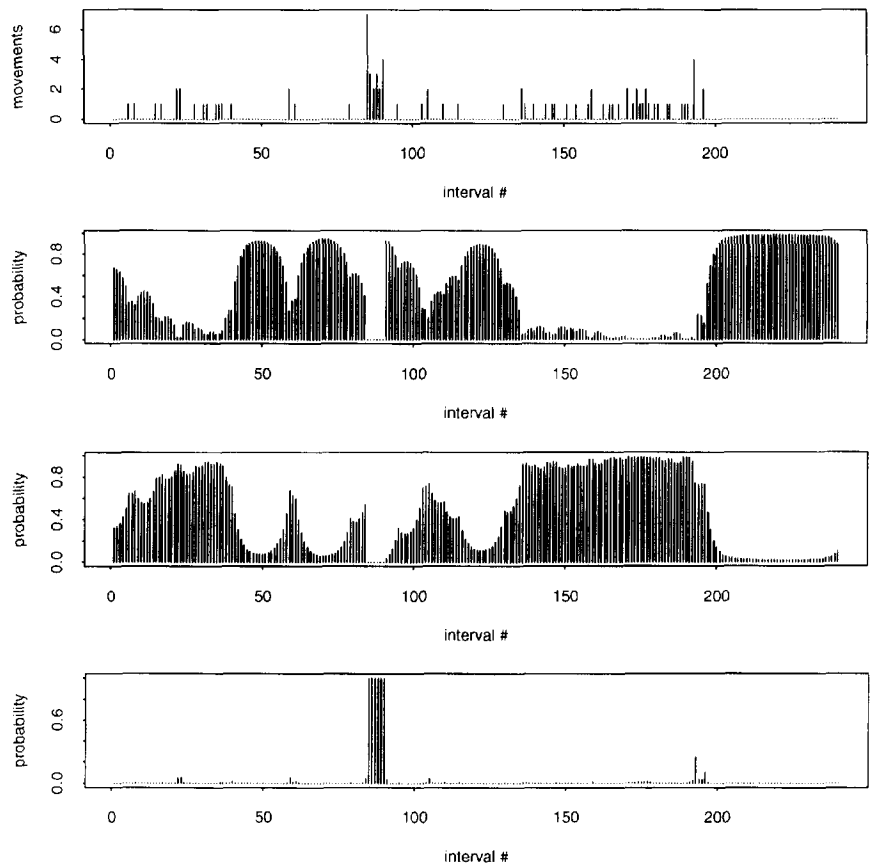


Fig. 4. Three-population Poisson mixture Top: data $Y_n$, second: $\Pr(s_t = 1 \mid Y_n)$, third: $\Pr(s_t = 2 \mid Y_n)$, bottom: $\Pr(s_t = 3 \mid Y_n)$.

populations are presented in Fig. 4. In this case only six observations appear to arise from the third population while the rest of the data is evenly distributed among the first two populations. We have presented the ML estimates and the full Bayes results in Table 3 but have suppressed the MCEM estimates to conserve space. They are all in close agreement with the full Bayes results.

### 4.2. *Autoregressive GNP data*

Now consider the data set on quarterly U.S. real GNP that has been analyzed earlier by Hamilton (1989) and Albert and Chib (1993) using a two-population model with a fourth-order stationary autoregression. The variable of interest is the percentage change (multiplied by 100) in the postwar real GNP for the period

Table 3
ML and MCMC estimates for Poisson three-population Markov model.

| | | Prior | | Posterior | | | |
|---|---|---|---|---|---|---|---|
| | MLE | Mean | Std. dev. | Mean | Std. dev. | Lower | Upper |
| $\lambda_1$ | 0.045 | 0.500 | 0.500 | 0.063 | 0.043 | 0.004 | 0.167 |
| $\lambda_2$ | 0.509 | 2.000 | 1.414 | 0.510 | 0.099 | 0.345 | 0.743 |
| $\lambda_3$ | 3.414 | 3.000 | 1.732 | 3.267 | 0.727 | 1.972 | 4.814 |
| $p_{11}$ | 0.947 | 0.732 | 0.196 | 0.933 | 0.038 | 0.842 | 0.986 |
| $p_{12}$ | 0.043 | 0.244 | 0.190 | 0.062 | 0.038 | 0.010 | 0.156 |
| $p_{21}$ | 0.042 | 0.333 | 0.298 | 0.051 | 0.041 | 0.002 | 0.158 |
| $p_{22}$ | 0.958 | 0.333 | 0.298 | 0.934 | 0.043 | 0.824 | 0.986 |
| $p_{31}$ | 0.184 | 0.244 | 0.190 | 0.228 | 0.131 | 0.034 | 0.520 |
| $p_{33}$ | 0.816 | 0.732 | 0.196 | 0.757 | 0.132 | 0.466 | 0.956 |

1951.2 to 1984.4. The objective is to fit autoregressive models in which the intercept can be drawn from one of four populations but all other parameters are constant across the populations. This is flexible structure that can capture Markov shifts in the level of the process. McCulloch and Tsay (1994) consider a similar model but they restrict attention to two populations and use a Gibbs sampler in which the population indices are not drawn jointly.

Specifically, let the conditional density of $y_t$, given $Y_{t-1}$ and $s_{t-1}$, be given by

$$f(y_t \mid Y_{t-1}, s_{t-1}, \alpha, \gamma) = \sum_{k=1}^{4} p(s_t = k \mid s_{t-1})\, f(y_t \mid Y_{t-1}, \alpha_k, \gamma, \sigma^2),$$

where $\alpha = (\alpha_1, \ldots, \alpha_4)$, $\gamma = (\gamma_1, \ldots, \gamma_p)$ and

$$f(y_t \mid Y_{t-1}, \alpha_k, \gamma, \sigma^2) = \phi(y_t \mid \alpha_k + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p}, \sigma^2). \tag{18}$$

Hence, at time $t$, the data is drawn from one of four Gaussian populations with (respective) conditional mean $\mathrm{E}(y_t \mid Y_{t-1}, \alpha, \gamma, \sigma^2) = \alpha_k + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p}$ and conditional variance that is constant across the populations. Note that this specification differs from that used by Hamilton (1989) and Albert and Chib (1993), where in the context of two populations, the conditional mean of $y_t$ depends on realizations of the states at previous time points.

The MCMC algorithm is again quite easily implemented, provided the analysis is conditioned on the first $p$ observations. As before, the states and the transition probability matrix are simulated according to (8) and (11). Then, given $S_n$, the other parameters, namely $(\alpha, \beta, \sigma^2)$, are simulated from distributions that are easily derived based on results presented in Chib (1993). In particular, under the prior $\alpha_k \sim N(\alpha_{0k}, A_{0k}^{-1})$, $\alpha_k$ is simulated from the distribution

$$\alpha_k \mid Y_n, S_n, \gamma, \sigma^2 \sim N\left(V_k(A_0\alpha_{0k} + \sigma^{-2}\sum_{t=1}^{n} z_t\, I[s_t = k]),\ V_k\right),$$

where $z_t = y_t - \gamma_1 y_{t-1} - \cdots - \gamma_p y_{t-p}$ and $V_k = (A_{0k} + N_k \sigma^{-2})^{-1}$. Next, under a $\mathcal{N}_p(\gamma_0, \Gamma_0^{-1})$ prior on $\gamma$ restricted to the stationary region, $\gamma$ is simulated from the distribution

$$\gamma \mid Y_n, S_n, \alpha, \sigma^2 \propto \mathcal{N}_p\left(V(\Gamma_0\gamma_0 + \sigma^{-2}\sum_{t=p+1}^{n} x_t(y_t - \alpha_t)),\ V\right),$$

where $x_t = (y_{t-1}, \ldots, y_{t-p})'$, $\alpha_t = \alpha_k$, when $s_t = k$, and $V = (\Gamma_0 + \sigma^{-2}\sum_{t=p-1}^{n} x_t x_t')^{-1}$. A drawing from this distribution is accepted only if all the roots of the polynomial $1 - \gamma_1 L - \cdots - \gamma_p L^p$ lie outside the unit circle. Finally, under the inverse-gamma prior $\mathcal{IG}(\nu_0/2, \delta_0/2)$, $\sigma^2$ is simulated from

$$\mathcal{IG}\left(\frac{\nu_0 + n}{2}, \frac{\delta_0 + \sum_{t=1}^{n}(y_t - \alpha_t - x_t'\beta)^2}{2}\right).$$

These results are applied to the GNP data set for different values of $p$ under weak priors on the parameters. For brevity, consider the case of a fourth-order autoregression ($p=4$). The Gibbs sampler is run for 7,000 iterations and the last 6,000 draws are used for purposes of summarizing the posterior distribution. The results on $(\alpha, \gamma, \sigma^2)$ are presented in Table 4 while those on $P$ are in Table 5.

The posterior moments of $\alpha$ relative to the prior moments appears to provide support for more than two populations. In addition, it is noted that the marginal posterior distributions of $\gamma_3$ and $\gamma_4$ are quite concentrated around 0. This suggests that an AR(2) specification with four populations is a parsimonious description for the data.

From the posterior distribution of the elements of the transition matrix it may be noted that the data is not informative about some elements of the matrix but that there is considerable evidence for switching between the populations.

Table 4
MCMC estimates of $(\alpha, \gamma, \sigma^2)$ in the AR(4) four-population Markov model

| | Prior | | Posterior | | | |
|---|---|---|---|---|---|---|
| | Mean | Std. dev. | Mean | Std. dev. | Lower | Upper |
| $\alpha_1$ | 0.000 | 1.414 | −0.153 | 0.185 | −0.522 | 0.198 |
| $\alpha_2$ | 0.400 | 1.414 | 0.344 | 0.249 | −0.133 | 0.857 |
| $\alpha_3$ | 1.000 | 1.414 | 0.800 | 0.264 | 0.283 | 1.342 |
| $\alpha_4$ | 1.500 | 1.414 | 1.258 | 0.348 | 0.571 | 1.954 |
| $\gamma_1$ | 0.000 | 2.000 | 0.398 | 0.090 | 0.219 | 0.570 |
| $\gamma_2$ | 0.000 | 2.000 | 0.206 | 0.092 | 0.024 | 0.382 |
| $\gamma_3$ | 0.000 | 2.000 | −0.067 | 0.095 | −0.258 | 0.119 |
| $\gamma_4$ | 0.000 | 2.000 | 0.003 | 0.085 | −0.166 | 0.168 |
| $\sigma^2$ | 1.333 | 0.943 | 0.841 | 0.134 | 0.602 | 1.129 |

Table 5
MCMC estimates of $P$ in AR(4) four-population Markov model

| | Prior | | Posterior | | | |
|---|---|---|---|---|---|---|
| | Mean | Std. dev. | Mean | Std. dev. | Lower | Upper |
| $p_{11}$ | 0.143 | 0.124 | 0.379 | 0.184 | 0.043 | 0.710 |
| $p_{12}$ | 0.286 | 0.160 | 0.295 | 0.171 | 0.042 | 0.681 |
| $p_{13}$ | 0.286 | 0.160 | 0.196 | 0.119 | 0.028 | 0.484 |
| $p_{21}$ | 0.143 | 0.124 | 0.287 | 0.178 | 0.017 | 0.669 |
| $p_{22}$ | 0.286 | 0.160 | 0.312 | 0.163 | 0.061 | 0.668 |
| $p_{23}$ | 0.286 | 0.160 | 0.231 | 0.145 | 0.029 | 0.575 |
| $p_{31}$ | 0.200 | 0.163 | 0.424 | 0.209 | 0.034 | 0.790 |
| $p_{32}$ | 0.200 | 0.163 | 0.193 | 0.167 | 0.005 | 0.624 |
| $p_{33}$ | 0.400 | 0.200 | 0.283 | 0.145 | 0.056 | 0.603 |
| $p_{41}$ | 0.200 | 0.163 | 0.341 | 0.204 | 0.017 | 0.750 |
| $p_{42}$ | 0.200 | 0.163 | 0.218 | 0.171 | 0.007 | 0.625 |
| $p_{43}$ | 0.200 | 0.163 | 0.172 | 0.150 | 0.005 | 0.551 |

### 4.3. *Bivariate Gaussian data*

Next the model in the previous two sections is generalized to a three-component mixture of multivariate normal distributions. In particular, consider a bivariate normal distribution and let

$$f(y_t \mid s_{t-1}, \mu, \Omega, P) = \sum_{k=1}^{3} p(s_t = k \mid s_{t-1})\, \phi_2(y_t \mid \mu_k, \Omega_k),$$

where $\phi_2$ is the density function of a bivariate normal distribution, $y_t$ is a 2 vector, and $\Omega_k$ is a $2 \times 2$ positive definite matrix. Titterington, Smith, and Makov (1985) contain a treatment of the work done on such models in the classical mixture set-up.

We generate 300 observations from this model under the following specifications: $\mu_1 = (1, 2)'$, $\mu_2 = (3, 0)'$, $\mu_3 = (5, 4)'$, $\text{vec}(\Omega_1) = (1.5, 0.5, 0.5, 1)'$, $\text{vec}(\Omega_2) = (2.0, 0.6, 0.6, 1.0)'$, $\text{vec}(\Omega_3) = (1.5, -0.5, -0.5, 2.0)'$. For the transition probability matrix, values are specified that imply some persistence in the choice of the populations:

$$P = \begin{pmatrix} 0.3 & 0.3 & 0.4 \\ 0.4 & 0.5 & 0.1 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}.$$

It should be noted that each of the components of $y_t$ satisfy different order relations in the mean and variance. The data used in the study is reproduced in Fig. 5.

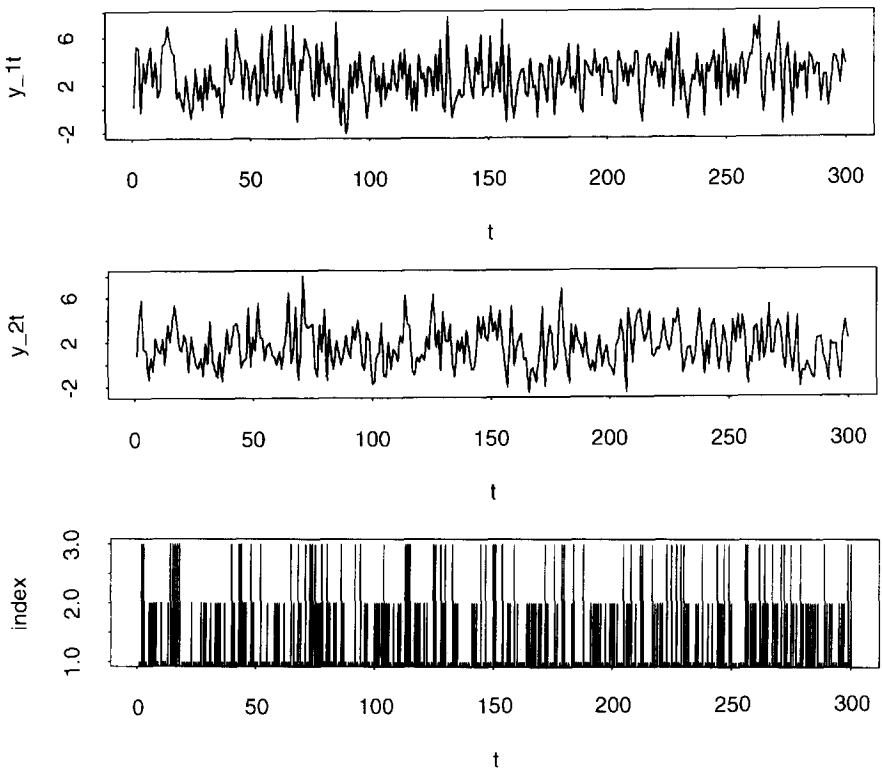


Fig. 5. Bivariate three-population Gaussian misture – Top: $y_{1t}$, middle: $y_{2t}$, bottom: true $s_t$.

The full conditional distribution of $\mu_k$ is easily derived. Under a bivariate normal prior given by $\mu_k \sim N_2(\mu_{0k}, A_{0k}^{-1})$, the updated distribution for $\mu_k$ is

$$N_2\left((A_{0k}+N_k\Omega^{-1})^{-1}\left(A_0\mu_{0k}+\Omega^{-1}\sum_{t=1}^{n}y_t I[s_t=k]\right),(A_{0k}+N_k\Omega^{-1})^{-1}\right). \tag{19}$$

Also, if $\Omega_k^{-1}$ is given a Wishart prior, say $W(v_{0k}, D_{0k})$, $\Omega^{-1}$ is simulated from

$$W\left(v_{0k}+N_k,\left(D_{0k}^{-1}+\sum_{t=1}^{n}(y_t I[s_t=k]-\mu_k)(y_t I[s_t=k]-\mu_k)'\right)^{-1}\right).$$

The SEM and MCEM updates are the sample mean and sample covariance if the prior is fully diffuse. These are easy to modify for the above priors. The full Bayes results for this model are obtained under a fairly diffuse specification. The results in Table 6 (relating to $\mu_k$), are not sensitive to the specification of

Table 6
MCMC estimates for $\mu_k$ in bivariate Gaussian three-population Markov model

| | | | Prior | | Posterior | |
|---|---|---|---|---|---|---|
| | True | SEM | Mean | Std. dev. | Mean | Std. dev. |
| $\mu_{11}$ | 1.0 | 1.718 | 0.500 | 2.000 | 0.929 | 0.161 |
| $\mu_{12}$ | 2.0 | 2.656 | 1.000 | 2.000 | 1.958 | 0.150 |
| $\mu_{21}$ | 3.0 | 2.286 | 3.500 | 2.000 | 2.922 | 0.011 |
| $\mu_{22}$ | 0.0 | 0.376 | 1.000 | 2.000 | −0.274 | 0.006 |
| $\mu_{31}$ | 5.0 | 5.070 | 4.000 | 2.000 | 4.684 | 0.012 |
| $\mu_{32}$ | 4.0 | 3.188 | 2.000 | 2.000 | 3.383 | 0.016 |

the prior. It should be noted that the Bayes estimates are more accurate than the SEM estimates.

The posterior probabilities in Fig. 6 are able to correctly uncover the membership for most of the observations. An interesting feature is observed in the simulation. Since there are no order relations between the population parameters and the numbering of the states is arbitrary, we find that $s_t = 2$ corresponds to the third population as defined above. The same feature is observed with the SEM results. In summary, we find that the Bayes results are very accurate, and they show clearly that even in this quite difficult problem, the MCMC approach developed in this paper is able to learn about the component densities and the component parameters.

## 5. Concluding remarks

This paper has developed a new Markov chain Monte Carlo method to estimate an important class of finite mixture distributions. For models described by (1)–(3), a approach is developed that relies, first, on data augmentation and, second, on the simulation of the unobserved population index from its joint distribution given the data and the remaining parameters. The paper also shows the value of stochastic versions of the EM algorithm in finding modal estimates and includes comparisons with results obtained from the full Bayesian approach. The ideas are illustrated with Poisson data, bivariate Gaussian data, and an autoregressive time series model applied to U.S. GNP data. In all the examples, the methods perform extremely well.

In future work, it will be of interest to consider the issue of model selection in this setting. Recently, Carlin and Chib (1995) have developed simulation based approaches to model selection in regression models and classical finite mixture models. Similar results on the model selection problem in Markov mixture models will be presented elsewhere.

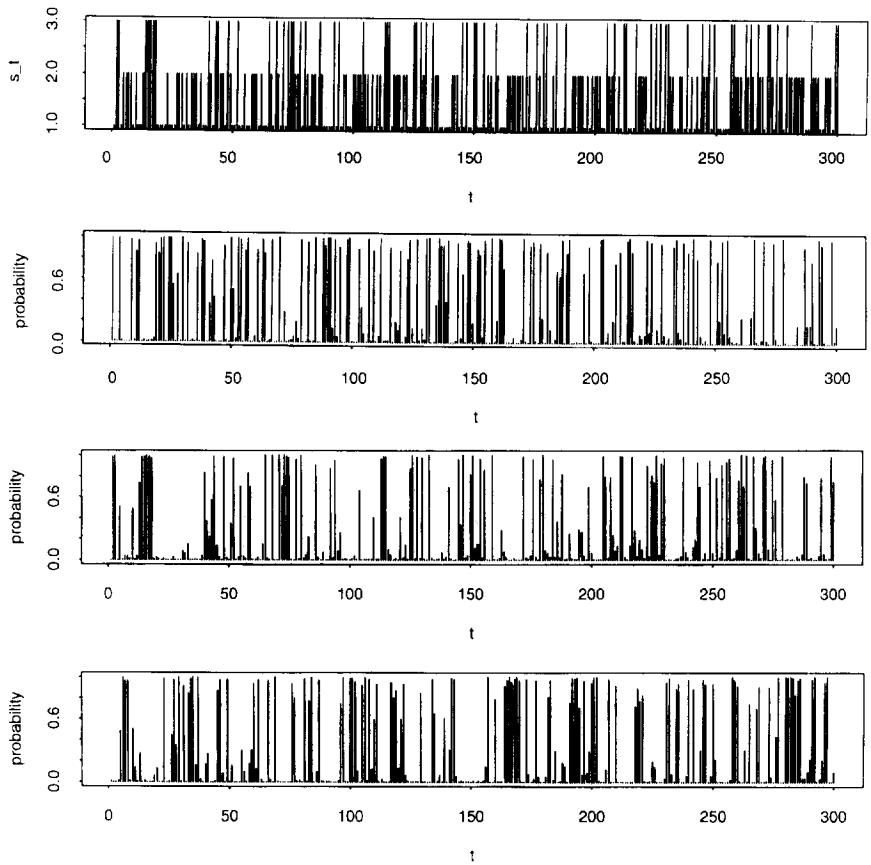


Fig. 6. Three-population bivariate Gaussian mixture – Top: true $s_t$, second: $\Pr(s_t = 1 \mid Y_n)$, third: $\Pr(s_t = 2 \mid Y_n)$, bottom: $\Pr(s_t = 3 \mid Y_n)$.

## References

Albert, J. and S. Chib, 1993, Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts, Journal of Business and Economic Statistics 11, 1–15.

Baum, L.E., T. Petrie, G. Soules, and N. Weiss, 1970, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Annals of Mathematical Statistics 41, 164–171.

Carlin, B. and S. Chib, 1995, Bayesian model choice via Markov chain Monte Carlo, Journal of the Royal Statistical Society B 57, 473–484.

Celeux, G. and J. Diebolt, 1985, The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, Computational Statistics Quarterly 2, 73–82.

Chib, S., 1993, Bayes regression with autoregressive errors: A Gibbs sampling approach, Journal of Econometrics 58, 275–294.

Diebolt, J. and C.P. Robert, 1994, Estimation of finite mixture distributions through Bayesian sampling, Journal of the Royal Statistical Society B 56, 363–375.

Dempster, A.P., N. Laird, and D.B. Rubin, 1977, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society B 39, 1–38.

Everitt, B. and D. Hand, 1981, Finite mixture distributions (Chapman and Hall, London).

Gelfand, A.E. and A.F.M. Smith, 1990, Sampling-based approaches to calculating marginal densities, Journal of the American Statistical Association 85, 398–409.

Goldfeld, S.M. and R.E. Quandt, 1973, A Markov model for switching regressions, Journal of Econometrics 1, 3–16.

Hamilton, J.D., 1989, A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica 57, 357–384.

Hamilton, J.D., 1994, Time series analysis (Princeton University Press, Princeton, NJ).

Leroux, B.G., 1992, Maximum-likelihood estimation for hidden Markov models, Stochastic Processes and their Applications 40, 127–143.

Leroux, B.G. and M.L. Puterman, 1992, Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models, Biometrics 48, 545–558.

McCulloch, R. and R. Tsay, 1994, Statistical analysis of macroeconomic time series via Markov switching models, Journal of Time Series Analysis 15, 523–539.

Titterington, D., A.F.M. Smith, and U. Makov, 1985, Statistical analysis of finite mixture distributions (Wiley, New York, NY).

Wei, G.C.G. and M.A. Tanner, 1990, A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm, Journal of the American Statistical Association 85, 699–704.