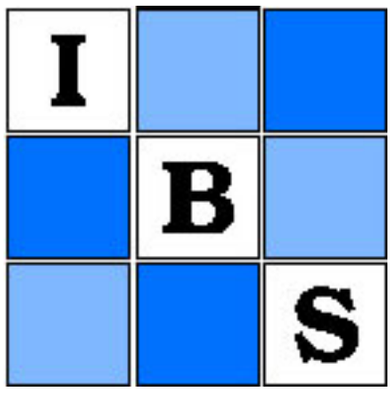

BIOMETRICS 48, 545–558
June 1992

Maximum-Penalized-Likelihood Estimation for Independent and Markov-Dependent Mixture Models

Brian G. Leroux*

Health and Welfare Canada, Environmental Health Centre, Tunney's Pasture, Ottawa, Ontario K1A 0L2, Canada

and

Martin L. Puterman

The Faculty of Commerce, The University of British Columbia, 2053 Main Mall, Vancouver, British Columbia V6T 1Z2, Canada

SUMMARY

This paper concerns the use and implementation of maximum-penalized-likelihood procedures for choosing the number of mixing components and estimating the parameters in independent and Markov-dependent mixture models. Computation of the estimates is achieved via algorithms for the automatic generation of starting values for the EM algorithm. Computation of the information matrix is also discussed. Poisson mixture models are applied to a sequence of counts of movements by a fetal lamb *in utero* obtained by ultrasound. The resulting estimates are seen to provide plausible mechanisms for the physiological process.

1. Introduction

The analysis of count data that are overdispersed relative to the Poisson distribution (i.e., variance > mean) has received considerable recent attention. Such data might arise in a clinical study in which overdispersion is caused by unexplained or random subject effects. Alternatively, we might observe a time series of counts in which temporal patterns in the data suggest that a Poisson model and its implied randomness are inappropriate. This paper is motivated by analysis of a time series of overdispersed count data generated in a study of central nervous system development in fetal lambs. Our data set consists of observed movement counts in 240 consecutive 5-second intervals obtained from a single animal. In analysing these data, we focus on the use of Poisson mixture models assuming independent observations and also Markov-dependent mixture models (or hidden Markov models). These models assume that the counts follow independent Poisson distributions conditional on the rates, which are generated from a mixing distribution either independently or with Markov dependence. We believe finite mixture models are particularly attractive because they provide plausible explanations for variation in the data.

This paper will emphasize the following issues concerning estimation, inference, and application of mixture models: (i) choosing the number of model components; (ii) applying the EM algorithm to obtain parameter estimates; (iii) generating sufficiently many starting values to identify a global maximum of the likelihood; (iv) avoiding numerical instability

* *Current address:* Department of Biostatistics, SC-32, University of Washington, Seattle, Washington 98195, U.S.A.

Key words: EM algorithm; Hidden Markov model; Markov chain; Maximum likelihood; Nonparametric mixture; Overdispersion; Poisson distribution.

of the algorithm; (v) computing the observed information matrix. Methods and results will be illustrated in the context of our application, which is introduced in the next section.

2. Fetal Movement Data

In a study of breathing and body movements in fetal lambs designed to examine the possible changes in the amount and pattern of fetal activity during the last two-thirds of gestation, the numbers of movements by a fetal lamb observed through ultrasound were recorded. Changes in activity may be due to physical factors such as reduction in amniotic fluid volume and empty space within the uterus, or the development of the central nervous system (Wittmann, Rurak, and Taylor, 1984). In this paper, we analyse one particular sequence of counts of the numbers of movements in 240 consecutive 5-second intervals. The analyses discussed herein provide a description of the pattern of fetal activity at a fixed point in gestation; the study of changes in activity would involve the comparison of two or more sequences of counts.

The data are displayed in Figure 1 (the height of each bar represents the number of movements in a 5-second interval). The counts are listed in Table 1 and summarized here:

Number of movements:	0	1	2	3	4	5	6	7
Number of intervals:	182	41	12	2	2	0	0	1

It is reasonable to assume that, in short time intervals, movements occur randomly in time as in a Poisson process. However, it appears from Figure 1 that the rate at which movements occur changes over time, and the index-of-dispersion test (Snedecor and Cochran, 1980, p. 198) indicates that the distribution of the counts is overdispersed relative to the Poisson distribution ($239s^2/\bar{y} = 438.7$, $P < .001$).

The Poisson mixture model explains overdispersion by allowing the unobserved Poisson

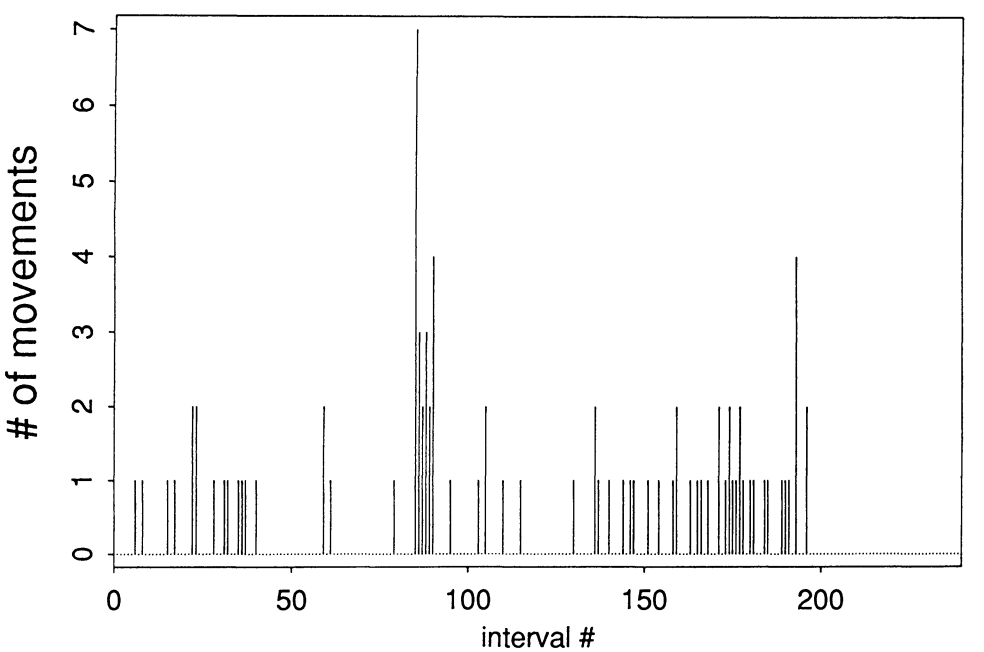

Figure 1. Numbers of movements by a fetal lamb observed through ultrasound. The height of each bar represents the number of movements in one of 240 consecutive 5-second intervals. The counts are listed in Table 1.

Table 1
Numbers of movements by a fetal lamb in 240 consecutive 5-second intervals[a]

0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	2	2	0	0	0	0	1	0	0	1	1	0	0	1	1	1	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	7	3	2	3	2	4	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	1	0	0	1	0	0	0	1	0	1	1	0	0	0	1	0	0	1	0	0	0	1	2	0
0	0	1	0	1	1	0	1	0	0	2	0	1	2	1	1	2	1	0	1	1	0	0	1	1	0	0	0	1	1	1	0	4	0	0	2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

[a] To be read from left to right within each row.

rate parameter to vary independently from interval to interval. For instance, we might postulate different physical states, such as a relaxed state with a background rate of movement that the fetus occupies most of the time and an excited state with movement occurring at a much higher rate, perhaps being triggered by external stimuli. Models with several states could approximate a smoothly varying movement rate.

The Markov-dependent mixture model can explain both overdispersion and autocorrelation in the counts by allowing Markov dependence in the sequence of rate parameters. Thus, the state occupied in any one interval might depend strongly on the state occupied in the previous interval.

3. Mixture Models

3.1 *Mixture Distributions for Independent Observations*

Let $y_1, \ldots, y_n$ represent a sequence of observations. Initially assume that the observations are independent with density $p(y)$ given by

$$p(y) = \sum_{j=1}^{m} \alpha_j f(y; \lambda_j), \tag{1}$$

where $\alpha_j \geq 0$, $\sum_j \alpha_j = 1$, and $\{f(y; \lambda): \lambda \in \Lambda\}$ is a parametric family of densities. Such a density is called a finite mixture.

We will model the fetal movement counts introduced in the previous section by a mixture of Poissons, for which $f(y; \lambda) = \lambda^y e^{-\lambda}/y!$ and (1) becomes

$$p(y) = \sum_{j=1}^{m} \alpha_j \lambda_j^y e^{-\lambda_j}/y!.$$

We posit the following model for generation of the data. At each time epoch i, an (unobserved) value $\lambda(i)$ is drawn from the set $\{\lambda_1, \ldots, \lambda_m\}$ with probabilities $\alpha_1, \ldots, \alpha_m$. Conditional on $\lambda(i)$, the count y_i is generated from a probability distribution with density $f(y; \lambda(i))$.

The density (1) is a special case of a general mixture density

$$p_F(y) = \int_\Lambda f(y; \lambda)\, dF(\lambda), \tag{2}$$

where F is an arbitrary distribution function. Other examples of this general model include mixtures in which F itself belongs to a parametric family, such as a gamma-mixture of Poissons, which is a negative-binomial density (probability mass function):

$$p_F(y) = \frac{\alpha(\alpha + 1) \cdots (\alpha + y - 1)}{y!} \left(\frac{\mu}{\alpha + \mu}\right)^y \left(\frac{\alpha}{\alpha + \mu}\right)^\alpha, \quad y = 0, 1, \ldots,$$

where the mean is μ and the variance is $\mu(1 + \mu/\alpha)$.

A series of papers (Simar, 1976; Laird, 1978; Lindsay, 1983) has shown that the likelihood function based on a random sample from the mixture density (2) is maximized over all mixing distributions F by a distribution function $\hat{F}$ with a finite number of components. Further, Lindsay showed that the number of components can be bounded above by the number of distinct observations in the sample.

3.2 *Mixture Models with Markov Dependency*

Markov-dependent mixture models (or hidden Markov models) generalize mixture distributions by introducing serial correlation through the sequence of unobserved parameter values $\lambda(i)$. In particular, this sequence is assumed to follow a Markov chain with stationary transition probabilities. Formally, let $\{X_i\}$ be a Markov chain with states denoted $1, \ldots, m$ and stationary transition probabilities. Then we assume that the Y_i are conditionally independent given the X_i, with conditional densities $f(y; \lambda_{X_i})$. Note that this definition extends to nonfinite mixtures, but we will not consider such models here. To fit such a model, the transition probabilities must be estimated along with the component parameters λ_j.

Baum and Eagon (1967) gave an algorithm for locating a local maximum of the likelihood for a special case in which the observed variables have finitely many values. Baum et al. (1970) developed the EM algorithm, proved that it increases the likelihood at each iteration, and applied it to the Markov-dependent mixture model. Leroux (1992a) proved the consistency of maximum likelihood estimators for general Markov-dependent mixture models. These models have been applied to blood cell differentiation (Guttorp, Newton, and Abkowitz, 1990), sequences of bases in a DNA molecule (Churchill, 1989), rainfall (Smith, 1987), and automatic speech recognition (Levinson, Rabiner, and Sondhi, 1983).

Markov-dependent mixture models are closely related to state-space models, in which unobserved state variables determine the distribution of the observations. In many applications, the goal is reconstruction of the state variable based on an observation set. This is achieved by the Kalman filter in Gaussian linear state-space models. Kitagawa (1987) gives recursive equations for filtering, smoothing, and prediction that apply to many non-Gaussian nonlinear models (see also Kohn and Ansley, 1987). The key elements seem to be a state process that is Markov and an observation sequence constructed from a conditionally independent sequence, given the state process. Thus we find overlap with Markov-dependent mixture models, for which the application of the EM algorithm involves reconstruction of the underlying Markov chain. Versions of the smoothing and filtering equations of Kitagawa (1987) were derived in Lindgren (1978) and Askar and Derin (1981) for Markov-dependent mixture models.

4. Selecting the Number of Components

We apply the general theory of model selection (Linhart and Zucchini, 1986) to the selection of the number of components, m, in a mixture distribution or a Markov-dependent mixture model. In particular, we propose that m be chosen to maximize a criterion of the form $l_m - a_{mn}$, where l_m is the log-likelihood maximized over models with m components and a_{mn} is a prespecified penalty term depending on m and the sample size n. Two examples are the Akaike information criterion (AIC), $l_m - d_m$, and the Bayesian information criterion (BIC), $l_m - (\log n)d_m/2$, where d_m is the number of free parameters in the model with m components (this is $2m - 1$ for Poisson mixtures and m^2 for Markov-dependent Poisson mixture models). The penalty term discourages the selection of an excessive number of components. When a finite number of components is physically plausible, this seems reasonable. In any case, a small number of components that provides a good fit yields a simple explanatory model for the data. The components may have convenient physical

interpretations and be useful in comparative studies involving more than one data set, or in clustering multivariate data (Aitkin, Anderson, and Hinde, 1981). Also, fitting a model with a large number of components requires estimation of a large number of parameters (particularly for the Markov-dependent model) and a consequent loss of precision in these estimates.

Theoretical justification for these criteria can be given for the estimation of a mixing distribution. Leroux (1992b) proved that, under mild conditions, the estimator obtained with the number of components selected using AIC or BIC (or certain other criteria) is consistent for the true mixing distribution (finite or not) and has, in the limit, at least as many components.

5. Estimation for Mixture Distributions with Independent Observations

Algorithms for the computation of the maximum likelihood estimate of a mixing distribution have been given by Simar (1976), Lindsay (1981), and DerSimonian (1986). These algorithms determine the number of components m required to maximize the likelihood and the estimates of the parameters $\alpha_1, \ldots, \alpha_m$ and $\lambda_1, \ldots, \lambda_m$. On the other hand, our penalized-likelihood procedure requires maximization of the likelihood for each number of components separately, up to the number of distinct observations. (A potential saving in computation results from the observation that whenever the addition of a component fails to produce an increase in likelihood, the maximum value of the likelihood has been found, i.e., $l_{m+1} = l_m$ implies $l_m = \max_k l_k$.)

5.1 *The EM Algorithm*

Algorithms for finding local maxima of the likelihood for a fixed number of components include EM, Newton–Raphson, and scoring. The latter two methods require more operations per iteration than the EM algorithm, but tend to require fewer iterations to achieve convergence (Redner and Walker, 1984). The EM algorithm consists of the following simple updating steps:

$$\alpha_j^{(\omega+1)} = \frac{\sum_{i=1}^{n} \hat{u}_j^{(\omega)}(i)}{\sum_{k=1}^{m} \sum_{i=1}^{n} \hat{u}_k^{(\omega)}(i)} \tag{3}$$

and set $\lambda_j^{(\omega+1)}$ equal to the value of λ_j ($j = 1, \ldots, m$) that maximizes the conditional expected log-likelihood,

$$\sum_{i=1}^{n} \hat{u}_j^{(\omega)}(i) \log f(y_i; \lambda_j). \tag{4}$$

In (3) and (4),

$$\hat{u}_j^{(\omega)}(i) = \frac{\alpha_j^{(\omega)} f(y_i; \lambda_j^{(\omega)})}{\sum_{k=1}^{m} \alpha_k^{(\omega)} f(y_i; \lambda_k^{(\omega)})}$$

is the current estimated value of the conditional probability that observation i was generated by the jth component, and ω is the iteration number.

The estimate for α_j in (3) can be thought of as a weighted empirical relative frequency of component j with each observation weighted by its conditional probability of component membership. If $f(y; \lambda)$ is a Poisson density with mean λ, then

$$\lambda_j^{(\omega+1)} = \frac{\sum_{i=1}^{n} \hat{u}_j^{(\omega)}(i)\, y_i}{\sum_{i=1}^{n} \hat{u}_j^{(\omega)}(i)}$$

is a weighted average of the observed counts with conditional probability weights.

The algorithm can be started with either initial estimates $(\alpha_1^{(0)}, \ldots, \alpha_m^{(0)}, \lambda_1^{(0)}, \ldots, \lambda_m^{(0)})$ or conditional probabilities $u_j^{(0)}(i)$, and terminated whenever the changes in parameter estimates are sufficiently small.

5.2 *Starting Values for the EM Algorithm*

We propose to use the EM algorithm with several sets of starting values to locate local maxima and then choose the one with the largest likelihood value. Because we wish to consider several values of m, we require a procedure for automatically generating starting values that can be expected to provide a good chance of finding the global maximum value of the likelihood function for a fixed m. Laird (1978) suggested grid search; however, even for moderate numbers of components, the parameter space has high dimension and the number of grid points becomes prohibitively high.

Our approach is as follows. We first group the distinct observed counts into m clusters, with counts considered to arise from the same underlying rate grouped together. An initial estimate of the mixing distribution is obtained from the clusters by an execution of an iteration of the EM algorithm with the posterior probabilities $\hat{u}_j(i)$ replaced by indicators of cluster membership, i.e., $\hat{u}_j(i) = 1$ if observation i has been placed in cluster j. McLachlan and Basford (1988, §1.7) suggest choosing starting values in this way, with clusters generated by a clustering algorithm.

A simple way of forming the clusters is via disjoint intervals with the observations classified according to which of the intervals they fall into. The set of all possible partitions of the data that can be formed this way will generate a set of initial parameter estimates covering a wide range of mixing distributions. We take this as our set of starting values. For example, for count data, one interval might include only counts of 0 while a second interval includes all other counts. An iteration of the EM algorithm in this case will produce a component with $\lambda_1 = 0$ and a second with $\lambda_2 > 0$. The corresponding proportions will be α_1, equal to the proportion of the observed counts that are equal to 0, and α_2 the complementary proportion. Note that all subsequent iterations of the EM algorithm will preserve the component with the zero rate; thus, in order to find a global maximum of the likelihood, it is necessary to include starting values with zero rates as well as starting values with no zero rate.

To implement our procedure, we use an algorithm of Nijenhuis and Wilf (1978, Chap. 5) for determining all possible partitions of the distinct observations into m groups. This algorithm decomposes the total number of distinct observations, K, into the sum $K = r_1 + \cdots + r_m$; thus, starting with the smallest observation, for each i we assign the next r_i distinct observations to the ith group.

5.3 *The Observed Information Matrix*

Louis (1982) showed how to compute the observed information matrix for general missing information problems when using the EM algorithm; certain simplifications, which we describe below, occur in the case of finite mixtures. Let $\mathbf{U}(i) = (U_1(i), \ldots, U_m(i))$, where $U_j(i)$ is 1 if Y_i is sampled from the jth component distribution and 0 otherwise. Then the density of $(Y_i, \mathbf{U}(i))$ is $p(y_i, \mathbf{u}(i); \boldsymbol{\Phi}) = \alpha_j f(y_i; \lambda_j)$, if $u_j(i) = 1$, where $\boldsymbol{\Phi} = (\alpha_1, \ldots, \alpha_{m-1}, \lambda_1, \ldots, \lambda_m)$ and $\alpha_m = 1 - \sum_{j=1}^{m-1} \alpha_j$. Define $\mathbf{S}_i = \partial \log p(y_i, \mathbf{u}(i); \boldsymbol{\Phi})/\partial\boldsymbol{\Phi}$ and $\mathbf{H}_i = -\partial^2 \log p(y_i, \mathbf{u}(i); \boldsymbol{\Phi})/\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}^{\mathrm{T}}$. Louis's formula for the matrix of negative second derivatives of the log-likelihood for the observations $y_1, \ldots, y_n$ is

$$\sum_{i=1}^{n} \mathrm{E}[\mathbf{H}_i - \mathbf{S}_i\mathbf{S}_i^{\mathrm{T}} \mid y_i] + \sum_{i=1}^{n} \mathrm{E}[\mathbf{S}_i \mid y_i]\mathrm{E}[\mathbf{S}_i^{\mathrm{T}} \mid y_i].$$

The first term reduces to

$$\begin{bmatrix} \mathbf{0} & \sum_i \mathbf{A}^{\mathrm{T}}(i) \\ \sum_i \mathbf{A}(i) & \sum_i \mathbf{B}(i) \end{bmatrix},$$

where $\mathbf{0}$ is an $(m-1) \times (m-1)$ matrix of zeros,

$$\mathbf{A}_{jk}(i) = \begin{cases} -\hat{u}_j(i)\alpha_j^{-1}\partial \log f(y_i; \lambda_j)/\partial\lambda_j, & 1 \leqslant j, k \leqslant m-1, \\ \hat{u}_m(i)\alpha_m^{-1}\partial \log f(y_i; \lambda_m)/\partial\lambda_m, & j = m, 1 \leqslant k \leqslant m-1, \end{cases}$$

and

$$\mathbf{B}_{jk}(i) = \begin{cases} -\hat{u}_j(i)(\partial^2 \log f(y_i; \lambda_j)/\partial\lambda_j^2 - [\partial \log f(y_i; \lambda_j)/\partial\lambda_j]^2), & 1 \leqslant j = k \leqslant m, \\ 0 & j \neq k. \end{cases}$$

The second term in Louis's formula is $\sum_{i=1}^n \hat{\mathbf{S}}_i \hat{\mathbf{S}}_i^{\mathrm{T}}$, where $\hat{\mathbf{S}}_i = \partial \log p(y_i, \hat{\mathbf{u}}(i); \Phi)/\partial\Phi$. The observed information results from the substitution of the maximum likelihood estimate for Φ. Note that, if each $\hat{\lambda}_j$ is an interior point, then $\sum_i \mathbf{A}(i) = \mathbf{0}$.

6. Estimation for Markov-Dependent Mixture Models

Let $y_1, \ldots, y_n$ be the realization of a Markov-dependent mixture model with underlying Markov chain $\{X_i\}$. Define Φ by $(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{mm}, \lambda_1, \ldots, \lambda_m)$, where $\alpha_{jk} = \Pr(X_i = k | X_{i-1} = j)$ denote the stationary transition probabilities of $\{X_i\}$. The likelihood function for Φ is

$$L(\Phi | y_1, \ldots, y_n) = \sum_{x_1=1}^{m} \cdots \sum_{x_n=1}^{m} \alpha_{x_1}^{(1)} f(y_1; \lambda_{x_1}(\Phi)) \prod_{i=2}^{n} \alpha_{x_{i-1}, x_i}(\Phi) f(y_i; \lambda_{x_i}(\Phi)),$$

where $\alpha_j^{(1)} = \Pr(X_1 = j)$ denote the initial probabilities of $\{X_i\}$. Since $L(\Phi | y_1, \ldots, y_n)$ is a convex combination of likelihood values obtained with a fixed initial state (i.e., with $\alpha_j^{(1)} = 1$ for some j), simultaneous maximization of $L(\Phi | y_1, \ldots, y_n)$ over Φ and $(\alpha_1^{(1)}, \ldots, \alpha_m^{(1)})$ can be accomplished by maximization over Φ with a fixed initial state. Thus, we will assume in what follows that the $\alpha_j^{(1)}$ are known. It is also possible to obtain maximum likelihood estimates subject to certain restrictions such as those resulting from setting some transition probabilities to zero or to a common value.

6.1 *The EM Algorithm*

The EM algorithm can be applied to likelihood maximization for the Markov-dependent mixture model almost as simply as for the independent mixture model. The log-likelihood function for (x_i, y_i), $i = 1, \ldots, n$ (called the complete-data log-likelihood) is

$$\begin{aligned} &\log L^{\mathrm{c}}(\Phi | x_1, y_1, \ldots, x_n, y_n) \\ &\quad = \log \alpha_{x_1}^{(1)} + \sum_{i=2}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} v_{jk}(i) \log \alpha_{jk} + \sum_{i=1}^{n} \sum_{j=1}^{m} u_j(i) \log f(y_i; \lambda_j), \end{aligned} \tag{5}$$

where $u_j(i)$ is 1 if $x_i = j$ and 0 otherwise, and $v_{jk}(i)$ is 1 if a transition from j to k occurred at i (i.e., $x_{i-1} = j$, $x_i = k$) and 0 otherwise (Φ is suppressed for ease of notation). This log-likelihood function consists of two parts, the log-likelihood for a Markov chain, depending only on the transition probabilities α_{jk}, and the log-likelihood for independent observations, depending only on the parameters $\lambda_1, \ldots, \lambda_m$. Note that when α_{jk} is independent of j, (5) gives the complete-data likelihood for the independent case, so that the independent model is nested in the Markov-dependent mixture model.

The M-step requires maximization of $E\{\log L^c(\Phi) | y_1, \ldots, y_n\}$, which is obtained by replacing the components of the missing data by their conditional means

$$\hat{v}_{jk}(i) = E\{v_{jk}(i) | y_1, \ldots, y_n\} = \Pr\{X_{i-1} = j, X_i = k | y_1, \ldots, y_n\}$$

and

$$\hat{u}_j(i) = E\{u_j(i) | y_1, \ldots, y_n\} = \Pr\{X_i = j | y_1, \ldots, y_n\}.$$

The maximizing values of the transition probabilities are

$$\alpha_{jk} = \frac{\sum_{i=2}^{n} \hat{v}_{jk}(i)}{\sum_{i=2}^{n} \sum_l \hat{v}_{jl}(i)}. \tag{6}$$

These equations, like the equations (3) for the mixing proportions in a mixture distribution, can be thought of as weighted empirical relative frequencies. The maximizing values of λ_j are obtained exactly as for independent observations, i.e., by maximization of (4). As before, the algorithm is terminated when changes in parameter estimates are small. The generation of starting values is considered later.

6.2 *The Forward–Backward Algorithm*

The forward–backward algorithm of Baum et al. (1970) is designed to calculate the conditional probabilities $\hat{u}_j(i)$ and $\hat{v}_{jk}(i)$. It is based on simple recursive formulae for

$$a_j(i) = p(y_1, \ldots, y_i, X_i = j)$$

and

$$b_j(i) = p(y_{i+1}, \ldots, y_n | X_i = j),$$

which yield the quantities of interest by

$$\hat{u}_j(i) = \frac{a_j(i) b_j(i)}{\sum_l a_l(n)}$$

and

$$\hat{v}_{jk}(i) = \alpha_{jk} f(y_i; \lambda_k) a_j(i-1) b_k(i) \Big/ \sum_l a_l(n).$$

The $a_j(i)$ and $b_j(i)$ are calculated recursively in i using the following formulae:

$$a_j(i) = \sum_{k=1}^{m} a_k(i-1) \alpha_{kj} f(y_i; \lambda_j) \tag{7}$$

$[a_j(1) = \alpha_j^{(1)} f(y_1; \lambda_j), j = 1, \ldots, m]$, and

$$b_j(i) = \sum_{k=1}^{m} \alpha_{jk} f(y_{i+1}; \lambda_k) b_k(i+1) \tag{8}$$

$[b_j(n) = 1, j = 1, \ldots, m]$. Note that the a's are computed by a forward pass through the observations and the b's by a backward pass after evaluating the a's. The likelihood is then simply calculated by the expression $\sum_j a_j(n)$. The total number of computations required is of linear order with respect to the sample size.

In theory, the above algorithm appears to provide the required conditional probabilities. However, for many situations the algorithm is numerically unstable because $a_j(i)$ converges rapidly to 0 or diverges to ∞ as i increases, thus making it impossible to calculate and store

long sequences. In fact, with probability 1, $p(y_1, \ldots, y_i) = \sum_{j=1}^{m} a_j(i)$ is approximately equal to $\exp(-iH)$ for large i, where H is a constant called the entropy of the process (Leroux, 1992a), and so $\sum_{j=1}^{m} a_j(i)$ converges to 0 or ∞ at an exponential rate. Typically, all of $a_1(i), \ldots, a_m(i)$ are about the same order of magnitude, so the same conclusion applies to each $a_j(i)$ individually. Various methods for avoiding this problem have been proposed in the field of speech recognition (Devijver, 1985). However, these methods are designed to compute $\hat{u}_j(i)$ and other probabilities such as $\Pr\{X_i = j \mid y_i, \ldots, y_1\}$, and appear unable to produce $\hat{v}_{jk}(i)$.

We use the following approach. For each i, we determine and store the order of magnitude of $\sum_j a_j(i)$, i.e., the integer p for which $10^{-p} \sum_j a_j(i)$ lies between .1 and 1.0, and multiply $a_j(i), j = 1, \ldots, m$, by 10^{-p}; then $a_j(i+1), j = 1, \ldots, m$ are computed. A similar procedure is applied to $b_j(i)$. Afterward, $a_j(i)$ and $b_j(i)$ can be reconstructed for the purpose of computing $\hat{u}_j(i)$ and $\hat{v}_{jk}(i)$.

6.3 *Starting Values for the EM Algorithm*

The choice of starting values is particularly important for hidden Markov models because of the large number of parameters and the tendency for there to exist many local maxima of the likelihood function. We present our procedure for automatically generating good starting values. In our experience, this procedure has performed well in the sense that the starting values generated have appeared to be sufficient to locate the global maximum of the likelihood (i.e., additional exploration of the likelihood surface did not find a larger likelihood value).

Our strategy for generating starting values uses the partitions of the observed distinct counts described in Section 5.2. For each of these partitions, we calculate the starting values for the rates λ_j and proportions α_j in the finite mixture model as described in Section 5.2, and generate four sets of values for the $v_{jk}(i)$ as follows:

(i) $v_{jk}(i) = \begin{cases} 1 & \text{if } y_{i-1} \text{ is in the } j\text{th component of the partition and } y_i \text{ is in the } k\text{th,} \\ 0 & \text{otherwise;} \end{cases}$

(ii) $v_{jk}(i) = \alpha_j \alpha_k$;

(iii) $v_{jk}(i) = \alpha_j$;

(iv) $v_{jk}(i)$ is proportional to $f(y_{i-1}; \lambda_j) f(y_i; \lambda_k)$.

Next we calculate convex combinations of all pairs of the above listed values (with weights .1 and .9), and with each resulting set of values of the $v_{jk}(i)$ the EM algorithm is entered at the M-step. For each set of starting values, the initial state of the Markov chain is allowed to take on all possible values.

It is necessary to consider several different values of $v_{jk}(i)$, and convex combinations of these, because some will have certain features that are preserved by the EM algorithm; only a subset of the parameter space could be explored using these. For instance, values produced by (i) tend to have many zero transition probabilities, which the EM algorithm preserves. Similarly, (ii) and (iii) generate models with independence and complete dependence, respectively, in the underlying Markov chain, both of which are also preserved by the EM algorithm.

7. Application to Fetal Movement Data

The maximum likelihood estimates (and standard errors) for Poisson mixture models with 1–4 components applied to the data described in Section 2 are given in Table 2. Although four components are required to maximize the likelihood, AIC and BIC indicate that two is the best choice. The two-component estimate has a clear interpretation in terms of a

state with a background rate of movement $\hat{\lambda}_1 = .2302$, which the fetus occupies about 94% of the time ($\hat{\alpha}_1 = .9388$), and an excited state occupied about 6% of the time ($\hat{\alpha} = .0612$), with movement occurring at rate $\hat{\lambda}_2 = 2.3242$, or approximately 10 times the background rate.

Table 3 contains the fitted frequency distributions for the maximum likelihood estimates with one, two, and three components, the unrestricted maximum likelihood estimate, and the maximum likelihood estimates for the negative-binomial family. The standardized deviations of the observed frequencies from the fitted for the Poisson distribution show the pattern, predicted by the result of Shaked (1980), of observed frequencies higher than predicted at the low and high values, and lower than predicted at the middle values. The two-component estimate provides a large improvement in fit over the Poisson distribution and there is some further improvement provided by three components that might be thought to be important. In this application there is little statistical evidence to favour the choice of a finite mixture as opposed to a negative-binomial distribution, although the physical interpretation of the former appears more meaningful.

Markov-dependent mixture models take into account the belief that observations in close time intervals will be related. The results in Table 4, obtained using the algorithm described in Section 6, indicate strong association between movement counts in adjacent time

Table 2
Independent mixture model estimates for fetal movement data

Number of components	Estimates		Log-likelihood	AIC	BIC
	Probability	Rate			
1	1 (0)	.3583 (.0386)	−174.26	−175.26	−177.00
2	.9388 (.0518)	.2302 (.0611)	−160.21	−163.21	−168.43
	.0612 (.0518)	2.3242 (1.0093)			
3	.4380[a]	0[a]	−159.01	−164.01	−172.71
	.5447[a]	.5320[a]			
	.0173[a]	3.9683[a]			
4	.4201[a]	0[a]	−159.00	−166.00	−178.18
	.5462[a]	.4918[a]			
	.0214[a]	1.6615[a]			
	.0123[a]	4.4074[a]			

[a] Standard errors are not attainable (or interpretable) for these parameters because the estimate is on the boundary of the parameter space and the information matrix is not nonnegative definite.

Table 3
Fitted frequency distributions for independent mixture models applied to the fetal movement data[a]

Count	Observed frequency	Poisson	2-component mixture	3-component mixture	MLE	Negative binomial
0	182	167.72 (1.1)	180.42 (.1)	182.00 (.0)	182.00 (.0)	182.55 (.0)
1	41	60.10 (−2.5)	44.54 (−.5)	41.16 (.0)	41.20 (.0)	39.00 (.3)
2	12	10.77 (.4)	8.62 (1.1)	11.48 (.2)	11.39 (.2)	12.04 (.0)
3	2	1.29 (.6)	3.37 (−.7)	2.74 (−.4)	2.85 (−.5)	4.10 (−1.0)
4	2	.12 (8.2)	1.77 (.0)	1.07 (.2)	1.07 (.3)	1.46 (.5)
5	0	.01	.81	.67	.62	.53
6	0	.00	.31	.43	.40	.20
7	1	.00	.10	.24	.24	.07
8+	0	.00	.04	.21	.24	.05

[a] Observed minus fitted frequencies divided by $[\text{fitted}]^{1/2}$ in parentheses (counts of 4 and higher are combined).

Table 4
Markov-dependent mixture model estimates for fetal movement data

	Parameter estimates								
m	Rate[a]	Transition probability matrix[b]				Stationary probabilities	Log-likelihood	AIC	BIC
1	.3583	1				1	−174.26	−175.26	−177.00
2	.2560*	.9884	.0116			.964	−150.70	−154.70	−161.66
	3.1006	.3083	.6917			.036			
3	.0447*	.9468	.0433	.0099		.482	−139.50	−148.50	−164.16
	.5090	.0424	.9576	0		.492			
	3.4138	.1838	0	.8162		.026			
4	0	1	0	0	0	1[c]	−134.97	−150.97	−178.81
	.2237*	0	.9841	.0070	.0089	0			
	.6689	.0206	0	.9794	0	0			
	3.3478	0	.1944	0	.8056	0			

[a] The likelihood is maximized with the initial state indicated by *.

[b] The (i, j)th element is the estimated probability $\hat{\alpha}_{ij}$ of transition from state i to j; e.g., for $m = 2$, .0116 is the estimated conditional probability that the chain is in state 2 in the next 5-second interval given that it is in state 1 in the current interval.

[c] State 1 is absorbing and all others are transient.

intervals. For instance, according to the two-component model, if the fetus is in state 1 (with $\hat{\lambda}_1 = .2560$), it remains in this state for the next time interval with probability $\hat{\alpha}_{11} = .9884$; for state 2 this probability is $\hat{\alpha}_{22} = .6917$.

Both the two- and three-component models have rates that are quite close to those of the estimated two- and three-component mixture distributions. The stationary probabilities of the estimated stochastic matrices (Table 4) are also quite similar to the corresponding estimated mixing proportions for $m = 2$ and $m = 3$. Although we have assumed stationarity, these data may in fact be nonstationary; for instance, the mixture distribution may change over time.

The Markov-dependent mixture models provide a much improved fit over the corresponding independent models, as indicated by the log-likelihood values; out of all of the models considered, AIC selects the Markov-dependent model with three components, whereas BIC selects that with two components.

8. Extensions

The fetal movement data could also have been modeled in continuous time using a continuous-time Markov chain. In such a model, the parameters of interest would be the Poisson intensities and the transition rates. Inference for such a model was considered by Meier-Hellstern (1987), where it is referred to as a Markov-modulated Poisson process.

We are currently exploring extensions of the models considered here, in which the distribution of the observations is allowed to depend on the values of covariates. There is great flexibility for modeling covariate effects in these models, since the component rates, the transition probabilities, or even the number of components, could all be allowed to vary.

The presence of overdispersion in counts is usually detected using omnibus tests such as the dispersion test or score tests against mixed Poisson alternatives (Dean and Lawless, 1989; Paul, Liang, and Self, 1989). The models presented here provide the basis for tests of overdispersion based on finite mixture (or Markov-dependent) alternatives.

FORTRAN codes for the algorithms herein are available free of charge from the first author.

Acknowledgements

This work is based on a part of Brian Leroux's Ph.D. dissertation in the Department of Statistics, The University of British Columbia. It has been partially supported by Grant A-5527 from the Natural Sciences and Engineering Research Council. The authors wish to thank Harry Joe and Randy Sitter for helpful comments and Dan Rurak for providing data. Two referees provided helpful comments on exposition.

Résumé

Cet article concerne l'utilisation et la mise en oeuvre de procédures de maximisation de la vraisemblance pénalisée pour choisir le nombre de composantes de mélange et estimer les paramètres provenant de modèles supposant soit l'indépendance soit une dépendance markovienne. Le calcul numérique des estimateurs est effectué au moyen d'algorithmes de génération automatique de valeurs de départ pour l'algorithme EM. Le calcul de la matrice d'information est aussi discuté. Des modèles de mélange de Poisson sont appliqués à une série de décomptes des mouvements *in utero* d'un foetus d'agneau obtenus par ultrasons. Les estimateurs qui en résultent s'interprètent de façon plausible en termes de mécanismes reliés au processus physiologique.

References

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, Series A* **144,** 419–461.

Askar, M. and Derin, H. (1981). A recursive algorithm for the Bayes solution of the smoothing problem. *IEEE Transactions on Automatic Control* **26,** 558–560.

Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society* **73,** 360–363.

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

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51,** 79–94.

Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* **84,** 467–472.

DerSimonian, R. (1986). Algorithm AS 221. Maximum likelihood estimation of a mixing distribution. *Applied Statistics* **35,** 302–309.

Devijver, P. A. (1985). Baum's forward–backward algorithm revisited. *Pattern Recognition Letters* **3,** 369–373.

Guttorp, P., Newton, M. A., and Abkowitz, J. L. (1990). A stochastic model for haematopoiesis in cats. *IMA Journal of Mathematics Applied in Medicine & Biology* **7,** 125–143.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association* **82,** 1032–1041.

Kohn, R. and Ansley, C. F. (1987). Comment on Kitagawa (1987). *Journal of the American Statistical Association* **82,** 1041–1044.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73,** 805–811.

Leroux, B. G. (1992a). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and Their Applications* **40,** 127–143.

Leroux, B. G. (1992b). Consistent estimation of a mixing distribution. *Annals of Statistics* **20,** in press.

Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process in automatic speech recognition. *Bell System Technical Journal* **62,** 1035–1074.

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics* **5,** 81–91.

Lindsay, B. G. (1981). Properties of the maximum likelihood estimator of a mixing distribution. In *Statistical Distributions in Scientific Work*, Vol. 5, C. Taillie et al. (eds), 95–109. New York: Reidel Holland.

Lindsay, B. G. (1983). The geometry of mixing likelihoods: A general theory. *Annals of Statistics* **11,** 86–94.

Linhart, H. and Zucchini, W. (1986). *Model Selection.* New York: Wiley.

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

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models.* New York: Dekker.

Meier-Hellstern, K. S. (1987). A fitting algorithm for Markov-modulated Poisson processes having two arrival rates. *European Journal of Operational Research* **29,** 370–377.

Nijenhuis, A. and Wilf, H. S. (1978). *Combinatorial Algorithms*, 2nd edition. New York: Academic Press.

Paul, S. R., Liang, K. Y., and Self, S. G. (1989). On testing departure from the binomial and multinomial assumptions. *Biometrics* **45,** 231–236.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *Society for Industrial and Applied Mathematics, Review* **26,** 195–239.

Shaked, M. (1980). On mixtures from exponential families. *Journal of the Royal Statistical Society, Series B* **42,** 192–198.

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Annals of Statistics* **4,** 1200–1209.

Smith, J. A. (1987). Statistical modeling of rainfall occurrences. *Water Resources Research* **23,** 885–893.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*, 7th edition. Ames, Iowa: University of Iowa Press.

Wittmann, B. K., Rurak, D. W., and Taylor, S. (1984). Real-time ultrasound observation of breathing and body movements in fetal lambs from 55 days gestation to term. Abstract presented at the XI Annual Conference, Society for the Study of Fetal Physiology, Oxford.

Received August 1990; *revised May* 1991; *accepted July* 1991.