Mathematical Population Studies, 12:135–157, 2005
Copyright © Taylor & Francis Inc.
ISSN: 0889-8480 print/1543-5253 online
DOI: 10.1080/08898480591005168

Routledge
Taylor & Francis Group

Model-based Clustering of Sequential Data with an Application to Contraceptive Use Dynamics

José G. Dias
Higher Institute of Social Sciences and Business Studies, Lisbon, Portugal

Frans Willekens
Netherlands Interdisciplinary Demographic Institute,
The Hague, The Netherlands

Multi-state models describe the transitions people experience as life unfolds. The transition probabilities depend on sex, age, and attributes of the person and the context. Empirical evidence suggests that attributes that cannot be measured directly may at most be inferred from a long list of observable characteristics. A cluster-based, discrete-time multi-state model is presented, where transition probabilities are estimated simultaneously for several subpopulations of a heterogeneous population. The subpopulations are not defined a priori but are determined on the basis of similarities in behavior in order to determine which women exhibit similar characteristics with respect to method choice, method switch, discontinuation and subsequent resumption of contraceptive use. The data are from the life history calendar based on the Brazilian Demographic and Health Survey 1996. The parameters of the model are estimated using the EM algorithm. Seven subpopulations with heterogeneous transition probabilities are identified.

Keywords: finite mixture models; Markov models; unobserved heterogeneity; contraceptive use dynamics; life history calendar

1. INTRODUCTION

Demographers are increasingly interested in understanding life histories or individual biographies with a focus on events, their

The authors are grateful to Sabu S. Padmadas and Vladimir Canudas Romo for helpful comments on earlier drafts.

This research was supported by Fundação para a Ciência e Tecnologia Grant no. SFRH/BD/890/2000 (Portugal).

Address correspondence to José G. Dias, Department of Quantitative Methods, ISCTE – Instituto Superior de Ciências do Trabalho e da Empresa, Av. das Forças Armadas, 1600 Lisboa, Portugal. E-mail: jose.dias@iscte.pt

sequence, ordering and transitions that people make from one state of life to another, for example, living to dead, single to married, or migration from one residence to another. The transitions in life generate life paths that encompass both continuity and change. Multi-state models have been developed to describe the dynamics and to determine the states occupied at different ages. There are several applications of multi-state models in family demography, labor economics, migration, and public health research. Contraceptive use dynamics is a relatively new area of application because age-specific data on method adoption, method switch, and discontinuation of use became available only recently (Islam, 1994). Most studies focus on reasons for stopping contraceptive use and apply competing risks models that include the multiple-decrement life table (Kost, 1993; Steele and Diamond, 1999). Some consider contraceptive behavior after discontinuation. In these studies, the unit of analysis is an episode or segment of method use or non-use (Kost, 1993: 111). A multistate approach addresses sequences of episodes. Both types of analyses take advantage of the availability of calendar-history data on contraceptive use.

In multi-state models, which include the increment-decrement life table, transitions are governed by transition probabilities. In demographic studies, the probabilities are usually functions of age and sex, but they may also vary by other attributes. The attributes are generally used to stratify the population, although they may also be used in a regression model as predictors of the transition probabilities. In the latter case, the attributes are generally referred to as covariates. Important assumptions include that attributes are observable and the state occupancies are observed, that people with the same set of attributes have the same transition probabilities (homogeneity property), and that the transition probability depends on current status and is independent of the past (Markov property).

In the model presented, group membership is not directly observed. It is a latent variable, the value of which depends on the observed characteristics. In other words, group membership is not defined *a priori* but rather is determined endogenously by the model. The number of subpopulations depends on the extent of heterogeneity of the observed population. In view of parsimony, the aim will be to find the smallest number of subpopulations that adequately describes the heterogeneous population. The latent variable representing group membership is a discrete variable. It defines a finite mixture model (McLachlan and Peel, 2000).

The development of finite mixture models dates to the nineteenth century (Newcomb, 1886). In recent decades, due to advances in computing, finite mixture models have proven powerful tools for the

analysis of a wide range of social and behavioral science data. In the social sciences, following the popularity of latent class models, which have a long tradition (following the seminal work by Lazarsfeld), more advanced finite mixture models have become popular. Recent applications of finite mixture models span such social science areas as economics (Wedel *et al*., 1993), psychology (Böckenholt, 1993), management (Rosbergen *et al*., 1997), and marketing (Wedel and Kamakura, 1999). In particular, finite mixtures for sequential data have become very popular in scientific fields such as machine learning (Rabiner and Juang, 1986), biology (Eddy, 1998), economics (Guha and Banerji, 1998/1999), and marketing (Poulsen, 1990). An early approach to sequential data is the mover-stayer model (Blumen *et al*., 1955). This model assumes that the population consists of two subpopulations: movers and stayers. Movers are assumed to follow a first-order Markov process with a constant transition probability, whereas stayers have a transition probability equal to the identity matrix. Thus, these models belong to the latent Markov family with an immune fraction (one of the subpopulations is immune to move) and can be estimated by the EM algorithm (Fuchs and Greenhouse, 1988).

The applications of finite mixture models in demography are few (Haughton and Haughton, 1996; Li and Choe, 1997; Willekens, 1999). Most studies of unobserved heterogeneity in demography assume continuous mixing distributions (Vaupel and Yashin, 1985).

We present a finite mixture model for demographic sequential data. In section 2, we discuss the main concepts on unobserved heterogeneity. In section 3, we introduce the finite mixture methodology and the estimation using the EM algorithm. In section 4, we explore this model by using simulated data that allow a better understanding of model selection criteria. In section 5, we illustrate the procedure using sequential data provided by the life history calendar implemented under the 1996 Brazil Demographic and Health Survey. The conclusions of this study are expected to yield new insights regarding potential substantive applications in demography.

2. UNOBSERVED HETEROGENEITY

Latent variable models can be defined as $f(\mathbf{x}_i) = \int f(\mathbf{x}_i|z_i)f(z_i)dz_i$ where $\mathbf{x}_i$ is an observed J – dimensional vector of variables, z_i is a latent unidimensional variable, and $f(z_i)$ is the mixing distribution with respect to z_i. Due to historical reasons, we focus on duration data with a parametric mixing distribution. This model, also known as the frailty model (Vaupel *et al*., 1979; Lancaster, 1979), usually assumes a conjugate mixing distribution such as the gamma distribution

(Lancaster, 1979; Vaupel *et al*., 1979). For a detailed analysis of this models, see Lancaster (1990). One of the problems of this model pertains to the flexibility of the mixing distribution, i.e., the model may not be robust concerning alternative parametric specifications of the mixing distribution. Indeed, Heckman and Singer (1984) showed that the results are sensitive to the choice of the mixing distribution. Another aspect is that the same $f(\mathbf{x}_i)$ can be obtained with different combinations of $f(\mathbf{x}_i; z_i)$ and $f(z_i)$ (identification problem (Heckman and Singer, 1984: 274)).

The parametric assumption of $f(z_i)$ can be relaxed by letting it be unspecified, resulting in the semi-parametric mixture model (Lindsay and Lesperance, 1995) that can be estimated by the nonparametric maximum likelihood estimator (NPMLE) introduced by Robbins (1950). This estimator, under a general set of assumptions including identifiability, is consistent (Kiefer and Wolfowitz, 1956; Laird, 1978). Lindsay (1983a) established that the NPMLE of $f(z_i)$ is a discrete distribution (Laird, 1978). Therefore, the NPMLE results in a finite mixture with the number of components (points of support) not specified *a priori*. For developments and understanding of the characteristics of this estimator, see Laird (1978) and Lindsay (1983a, 1983b), who analyzed the NPMLE from the geometric viewpoint of the likelihood as a convex combination of the mixed distribution. For a review, see Lindsay (1995).

3. THE FINITE MIXTURE MODEL

3.1. Model specification

Consider a sample of n respondents. A respondent will be denoted by i $(i = 1, \ldots, n)$. Each respondent is characterized by J attributes. An attribute is denoted by j $(j = 1, \ldots, J)$. The j-th attribute of respondent i is denoted by the random variable $\mathbf{X}_{ij}$ and the sample value is the realization $\mathbf{x}_{ij}$. The vector $\mathbf{X}_i$ consists of elements $\mathbf{X}_{ij}$ with $j = 1, \ldots, J$. The vector $\mathbf{x}_i$ is defined similarly. Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote a J-dimensional sample of size n.

Respondents are grouped into S subpopulations, with a subpopulation denoted by s $(s = 1, \ldots, S)$. The subpopulations, including their numbers, are not defined *a priori*, but are an outcome of the analysis. Thus, in advance one does not know how the population will be divided into subpopulations. The subpopulation which respondent i belongs to is denoted by the unobserved finite discrete variable Z_i with $Z_i \in \Im$ and $\Im = \{1, 2, \ldots, S\}$. The realization of the random value Z_i is z_i, and $\mathbf{z} = (z_1, \ldots, z_n)$. Let $\{(Z_1, \mathbf{X}_1), \ldots, (Z_n, \mathbf{X}_n)\} = \{(Z_i, \mathbf{X}_i)\}_{i=1}^n$ be a sequence

of independent pairs of random variables identically distributed and assuming values on $\Im \times \aleph$, with $\Im$ defined above and $\aleph \subseteq \Re$. Because $\mathbf{z}$ is hidden or missing, the inference problem is to estimate the parameters of the model, say φ, when only $\mathbf{x}$ is observed. Thus, the estimation procedure has to be based on the marginal distribution of $\mathbf{x}_i$.

$$f(\mathbf{x}_i; \varphi) = \sum_{s=1}^{S} \pi_s f_s(\mathbf{x}_i; \theta_s), \tag{1}$$

which defines a finite mixture (FM) model with S subpopulations. The mixture proportions, $\pi_s = p(Z_i = s; \varphi)$, correspond to the *a priori* probability that individual i belongs to the subpopulation s, and gives the subpopulation relative size. This mixing distribution, $\{\pi_s\}_{s=1}^{S}$, since it is a weighting function, satisfies $\pi_s > 0$ and $\sum_{s=1}^{S} \pi_s = 1$. Within each subpopulation (conditional on belonging to subpopulation s) observation $\mathbf{x}_i$ is characterized by the density $f_s(\mathbf{X}_i; \theta_s) = p(\mathbf{X}_i = \mathbf{x}_i \mid \mathbf{Z}_i = s; \theta_s)$. The density is the probability that individual i in subpopulation s has attributes $\mathbf{x}_i$. The functions $f_s(\mathbf{x}_i; \theta_s)$ imply that all individuals in one of the subpopulations have the same probability distribution, only the parameters θ_s vary across subpopulations. The choice of $f_s(\mathbf{x}_i; \theta_s)$ depends on the nature of data. The parameters of the finite mixture model to be estimated are $\varphi = (\pi_1, \ldots, \pi_{s-1}, \theta_1, \ldots, \theta_s)$. Therefore, finite mixture models can be seen as something between unmixed or homogeneous model ($S = 1$) and the NPMLE, since the mixing distribution is the multinomial distribution with S points of support (known *a priori*) and gives a more parsimonious representation of data than the NPMLE.

We now introduce time (or age). Let X_{ijt} denote the value of the attribute j of individual i at time t. We will assume a discrete-time process, since we observe the state at discrete times. The attribute is measured repeatedly during a period from 0 to T_i ($t = 0, 1, \ldots, T_i$). The length of the observation window may differ among individuals. To simplify notation, we consider a single attribute ($J = 1$); however, the extension for $J > 1$ is straightforward. Thus, the vectors $\mathbf{X}_i$ and $\mathbf{x}_i$ denote the consecutive values of the single attribute – respectively, X_{it} and x_{it} – , with $t = 0, \ldots, T_i$. The value assumed by X_{it} is called *state*, and the set of all possible values is the *state space* $\aleph$. For a finite discrete state space the attribute has finite states, say K categories, and $\aleph = \{1, \ldots, K\}$.

The probability function of $\mathbf{x}_i = (X_{i0}, X_{i1}, \ldots, X_{iT_i})$ is extremely difficult to characterize, due to its dimension $(T_i + 1)$. A common procedure to simplifying this expression is by assuming the Markov property that states that the event $X_t = x_t$, only depends upon the

previous state $X_{t-1} = x_{t-1}$. The transition probability of a stationary Markov chain (the transition probabilities are not dependent upon t) is $a_{jk} = P(X_t = k \mid X_{t-1} = j)$, where $t > 0$, and $j, k \in \aleph$. A Markov chain is completely specified by its transition probabilities and initial distribution (Taylor and Karlin, 1994; Norris, 1997; Ross, 2000).

For individual i, the distribution of data conditional on belonging to subpopulation s is $f_s(\mathbf{x}_i; \theta_s) = p(\mathbf{X}_i = \mathbf{x}_i \mid \mathrm{Z}_i = s; \theta_s)$. It defines an extension of the Markov chain incorporating unobserved heterogeneity, which is represented by the latent variable Z in Figure 1, i.e., the observed dynamics of individual i (represented by $\mathbf{X}_i$) are conditional on the subpopulation he belongs to. The random variables $(X_0, X_1, \ldots, X_T)$ are not independent from each other (see Figure 1b). This first-order Markov model is an extension of the latent class model (Lazarsfeld and Henry, 1968), which corresponds to a zero-order Markov model assuming local independence.

To simplify notation, we denote $P(X = x; \theta)$ as $p(x; \theta)$. From the Markov property, it comes

$$f_s(\mathbf{x}_i; \theta_s) = p(x_{i0}; \theta_s) \prod_{t=1}^{T_i} p(x_{it}; x_{i,t-1}, \theta_s), \tag{2}$$

where $p(x_{it};\ x_{i,t-1}, \theta_s)$ is the transition probability that individual i is in state x_{it} at time t, given that he belongs to subpopulation s and is in state $x_{i,t-1}$ at time $t-1$. The assumption of stationarity, i.e., the assumption that the pattern of change persists over time, and the assumption that all individuals in subpopulation s share the same pattern, imply that the probability of making a transition from j to k is independent of t and is the same for all individuals in a particular subpopulation:

$$a_{sjk} = P(X_{it} = k \mid X_{i,t-1} = j, \theta_s) = p(k;\ j, \theta_s), \tag{3}$$

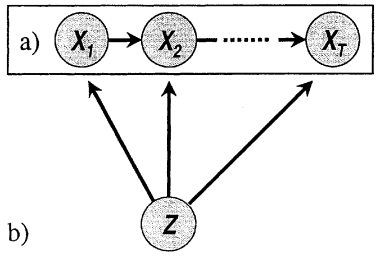

FIGURE 1 Extension from a (a) Markov chain to a (b) finite mixture of Markov chains.

where $t > 0$ and $j, k \in \{1, 2, \ldots, K\}$. Each subpopulation has a different transition matrix $\mathbf{A}_s = (a_{sjk})_{j=1,k=1}^{K,K}$. The j-th row of $\mathbf{A}_s$ includes all the conditional probability distributions of X_t given that $X_{t-1} = j$ and the individual i belongs to subpopulation s. Associated with each initial response x_{i0}, we have K binary variables y_{ij} $(i = 1, \ldots, n; \quad j = 1, \ldots, K)$ defined by the indicator function $y_{ij} = I(x_{i0} = j)$. The initial distribution $\mathbf{Y}_{i0} = (Y_{i1}, \ldots, Y_{iK})$ of the dynamic sequence follows a multinomial distribution, $\mathbf{Y}_{i0} \mid \theta_s \sim \text{Multi}_K(1, \lambda_{s1}, \ldots, \lambda_{sK})$, whose density is

$$p(I(x_{i0} = 1), \ldots, I(x_{i0} = K); \theta_s) = \prod_{j=1}^{K} \lambda_{sj}^{I(x_{i0}=j)}, \tag{4}$$

with $\lambda_{sj} = P(I(X_{i0} = j) \mid Z_i = s)$. From (2), and using (3) and (4), we conclude that the probability of an individual sequence belonging to subpopulation s is

$$f_s(\mathbf{x}_i; \theta_s) = \prod_{j=1}^{K} \lambda_{sj}^{I(x_{i0}=j)} \prod_{j=1}^{K} \prod_{k=1}^{K} a_{sjk}^{n_{ijk}}, \tag{5}$$

where n_{ijk} is the number of transitions from j to k for individual i. Finally, from (1), the finite mixture model for observation $\mathbf{x}_i$ is given by

$$f(\mathbf{x}_i; \varphi) = \sum_{s=1}^{S} \pi_s \prod_{j=1}^{K} \lambda_{sj}^{I(x_{i0}=j)} \prod_{j=1}^{k} \prod_{k=1}^{k} a_{sjk}^{n_{ijk}}, \tag{6}$$

with $\varphi = (\pi_1, \ldots, \pi_{S-1}, \lambda_{11}, \ldots, \lambda_{SK}, a_{111}, \ldots, a_{SKK})$. As independent parameters of this model, we have $S - 1$ prior probabilities, $S(K - 1)$ initial probabilities and $SK(K - 1)$ transition probabilities. Thus, the total number of independent parameters is $SK^2 - 1$. A finite mixture of Markov chains is not a Markov chain, which enables the modeling of more complex patterns (see Appendix I).

Before finishing the specification of the model, we briefly discuss its identification. In general, a parametric family of densities $f(\mathbf{x_i}; \varphi)$ is identified if distinct values of the parameter φ determine distinct members of the family of densities, i.e., $f(\mathbf{x}_i; \varphi) = f(\mathbf{x}_i; \varphi^*)$ if and only if $\varphi = \varphi^*$ (besides label-switching problems). Thus, in a non-identified model, an infinite number of solutions exists. The identifiability of finite mixture models in the exponential family (normal and Poisson) is generally ensured (see Titterington *et al.*, 1985). However, some models such as the latent class model may suffer from identifiability problems (Goodman, 1974; Clogg, 1995).

3.2. Model Estimation

The log-likelihood function for φ, given that $\mathbf{x}_i$ are independent from each other, is

$$\ell_S(\varphi;\mathbf{x}) = \sum_{i=1}^{n} \log \sum_{s=1}^{S} \pi_s \prod_{j=1}^{K} \lambda_{sj}^{I(x_{i0}=j)} \prod_{j=1}^{K}\prod_{k=1}^{K} a_{sjk}^{n_{ijk}}, \tag{7}$$

and the maximum likelihood estimator (MLE) is $\hat{\varphi} = \arg\max_{\varphi} \ell(\varphi;\mathbf{x})$. The MLE of φ is obtained by iterative procedures, such as the Newton-Raphson algorithm (Everitt, 1987). However, an attractive alternative in the context of finite mixture models is the EM algorithm (Dempster *et al.*, 1977; McLachlan and Krishnan, 1997).

The EM algorithm simplifies a complex log-likelihood function into a set of *easily* solvable log-likelihood functions by introducing the "missing variable" $\mathbf{z}$. This algorithm iterates between two steps: in the *E–Step* (Expectation Step), it estimates the conditional mean of the missing variable given the previous estimate of the model parameters and the observations, and in the *M–Step* (Maximization Step), it re-estimates the model parameters given the observations and the soft clustering done by the E–Step. McLachlan and Krishnan (1997) present a very detailed discussion of the EM algorithm. In Appendix II, we present the implementation of the EM algorithm for this model.

3.3. Number of Subpopulations

A main issue is how to estimate S or how many subpopulations do we need to consider in the analysis? A traditional approach to select the best among different models is using a likelihood ratio test. However, in the context of finite mixture models this approach is problematic. The null hypothesis under test is defined on the boundary of the parameter space, and consequently the regularity condition of Cramer on the asymptotic properties of the MLE is not valid under the null hypothesis. Some recent results have been achieved (Chen *et al.*, 2001; Lo *et al.*, 2001). However, most of these results are of difficult implementation and usually derived for finite mixtures of normal distributions.

As an alternative, there has been a recent interest in assessing the finite mixture model fit via information statistics. These statistics are based on the value of $-2\ell_s(\hat{\varphi};\mathbf{x})$ of the model, where $\hat{\varphi}$ represents the maximum likelihood estimate adjusted for the number of free parameters in the model (and other factors such as the sample size). The basic principle under these information criteria is parsimony: all other things being the same, we choose the model with fewer

parameters. Thus, we select S, which minimizes the following criterion $C_s = -2\ell_s(\hat{\varphi}; \mathbf{x}) + dN_s$, where N_s is the total number of free parameters of the model. According to the different values of d, we have the Akaike Information Criterion (AIC: Akaike, 1974) ($d = 2$), the Bayesian Information Criterion (BIC: Schwarz, 1978) ($d = \log n$), and the Consistent Akaike Information Criterion (CAIC: Bozdogan, 1987) ($d = \log n + 1$). Bozdogan (1993) suggested the modified AIC (AIC3) criterion, using 3 instead of 2 as penalizing factor. For these heuristic criteria, smaller values mean more parsimonious models. BIC and CAIC criteria have the advantage of being dimension consistent, they point to the right model with probability one as the sample size increases.

4. MONTE CARLO STUDY

4.1. Experimental Design

To evaluate the performance of the information criteria—AIC, AIC3, CAIC, and BIC—and robustness across experimental conditions, a Monte Carlo (MC) study was conducted.

The Monte Carlo (MC) experimental design controls the number of components, number of variables, and the sample size. We considered 2- and 3-component models: $S \in \{2, 3\}$. The number of repeated measurements ($T_i + 1 = T + 1$) was set at levels 30 and 60. The factor sample size n is set at 200, 500, and 1000. Furthermore, we set $K = 4$ and equal component sizes ($\pi_s = 1/S$). The true parameter values are shown in Table 1. These values mimic the values shown in our application which includes heavy retention probabilities and absorbing states.

This MC study sets a $2^2 \times 3$ factorial design with 12 cells. The main performance measure used is the frequency with which each criterion picks the correct model. For each dataset, each criterion is classified as *under-fitting*, *fitting*, or *over-fitting*, based on the relation between S and the estimated S by those criteria. Apart from the four information criteria mentioned, we also investigated a different definition of the BIC and CAIC criteria. Ramaswamy *et al*. (1993) and DeSarbo *et al*. (2004) have also considered as "sample size" the repeated measurements from each individuals. Therefore, the penalization would be function of nT, instead of n.

Special care needs to be taken before arriving at conclusions based on MC results. We performed 200 replications within each cell to obtain the frequency distribution of selecting the true model, resulting in a total of 2400 datasets. To avoid local optima, for each number of

TABLE 1 The True Parameter Values for the Monte Carlo Study

	Two-component model ($S = 2$)		Three-component model ($S = 3$)		
	$s = 1$	$s = 2$	$s = 1$	$s = 2$	$s = 3$
λ_{s1}	0.20	0.30	0.20	0.30	0.10
λ_{s2}	0.30	0.30	0.30	0.30	0.40
λ_{s3}	0.10	0.30	0.10	0.30	0.40
a_{s11}	0.80	0.85	0.80	0.85	1.00
a_{s12}	0.05	0.10	0.05	0.10	0.00
a_{s13}	0.05	0.02	0.05	0.02	0.00
a_{s21}	0.10	0.02	0.10	0.02	0.02
a_{s22}	0.70	0.85	0.70	0.85	0.95
a_{s23}	0.10	0.03	0.10	0.03	0.01
a_{s31}	0.10	0.05	0.10	0.05	0.03
a_{s32}	0.01	0.05	0.01	0.05	0.002
a_{s33}	0.80	0.85	0.80	0.85	0.90
a_{s41}	0.01	0.05	0.01	0.05	0.04
a_{s42}	0.01	0.02	0.01	0.02	0.02
a_{s43}	0.13	0.03	0.13	0.03	0.04

Note: $\lambda_{s4} = 1 - \lambda_{s1} - \lambda_{s2} - \lambda_{s3}$ and $a_{sj4} = 1 - a_{sj1} - a_{sj2} - a_{sj3}$.

components (2, 3, and 4) the EM algorithm was repeated 5 times with random starting centers, and the best solution (maximum likelihood value out of those 5 runs) and model selection results were kept. The EM algorithm ran until the difference between log-likelihoods being smaller than 10^{-6} (tolerance level).

4.2. Results

The key feature of the results is the overall remarkable performance of AIC3 (Table 2). While most of the criteria perform satisfactory, AIC3 identifies the true model 99.4% and 97.8% for the two- and three-component model, respectively. For both CAIC and BIC, nT reduces their performance and it is not considered hereafter. Overall, BIC performs well with 99.7% and 85.8% for the two- and three-component model respectively. As in other studies, our results document the tendency of AIC to over-fit. BIC and CAIC tend to under-fit, especially for the three-component model.

A second objective of the study was to compare these criteria across the factors in the design. Increasing the sample size almost always improves the performance of the information criteria. However, for AIC increasing the sample size tends to increase the over-fit, without improvement in fit. Increasing the number of measurements (T)

TABLE 2 Results of the Monte Carlo Study

		Criteria					
				CAIC		BIC	
Factors		AIC	AIC3	n	nT	n	nT
Number of components ($S = 2$)							
Sample size (n)							
200	Underfit	0.0	0.0	8.8	48.0	0.8	41.8
	Fit	77.7	99.0	91.2	52.0	99.2	58.2
	Overfit	22.3	1.0	0.0	0.0	0.0	0.0
500	Underfit	0.0	0.0	0.0	0.0	0.0	0.0
	Fit	78.5	99.3	100.0	100.0	100.0	100.0
	Overfit	21.5	0.7	0.0	0.0	0.0	0.0
1000	Underfit	0.0	0.0	0.0	0.0	0.0	0.0
	Fit	75.0	100.0	100.0	100.0	100.0	100.0
	Overfit	25.0	0.0	0.0	0.0	0.0	0.0
Number of variables (T)							
30	Underfit	0.0	0.0	5.8	32.0	0.5	27.8
	Fit	75.0	99.5	94.2	68.0	99.5	72.2
	Overfit	25.0	0.5	0.0	0.0	0.0	0.0
60	Underfit	0.0	0.0	0.0	0.0	0.0	0.0
	Fit	79.0	99.3	100.0	100.0	100.0	100.0
	Overfit	21.0	0.7	0.0	0.0	0.0	0.0
Total							
	Underfit	0.0	0.0	2.9	16.0	0.3	13.9
	Fit	77.0	99.4	97.1	84.0	99.7	86.1
	Overfit	23.0	0.6	0.0	0.0	0.0	0.0
Number of components ($S = 3$)							
Sample size (n)							
200	Underfit	0.0	4.0	48.2	55.2	41.2	52.2
	Fit	87.8	95.5	51.8	44.8	58.8	47.8
	Overfit	12.2	0.5	0.0	0.0	0.0	0.0
500	Underfit	0.0	0.0	1.2	34.5	0.2	21.5
	Fit	82.0	98.8	97.8	64.5	98.8	77.5
	Overfit	18.0	1.2	1.0	1.0	1.0	1.0
1000	Underfit	0.0	0.0	0.0	0.0	0.0	0.0
	Fit	79.8	99.5	100.0	100.0	100.0	100.0
	Overfit	20.2	0.5	0.0	0.0	0.0	0.0
Number of variables (T)							
30	Underfit	0.0	2.7	33.0	56.3	27.7	47.7
	Fit	80.8	96.5	67.0	43.7	72.3	52.3
	Overfit	19.2	0.8	0.0	0.0	0.0	0.0
60	Underfit	0.0	0.0	0.0	3.5	0.0	1.5
	Fit	85.5	99.3	99.3	95.8	99.3	97.8
	Overfit	14.5	0.7	0.7	0.7	0.7	0.7
Total							
	Underfit	0.0	1.3	16.5	29.9	13.9	24.6
	Fit	83.2	97.9	83.2	69.8	85.8	75.1
	Overfit	16.8	0.8	0.3	0.3	0.3	0.3

mostly improves the performance of the information criteria, reduces the under-fitting and over-fitting for CAIC/BIC and AIC, respectively.

Comparing the results for the two- and three-component model, we observed that the identification of the correct model is easier for $S = 2$, with exception for AIC. However, in general, across our design AIC3 presents a balanced result across the design, and will be used in the selection of the model in our application.

5. APPLICATION

We analyzed contraceptive use dynamics as an illustration of our approach to model-based clustering of sequential data with demographic purposes. First, the data are described; then, we estimate the transition probabilities under the assumption of homogeneity, when the process is represented by a single Markov chain. Heterogeneity is thereafter introduced by allowing more than one subpopulation.

We use the life history calendar (LHC), which is a major and relatively new instrument for the collection of retrospective data (Belli *et al*., 2001). Data for this application come from the Brazil Demographic and Health Survey (BDHS) conducted between March 1996 and June 1996. The BDHS includes a calendar of monthly data on contraceptive use and pregnancy status. More than 20 countries with DHS surveys adopted the LHC. The BDHS is nationally representative, stratified two-stage sample. A total of 12612 women of age 15–49 were interviewed. The calendar covers the period from January 1991 to the month of interview (from March to June 1996). Details of the BDHS are available in the main report (BENFAM and Macro International 1997). We selected the 20–34 age interval at the time of the interview. The reason for this selection is twofold: firstly, to avoid that subpopulations would just be picking women at different stages of the life course, we decided to select a shorter age interval; secondly, there is a decline in contraceptive usage after the age 35 (Dias and Kathun, 2002). We focused on the Northeast region of Brazil that plays an important role in the declining fertility in Brazil (Gupta and Leite, 1999). Therefore, the final sample size is 2228 women.

Table 3 presents the categories defined in the BDHS (BDHS Label). Since this state space is substantially too large to extract meaningful conclusions, and some of the states are related, we aggregate these categories into 5 states: 1 – Non-use of contraception, 2 – Sterilization (Female sterilization, Male sterilization), 3 – Pregnancy (Pregnancy, Birth, Terminated pregnancy/non-live birth), 4 – Pill, and 5 – Other temporary methods (IUD-Intrauterine device, Injection, Diaphragm/foam/jelly, Condom, Periodic abstinence/rhythm, Withdrawal, and

TABLE 3 State Space in DHS Data and Aggregated

State space	DHS label	BDHS code
1 Non-use of contraception		0
2 Sterilization	Female sterilization	6
	Male sterilization	7
3 Pregnancy	Pregnancy	P
	Birth	B
	Terminated pregnancy/ non-live birth	T
4 Pill		1
5 Other temporary methods	IUD	2
	Injections	3
	Diaphragm/foam/jelly	4
	Condom	5
	Periodic abstinence/rhythm	8
	Withdrawal	9
Total	Other traditional methods	W

Other traditional methods). Table 3 summarizes the state space used to model the dynamics of contraceptive use in this application.

We examined the data using a finite mixture of Markov chains to incorporate unobserved heterogeneity. Women are grouped in subpopulations on the basis of similarity of their behavior. First, we address the selection and estimation of the model; we then focus on the interpretation.

We estimated the finite mixture of Markov chains with a different number of subpopulations from 1 to 10, using 20 different starting values to avoid local maxima. If there is a single population, the population is homogeneous and all individuals have the same transition probabilities. From the information criteria presented in Table 4, at least 3 subpopulations have to be included in the model ($S \geq 3$), which corresponds to at least 74 free parameters (Figure 2). However, based on the simulation study presented earlier, we conclude that the selection based on the AIC3 more likely recovers the true dimension of the model. Therefore, we set a solution with 7 subpopulations to obtain a better representation of unobserved heterogeneity.

Tables 5 and 6 present the estimate of the parameters of the model. Results for the homogeneous population (aggregate results: $S = 1$) are reported as well. We first analyze the homogeneous population as a benchmark to show the advantages of the proposed method. For the homogeneous population, more than half of the women in January 1991 were not using any method of contraception, 8.4% were sterilized

TABLE 4 Information Criteria

No. of subpopulations	Log-likelihood	No. of free parameters	Information criteria			
			BIC	AIC	AIC3	CAIC
1	−33251.8	24	66688.7	66551.7	66575.7	66712.7
2	−31959.5	49	64296.7	64017.0	64066.0	64345.7
3	−31608.5	74	63787.5	63365.1	63439.1	63861.5
4	−31505.5	99	63774.2	63209.0	63308.0	63873.2
5	−31414.5	124	63785.0	63077.1	63201.1	63909.0
6	−31379.6	149	63907.9	63057.2	63206.2	64056.9
7	−31329.6	174	64000.5	63007.2	63181.2	64174.5
8	−31294.3	199	64122.6	62986.5	63185.5	64321.6
9	−31262.6	224	64252.0	62973.2	63197.2	64476.0
10	−31216.9	249	64353.4	62931.9	63180.9	64602.4

(or their partner), and 16.6% were using the pill. The survey also shows that 11.4% of the women were pregnant in January 1991. Results for the homogeneous population in Table 6 show a strong persistence of staying in the same state. Indeed, excluding pregnancy ($\hat{a}_{33} = 0.879$), the probability that the process remains in the same state is always greater than 0.95. Note that sterilization is an absorbing state ($\hat{a}_{22} = 1$). This description of the dynamics of contraceptive use is not very informative, because all women are assumed to follow exactly the same pattern over time. Apart from that, the Markov property under the homogeneous population might be problematic in general, and for the pregnancy sequence in particular. However, extending the Markov property incorporating unobserved heterogeneity, the

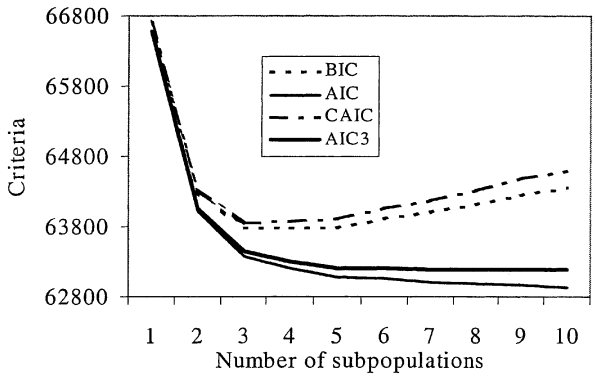

FIGURE 2 Information criteria.

TABLE 5 Estimates of the Initial Distribution and Prior Probabilities

Contraceptive use	Homogeneous population	Heterogeneous population						
		1	2	3	4	5	6	7
Non-use of contraception	0.568	0.876	0.037	0.067	0.023	0.896	0.313	0.625
Sterilization	0.084	0.088	0.173	0.095	0.124	0.084	0.093	0.039
Pregnancy	0.114	0.000	0.057	0.141	0.258	0.020	0.488	0.095
Pill	0.166	0.035	0.473	0.687	0.184	0.000	0.000	0.202
Other temporary methods	0.068	0.001	0.261	0.010	0.412	0.000	0.106	0.039
Subpopulation prior probability	1.000	0.302	0.054	0.096	0.079	0.119	0.104	0.247

states become dependent through the latent variable Z (see Figure 1), relaxing this assumption (see Appendix I).

We consider the case $S = 7$. For the heterogeneous population, the prior probability of each subpopulation ($\hat{\pi}_s$) indicates that the subpopulation size is ranged from 5.4% to 30.2% of the complete sample (Table 5). Subpopulation I, the largest (30.2% of the sample), consists predominantly of non-users of contraception. Indeed, 87.6% of the women in this subpopulation stay in the non-users state in this period, since this state is almost absorbing to this subpopulation (0.996). For 8.8% of the women, the process is similar; women remain in the sterilization state, which is an absorbing state. Thus, from a period perspective, this subpopulation presents stable processes, in which pregnancies are absent. Subpopulation II (5.4% of the sample) represents users of contraception (starting as pill user: 47.3%; other temporary methods: 26.1%; and sterilization: 17.3%). Given the high retention probability, the subpopulation is rather stable. Subpopulation III (9.6% of the sample) represents the intensive users of contraceptive methods. Indeed, only 6.7% of the women in this subpopulation begin without a contraceptive method. Pill users represent the largest group in this subpopulation (68.7%). When women in this subpopulation do not use contraceptive methods, their intention might likely be related to a desired pregnancy. Women in subpopulation IV (7.9% of the sample) corresponds to a more diverse subpopulation which prefers other temporary methods at the beginning (41.2%) and tends to be pregnant (25.8%). Women in subpopulation V (11.9% of the sample) tend to not use contraception at the beginning (89.6%). Subpopulation VI (10.4% of the sample) does not use pill and present 48.8% of the women pregnant at the beginning. Finally,

TABLE 6 Estimates of the Transition Probabilities

Contraceptive use Origin	Destination (1)	(2)	(3)	(4)	(5)
Homogeneous population					
Non-use of contraception (1)	0.958	0.001	0.022	0.012	0.006
Sterilization (2)	0.000	1.000	0.000	0.000	0.000
Pregnancy (3)	0.095	0.015	0.879	0.006	0.005
Pill (4)	0.028	0.002	0.007	0.954	0.009
Other temporary methods (5)	0.019	0.002	0.018	0.011	0.951
Heterogeneous population					
Subpopulation 1					
Non-use of contraception (1)	0.996	0.000	0.001	0.008	0.002
Sterilization (2)	0.000	1.000	0.000	0.000	0.000
Pregnancy (3)	0.052	0.010	0.899	0.015	0.025
Pill (4)	0.019	0.000	0.000	0.967	0.013
Other temporary methods (5)	0.008	0.000	0.024	0.010	0.959
Subpopulation 2					
Non-use of contraception (1)	0.744	0.000	0.256	0.000	0.000
Sterilization (2)	0.000	1.000	0.000	0.000	0.000
Pregnancy (3)	0.000	0.057	0.875	0.029	0.038
Pill (4)	0.017	0.001	0.006	0.971	0.004
Other temporary methods (5)	0.010	0.005	0.019	0.004	0.963
Subpopulation 3					
Non-use of contraception (1)	0.580	0.012	0.196	0.183	0.030
Sterilization (2)	0.000	1.000	0.000	0.000	0.000
Pregnancy (3)	0.096	0.018	0.877	0.006	0.002
Pill (4)	0.026	0.004	0.008	0.955	0.008
Other temporary methods (5)	0.016	0.010	0.027	0.146	0.802
Subpopulation 4					
Non-use of contraception (1)	0.451	0.014	0.141	0.175	0.219
Sterilization (2)	0.000	1.000	0.000	0.000	0.000
Pregnancy (3)	0.106	0.012	0.868	0.005	0.009
Pill (4)	0.010	0.000	0.007	0.964	0.018
Other temporary methods (5)	0.005	0.001	0.016	0.003	0.975
Subpopulation 5					
Non-use of contraception (1)	0.964	0.000	0.017	0.011	0.007
Sterilization (2)	0.000	1.000	0.000	0.000	0.000
Pregnancy (3)	0.099	0.005	0.887	0.007	0.002
Pill (4)	0.090	0.002	0.017	0.880	0.012
Other temporary methods (5)	0.258	0.000	0.065	0.072	0.605
Subpopulation 6					
Non-use of contraception (1)	0.941	0.004	0.046	0.005	0.003
Sterilization (2)	0.000	1.000	0.000	0.000	0.000
Pregnancy (3)	0.094	0.028	0.867	0.007	0.004
Pill (4)	0.053	0.016	0.000	0.888	0.043
Other temporary methods (5)	0.024	0.002	0.008	0.006	0.960
Subpopulation 7					
Non-use of contraception (1)	0.896	0.004	0.057	0.030	0.013
Sterilization (2)	0.000	1.000	0.000	0.000	0.000
Pregnancy (3)	0.106	0.009	0.882	0.002	0.001
Pill (4)	0.031	0.000	0.006	0.957	0.006
Other temporary methods (5)	0.028	0.000	0.017	0.009	0.946

TABLE 7 Characterization of the Identified Subpopulations (Percentages)

Background characteristics	Homogeneous population	Heterogeneous population						
		1	2	3	4	5	6	7
Age of respondent (***)								
20–24	37.21	43.07	14.71	27.78	20.65	52.24	20.99	40.73
25–29	33.17	26.24	40.44	38.33	38.71	26.12	46.96	37.09
30–34	29.62	30.69	44.85	33.89	40.65	21.63	32.04	22.18
Place of residence (***)								
Capital, large city	32.27	34.16	46.32	29.44	38.71	29.80	21.55	29.64
Small city	17.24	19.68	14.71	22.78	18.71	12.65	17.13	13.96
Town	24.01	24.38	20.59	28.33	17.42	27.76	22.10	23.71
Countryside	26.48	21.78	18.38	19.44	25.16	29.80	39.23	32.70
Education of respondent (***)								
No education	7.76	6.31	4.41	5.00	6.45	5.31	16.57	10.33
Primary	35.41	30.69	28.68	41.11	34.19	36.73	42.54	39.77
Secondary	53.14	57.92	64.71	50.56	52.90	53.88	40.33	47.80
Higher	3.68	5.07	2.21	3.33	6.45	4.08	0.55	2.10
Current marital status (***)								
Never married	26.93	58.29	1.47	2.22	7.74	25.31	4.97	7.65
Married	41.88	24.75	63.24	63.33	60.65	35.92	43.65	52.01
Living together	21.95	10.52	24.26	28.89	24.52	23.27	37.57	29.83
Widowed	0.72	0.87	0.74	0.00	0.00	0.41	0.55	1.15
Divorced	0.31	0.00	0.74	0.00	1.29	0.82	0.00	0.38
Not living together	8.21	5.57	9.56	5.56	5.81	14.29	13.26	8.99
Occupation of respondent (***)								
Not working	42.71	36.15	41.18	51.67	35.48	45.90	51.11	47.89
Prof., Tech., Manag.	9.14	12.92	10.29	7.22	12.90	7.38	4.44	4.98
Clerical	8.64	10.43	8.09	8.89	11.61	8.20	6.67	5.94
Sales	12.42	13.54	19.12	11.67	13.55	6.97	9.44	12.45
Agriculture, self employed	6.57	6.09	1.47	3.33	7.74	5.33	12.22	8.05
Agriculture, employee	0.18	0.12	0.00	0.00	0.65	0.41	0.00	0.19
Household and domestic	10.35	11.80	6.62	6.67	6.45	16.39	9.44	9.00
Services	4.95	4.22	5.15	3.89	7.74	6.15	3.33	5.56
Skilled manual	1.17	1.49	1.47	1.11	1.94	0.41	1.11	0.77
Unskilled manual	3.87	3.23	6.62	5.56	1.94	2.87	2.22	5.17

$^{***}p < 0.001$.

subpopulation VII, the second largest group (24.7%), consists largely of non-users of contraception at the beginning (62.5%), and tends to switch to the pregnancy state (0.057).

Once the subpopulations are identified—and we did not know in advance their number and behaviors—each subpopulation is characterized through observable variables, which are needed for the setting of specific communication strategies and family planning services. A large number of subpopulations (S) allows a more specific targeting. Sometimes, there are good proxy variables for the unobserved heterogeneity as this application shows. Variables are age, region,

place of residence, education, current marital status, and occupation of the women. Women were allocated to the subpopulations based on the posterior probability (optimal Bayesian classification) given by Eq. (10). We used chi-square tests (Pearson's statistic) to test the level of association between the classification of women into these three subpopulations and the background variables. All these variables are statistically related with the classification rule ($p < 0.001$). The results are summarized in Table 7. Subpopulation I is characterized mostly by younger women from urban places with secondary or high education, never married with high qualified occupation. Subpopulation II is characterized mostly by older married women from large cities, with secondary education. Woman in Subpopulation III tend to be in the 25–29 age range from small cities or towns and with primary or secondary education. They are married and they do not work. Subpopulation IV consists of older married women from large cities with secondary or high education. Subpopulation V tends to be rural with household and domestic jobs. They are younger with primary and secondary education. Subpopulation VI is rural with low education and mostly not working. Finally, subpopulation VII tends to be younger with low education and married. They live mostly in rural areas. We conclude that age, place of residence, education, and current marital status discriminate these subpopulations. The occupation of respondent is not so strong in discriminating the subpopulations, but it helps in giving a more precise picture of the segments. It should be noted that other background variables not considered in this application might play a role in characterizing these subpopulations as proxies of the unobserved heterogeneity.

These results have implications for family planning policies and programs. This illustration demonstrates the advantage of using finite mixture models for describing dynamic patterns of behavior that helps in developing subpopulation-tailored family planning campaigns. It is an aid in identifying subpopulations with special needs that need to be addressed using different programs.

6. CONCLUSION

We provide a new approach to modeling demographic sequential data incorporating unobserved heterogeneity. Because of the dependency over time, heuristic procedures such as cluster analysis are not appropriate. Finite mixtures as a model-based clustering procedure provide an attractive alternative whenever one needs to identify subpopulations with heterogeneous behavior. In particular, we propose the analysis of discrete sequential data using a finite mixture of Markov chains.

The adoption of a contraceptive method, contraceptive switching, and discontinuation of contraceptive practice involve transitions to new contraceptive use states. Many family planning programs are directed towards these transitions. To be effective, programs should treat people with different behavior differently. Model-based clustering techniques are instruments for designing differential policies and programs.

We applied this model to the contraceptive use calendar of the Brazilian Demographic and Health Survey 1996 to identify groups of women with similar contraceptive use patterns. We found seven subpopulations with differential contraceptive use and dynamics. The seven subpopulations were described further using background characteristics.

APPENDIX I

Let us show that a mixture of Markov chains is not a Markov chain. As before, let x_{it} and $\tilde{\mathbf{x}}_{it} = (x_{i1}, \ldots, x_{i,t-1)}$. Then, we have

$$\begin{aligned} p(x_{it}|\tilde{\mathbf{x}}_{it}; \varphi) &= \sum_{s=1}^{s} p(x_{it}, z_s|\tilde{\mathbf{x}}_{it}; \varphi) \\ &= \sum_{s=1}^{s} p(z_s|\tilde{\mathbf{x}}_{it}; \varphi) p(x_{it}|z_s, \tilde{\mathbf{x}}_{it}; \varphi). \end{aligned} \tag{8}$$

Because, it is assumed that within each component (given z_s) the process is Markovian,

$$\begin{aligned} p(x_{it}|z_s, \tilde{\mathbf{x}}_{it}; \varphi) &= p(x_{it}|z_s, x_{i,t-1}; \varphi), \\ p(x_{it}|\tilde{\mathbf{x}}_{it}; \varphi) &= \sum_{s=1}^{s} p(z_s|\tilde{\mathbf{x}}_{it}; \varphi) p(x_{it}|z_s, x_{i,t-1}; \varphi). \end{aligned} \tag{9}$$

Therefore, the process is no longer Markovian and can mimic complex sequential patterns through the latent variable. Note that the same conclusion can be reached by Figure 1; that is, conditional on z, the variables are independent from the past; unconditional on z (finite mixture model) all the past models x_{it} through z.

APPENDIX II

The EM algorithm proceeds as follows. Let us assume that we have obtained an approximation $\varphi^{(h)}$ to estimate φ given by $\varphi^{(h)} = (\pi_1^{(h)}, \ldots, \pi_{S-1}^{(h)}, \theta_1^{(h)}, \ldots, \theta_S^{(h)})$. The general objective is that the next

estimate $\varphi^{(h+1)}$ will be closer to $\hat{\varphi}$. The auxiliary Q function, defined as $Q(\varphi;\varphi^{(h)}) = E[\log p(\mathbf{Z},\mathbf{X}|\varphi)|\mathbf{X}=\mathbf{x},\varphi^{(h)}]$, corresponds to the conditional expectation of the missing variable $\mathbf{Z}$ given observed data and parameters, where $p(\mathbf{Z},\mathbf{X}|\varphi)$ is the density of the complete data. For the E-step, we need to compute $Q(\varphi;\varphi^{(h)}) = \sum_{i=1}^{n}\sum_{s=1}^{S} E[Z_{is}|\mathbf{X}_i=\mathbf{x}_i,\varphi^{(h)}]$ $\log(\pi_s f_s(\mathbf{x}_i;\theta_s))$. Let

$$\alpha_{is}^{(h+1)} = E[Z_{is}|\mathbf{X}_i=\mathbf{x}_i,\varphi^{(h)}] = \frac{\pi_s^{(h)} f_s(\mathbf{x}_i;\theta_s^{(h)})}{\sum_{r=1}^{S}\pi_r^{(h)} f_r(\mathbf{x}_i;\theta_r^{(h)})} \tag{10}$$

be the conditional expectation of Z_{is} at the $(h+1)^{th}$ iteration. Then, we obtain

$$\begin{aligned} Q(\varphi;\varphi^{(h)}) = &\sum_{i=1}^{n}\sum_{s=1}^{S}\alpha_{is}^{(h+1)}\log\pi_s \\ &+\sum_{i=1}^{n}\sum_{s=1}^{S}\alpha_{is}^{(h+1)}\log f_s(\mathbf{x}_i;\theta_s). \end{aligned} \tag{11}$$

For the M-step, we have to maximize (11) over φ. At the $(h+1)^{th}$ iteration in the M-step, we choose the value of φ, say $\varphi^{(h+1)}$, which maximizes $Q(\varphi;\varphi^{(h)})$, resulting $\pi_s^{(h+1)} = \frac{1}{n}\sum_{i=1}^{n}\alpha_{is}^{(h+1)}$ and $\theta_s^{(h+1)} = \arg\max_{\theta_s}\sum_{i=1}^{n}\alpha_{is}^{(h+1)}\log f_s(\mathbf{x}_i;\theta_s)$, for $s=1,\ldots,S$. Under suitable regular conditions, $\{\varphi^{(h)}\}$ converges to a stationary point of $\ell(\varphi;\mathbf{x})$ (Dempster *et al.*, 1977; McLachlan and Krishnan, 1997). The convergence of the EM algorithm as well as other iterative algorithms means convergence to a local maximum of $\ell(\varphi;\mathbf{x})$.

For our model, the estimate of the parameters can be found using the EM iterative process defined by $\alpha_{is}^{(h+1)}$, given by (10), and

$$\begin{aligned} \pi_s^{(h+1)} &= \frac{1}{n}\sum_{i=1}^{n}\alpha_{is}^{(h+1)} \\ \lambda_{sj}^{(h+1)} &= \frac{\sum_{i=1}^{n}\alpha_{is}^{(h+1)} I(x_{i0}=j)}{\sum_{i=1}^{n}\alpha_{is}^{(h+1)}} \\ a_{sjk}^{(h+1)} &= \frac{\sum_{i=1}^{n}\alpha_{is}^{(h+1)} n_{ijk}}{\sum_{r=1}^{K}\sum_{i=1}^{n}\alpha_{is}^{(h+1)} n_{ijr}}. \end{aligned} \tag{12}$$

Thus, (10) and (12) define the EM iterative process to estimate this model. For $S=1$, we have the homogeneous population with a unique Markov chain, and these results reduce to the maximum likelihood estimates of a single Markov chain (Bishop *et al.*, 1975). The EM algorithm for this model was programmed in MATLAB 6.5 (MathWorks, 2002).

REFERENCES

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control AC-19*: 716–723.

Belli, R.F., Shay, W.L., and Stafford, F.P. (2001). Event history calendars and question list surveys. *Public Opinion Quarterly 65*: 45–74.

BENFAM and Macro international (1997). *Pesquisa Nacional Sobre Demografia e Saúde (PNDS), Brasil, 1996.* Brasil: Rio de Janeiro BENFAM/Macro International.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Massachusetts: The MIT Press.

Blumen, I., Kogan, M., and McCarth, P.J. (1955). *The Industrial Mobility of Labor Relations.* Ithaca, NY: Cornell University Press.

Böckenholt, U. (1993). A latent class regression approach for the analysis of recurrent choice data. *British Journal of Mathematical and Statistical Psychology 46*: 95–118.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika 52*: 345–370.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In O. Opitz, B. Lausen, and R. Klar (Eds.), *Information and Classification, Concepts, Methods and Applications.* Berlin: Springer-Verlag, pp. 40–54.

Chen, H., Chen, J., and Kalbfleisch, J.D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of Royal Statistical Society B 63*: 19–29.

Clogg, C.C. (1995). Latent class models. In G. Arminger, C.C. Clogg, and M.E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences.* New York: Plenum Press, pp. 311–359.

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

DeSarbo, W.S., Lehmann, D.R., and Hollman, F.G. (2004). Modeling dynamic effects in repeated-measures experiments involving preference/choice: An illustration involving stated preference analysis. *Applied Psychological Measurement 28*: 186–209.

Dias, J.G. and Kathun, M. (2002). Modelling the choice of contraceptive methods incorporating unobserved heterogeneity. *Working paper*, SOM Research Centre, University of Groningen.

Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics 14*: 755–763.

Everitt, B.S. (1987). *Introduction to Optimization Methods and their Application in Statistics.* London: Chapman and Hall.

Fuchs, C. and Greenhouse, J.B. (1988). The EM algorithm for maximum likelihood estimation in the mover-stayer model. *Biometrics 44*: 605–613.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika 61*: 215–231.

Guha, D. and Banerji, A. (1998/1999). Testing for regional cycles: A Markov-switching approach. *Journal of Economics and Social Measurement 25*: 163–182.

Gupta, N. and Leite, I.C. (1999). Adolescent fertility behavior: Trends and determinants in Northeastern Brazil. *International Family Planning Perspectives 25*: 125–130.

Haughton, D. and Haughton, J. (1996). Using a mixture model to detect son preference in Vietnam. *Journal of Biosocial Sciences 28*: 355–65.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica 52*: 271–320.

Islam, M.A. (1994). Multistate survival models for transitions and reverse transitions: An application to contraceptive use data. *Journal of the Royal Statistical Society 157*: 441–456.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics 27*: 887–906.

Kost, K. (1993). The dynamics of contraceptive use in Peru. *Studies in Family Planning 24*: 109–119.

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

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica 47*: 939–956.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data.* Cambridge: Cambridge University Press.

Lazarsfeld, P.F. and Henry, N.W. (1968). *Latent Structure Analysis.* New York: Houghton Mifflin.

Li, L. and Choe, M.K. (1997). A mixture model for duration data: Analysis of second births in China. *Demography 34*: 189–197.

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

Lindsay, B.G. (1983b). The geometry of mixture likelihoods. Part II: The exponential family. *The Annals of Statistics 11*: 783–792.

Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry and Applications.* CA: IMS Hayward.

Lindsay, B.G. and Lesperance, M.L. (1995). A review of semiparametric mixture models. *Journal of Statistical Planning and Inference 47*: 29–39.

Lo, Y., Mendell, N.R., and Rubin, D.B. (2001). Testing the number of components in a normal mixture. *Biometrika 88*: 767–778.

Math Works (2002). *MATLAB 6.5. Natick.* MA: The Math Works Inc.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models.* New York: John Wiley and Sons.

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* New York: John Wiley and Sons.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics 8*: 343–366.

Norris, J.R. (1997). *Markov Chains.* Cambridge: Cambridge University Press.

Poulsen, C.S. (1990). Mixed Markov and latent Markov modeling applied to brand choice behavior. *International Journal of Research in Marketing 72: 5–19.*

Rabiner, L.R. and Juang, B.H. (1986). An introduction to hidden Markov models. *IEEE Acoustics, Speech and Signal Processing Magazine 3*: 4–16.

Ramaswamy, V., DeSarbo, W.S., Reibstein, D.J., and Robinson, W.T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science 12*: 103–124.

Robbins, H. (1950). A generalization of the method of maximum likelihood: Estimating a mixing distribution (abstract). *The Annals of Mathematical Statistics 12*: 314–315.

Rosbergen, E., Pieters, R., and Wedel, M. (1997). Visual attention to advertising: A segment-level analysis. *Journal of Consumer Research 24*: 305–314.

Ross, S.M. (2000). *Introduction to Probability Models (7th Edition).* San Diego: Harcourt/Academic Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*: 461–464.

Steele, F. and Diamond, I. (1999). Contraceptive switches in Bangladesh. *Studies in Family Planning 30*: 315–328.

Taylor, H.M. and Karlin, S. (1994). *An Introduction to Stochastic Modeling (Revised Edition).* San Diego: Academic Press.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: John Wiley and sons.

Vaupel, J.W. and Yashin, A.I. (1985). The deviant dynamics of death in heterogeneous populations. In N.B. Tuma (Ed.), *Sociological Methodology.* London: Jossey-Bass, pp. 176–185.

Vaupel, J.W., Manton, K.G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*: 439–454.

Wedel, M., and Kamakura, W. (1999). *Market Segmentation. Conceptual and Methodological Foundations.* Boston: Kluwer Academic Publishers.

Wedel, M., DeSarbo, W.S., Bult, J.R., and Ramaswamy, V. (1993). A latent class Poisson regression model for heterogeneous count data. *Journal of Applied Econometrics 8*: 397–411.

Willekens, F.J. (1999). Modeling approaches to the indirect estimation of migration flows: from entropy to EM. *Mathematical Population Studies 7*: 239–278.