 DOI: 10.1111/biom.12224

A Discrete Time Event-History Approach to Informative Drop-Out in Mixed Latent Markov Models with Covariates

Francesco Bartolucci[1,*] and Alessio Farcomeni[2]

[1]Department of Economics, University of Perugia (IT), Perugia, Italy
[2]Department of Public Health and Infectious Diseases, Sapienza University of Rome (IT), Rome, Italy
[*]*email:* bart@stat.unipg.it

Summary. Mixed latent Markov (MLM) models represent an important tool of analysis of longitudinal data when response variables are affected by time-fixed and time-varying unobserved heterogeneity, in which the latter is accounted for by a hidden Markov chain. In order to avoid bias when using a model of this type in the presence of informative drop-out, we propose an event-history (EH) extension of the latent Markov approach that may be used with multivariate longitudinal data, in which one or more outcomes of a different nature are observed at each time occasion. The EH component of the resulting model is referred to the interval-censored drop-out, and bias in MLM modeling is avoided by correlated random effects, included in the different model components, which follow common latent distributions. In order to perform maximum likelihood estimation of the proposed model by the expectation–maximization algorithm, we extend the usual forward-backward recursions of Baum and Welch. The algorithm has the same complexity as the one adopted in cases of non-informative drop-out. We illustrate the proposed approach through simulations and an application based on data coming from a medical study about primary biliary cirrhosis in which there are two outcomes of interest, one continuous and the other binary.

Key words: Discrete latent variables; Expectation–maximization algorithm; Hidden Markov models; Shared-parameter models.

1. Introduction

In longitudinal studies, subjects may be lost to follow-up due to events, such as death, which are associated with the outcome of interest. In these cases an informative drop-out arises that must be properly modeled. From the reverse perspective, the time trend of a longitudinal measurement may predict the risk of an event (e.g., a steadily decreasing CD4 count is predictive of adverse events in HIV patients); see, for instance, Follmann and Wu (1995) for a general account of related longitudinal and survival processes.

A common approach to deal with informative drop-out is via shared-parameter models (e.g., Wu and Carrol, 1988; Follmann and Wu, 1995), where both longitudinal and survival mechanisms are assumed to share a latent Gaussian variable. Random effects corresponding to latent variables having a discrete distribution, along the lines of this work, are adopted by Roy (2003) to deal with an ordinal latent class model. Another approach to model informative drop-out is that of Wulfsohn and Tsiatis (1997) and Rizopoulos (2010), where the risk of an event at a time is influenced by the expected value of the longitudinal response at the same time. The resulting Joint Model (JM) uses both fixed and random effects in the hazard function. There are very few generalizations of JMs to the case of discrete longitudinal outcomes. Notable exceptions are those of Rizopoulos and Ghosh (2011), who propose generalized linear models in a Bayesian framework, and Viviani, Alfó, and Rizopoulos (2014), who rely on a classical maximum likelihood framework.

A limitation of shared-parameter models and JMs is that latent variables, in the form of subject-specific parameters, are time constant. The effect of time is usually captured by a fixed function of time (usually, polynomial) that must be pre-specified and the deviation of the behavior of a subject with respect to the average behavior may not change during the period of observation. For an exception see Henderson, Diggle, and Dobson (2000).

Latent Markov (LM) models represent a flexible and convenient way of modeling outcomes of a different nature which are repeatedly measured over time; see Bartolucci, Farcomeni, and Pennoni (2013) for an overview. The basic assumption of these models is that the response variables, which are longitudinally observed, are conditionally independent given a hidden first-order Markov chain which accounts for the possibly time-varying unobserved heterogeneity. This approach is extended by also including time-constant individual effects having a discrete distribution, giving rise to the class of mixed latent (or hidden) Markov (MLM) models as originally proposed by Altman (2007); for a review see Maruotti (2011).

Despite the relevance of LM and MLM models, there are very few extensions of these models that can deal with informative drop-out. We consider, in particular, that of Albert (2000), which jointly models the outcome and missing mechanisms. The latter is assumed to follow a manifest (not latent) first-order Markov chain, and the two processes are linked because the outcome is used to model the missingness indicators.

In this article, we propose a different approach with respect to the ones mentioned above to extend MLM models for informative drop-out, in which the manifest distribution is jointly referred to the longitudinal and drop-out processes;

the corresponding time-varying unobserved heterogeneity structure evolves according to the same initial and transition distributions. Our approach falls into the class of non-ignorable random-coefficient-based drop-out models as defined in Little (1995). The proposed model can be characterized as a selection model, where instead of directly modeling the drop-out mechanism as a function of unobserved responses as in Diggle and Kenward (1994), we do this indirectly as in the seminal article of Wu and Carrol (1988). Note that some authors (e.g., Creemers et al., 2010; Viviani, Rizopoulos, and Alfó, 2014) consider shared-parameter models as a separate framework. Selection models for discrete longitudinal data have been proposed, among others, by Molenberghs, Kenward, and Lesaffre (1997) and Ten Have et al. (1998). The advantages of shared-parameter models like the one we propose are that a directly interpretable marginal model is obtained for the observed outcomes, even if at the price of a heavier estimation procedure. In contrast, a pattern mixture model (e.g. Wu and Bailey, 1989; Little and Wang, 1996) is simpler to estimate but specifies the outcome distribution only conditionally on the time of drop-out.

As a motivation, we consider an example where patients with primary biliary cirrhosis were randomized either to a placebo or a treatment based on D-penicillamine. Outcomes were repeatedly measured after randomization but only 1% of patients have a complete record. In this application it is reasonable to expect drop-out to be informative when it is due to death related to the illness or to transplant. As in the MLM formulation, we account for two different types of unobserved heterogeneity. The first is time-constant and is represented by a latent variable having a discrete distribution with a suitable number of support points. The second is time-variant and is represented by a hidden Markov chain with a suitable number of states.

In the proposed Missing Not at Random (MNAR) approach, the longitudinal outcomes are modeled through a generalized linear mixed effects parametrization (e.g., Fitzmaurice, Laird, and Ware, 2004) and a discrete event-history (EH) model (see Steele, 2011, and references therein) is used for the drop-out process. We also allow for multivariate longitudinal data, so that more than one outcome, and of a different nature, can be considered at each time occasion. In the application, a continuous and a binary variable were observed at each time occasion. These outcomes are assumed to be conditionally independent given the latent variables, even if we outline how to relax this assumption. A flexible dependence structure is obtained as the random effects are distributed according to a single first-order latent Markov chain with a finite number of states, additionally to the discrete latent variable for the time-constant unobserved heterogeneity. Subjects in the same latent group share class-specific intercepts for the longitudinal models, and a class-specific intercept for the EH model, and they share common regression coefficients for the covariates. The resulting estimates are easily interpretable, and the model is reasonable as it is natural to expect that longitudinal outcomes and drop-out share the same sources of unobserved heterogeneity, which can have different effects on each of them.

Note that our proposal operates in discrete time, and not in continuous time, and therefore it should be used for longitudinal data where the missingness pattern consists of informative drop-out, or when the time-to-event is interval censored between two measurement occasions. The model can in principle be used also with right-censored data, but at the price of a loss of information and efficiency. The discrete time approach is simpler than the joint modeling one. Its interpretation is rather different given that joint longitudinal survival models are typically formulated in terms of hazard rate.

For the proposed model we perform maximum likelihood estimation by an expectation–maximization (EM) algorithm (Baum et al., 1970; Dempster, Laird, and Rubin, 1977). This requires an extension of the forward–backward recursions (Baum et al., 1970; Welch, 2003; Bartolucci et al., 2013) to account for informative drop-out. We also pay attention to the computation of the standard errors for the parameter estimates by employing a method proposed in Bartolucci and Farcomeni (2009).

The remainder of the article is organized as follows. In Section 2 we illustrate the proposed class of models. In Section 3 we describe likelihood inference for these models. Finally, the approach is illustrated in Section 4 through the analysis of primary biliary cirrhosis data and we provide some concluding remarks in Section 5. `R` code with the methods developed in this article, and the results of a brief simulation study, are available as Supplementary Material.

2. MLM Models with Informative Drop-Out

We consider a longitudinal study on a sample of n subjects, or more generally sample units, in which s_i follow-up time occasions are scheduled for $i = 1, \ldots, n$. We assume that s_i is known in advance for each subject, while if it is not known in advance all proposed inference can be thought of as being conditional on the number of follow-up times as is clarified in the following.

For each subject i and occasion j we observe r response variables, denoted by Y_{hij}, $h = 1, \ldots, r$; we also denote by J_i the last time occasion of observation for subject i, so that drop-out occurs before occasion $J_i + 1$. Also J_i is a random variable, the distribution of which depends on observable covariates and latent variables for the unobserved heterogeneity. A realization of the hth response variable is denoted by y_{hij} and, accordingly, a realization of J_i is denoted by j_i. The observed outcomes for the same subject i and occasion j are collected in the column vector $\boldsymbol{y}_{ij} = (y_{1ij}, \ldots, y_{rij})'$. Also note that if there is drop-out, then $j_i < s_i$, whereas $j_i = s_i$ indicates that a complete record of outcomes is observed for the ith subject. Then, if s_i is not known in advance we fix s_i equal to j_i for those subjects that do not drop out, whereas we can fix s_i equal to an arbitrary value greater than j_i for those subjects characterized by drop-out. This drop-out is informative, as it is assumed to depend on the latent variables affecting the responses.

Let D_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, s_i$, denote a binary random variable equal to 1 if subject i drops out from the study after occasion j and before occasion $j + 1$, that is, $J_i = j$, and to 0 otherwise, with $D_{ij} = 1$ implying that $D_{i,j+1} = \cdots = D_{i,s_i} = 1$ for $j < s_i$. The basic assumption of the proposed model is that, given the discrete latent variable U_i with k_1 support points, the discrete latent variable V_{ij} with k_2 support points, and the available covariates, the response variables $Y_{1ij}, \ldots, Y_{rij}$ are

conditionally independent and they are also independent of D_{ij}. We denote by $\boldsymbol{x}_{hij}$ the column vector of covariates affecting Y_{hij} and by $\mu_{hij}(u,v)$ the conditional expected value of this response variable given these covariates, $U_i = u$, and $V_{ij} = v$. Similarly, we denote by $\boldsymbol{z}_{ij}$ the column vector of covariates affecting D_{ij} and by $p_{ij}(u,v)$ the conditional probability that $D_{ij} = 1$ given these covariates, $U_i = u$, and $V_{ij} = v$. Then, for $u = 1, \ldots, k_1$ and $v = 1, \ldots, k_2$, we assume that

$$\begin{cases} g_h\{\mu_{hij}(u,v)\} = \alpha_{hu}^{(1)} + \alpha_{hv}^{(2)} + \boldsymbol{x}_{hij}'\boldsymbol{\beta}_h, & h = 1, \ldots, r,\ j = 1, \ldots, s_i, \\ \text{logit}\{p_{ij}(u,v)\} = \gamma_u^{(1)} + \gamma_v^{(2)} + \boldsymbol{z}_{ij}'\boldsymbol{\delta}, & h = 1, \ldots, j_i - 1, \\ p_{ij}(u,v) = 1, & j = j_i + 1, \ldots, s_i, \end{cases}$$

where $g_h(\cdot)$, $h = 1, \ldots, r$, are appropriate link functions and each Y_{hij} is assumed to follow a conditional distribution belonging to the regular exponential family. The reason why $p_{is_i}(u,v) = 1$ is that the follow-up surely stops after occasion s_i for all subjects. Note that the covariates in $\boldsymbol{x}_{hij}$ and $\boldsymbol{z}_{ij}$ are considered as given and fixed and then these vectors will be not explicitly indicated when we express conditional probability and density functions given the covariates.

The marginal model for D_{ij} based on the above assumptions represents the *missing data mechanism*, which is therefore specified conditionally on observed covariates and indirectly on unobserved outcomes. The underlying assumption is that all information about unobserved outcomes is summarized by the random effects, as usual in shared-parameter models approaches. The assumption that outcomes are conditionally independent given the latent variables and covariates is typically not restrictive given that we formulate an MLM model; see for instance Ip et al. (2013). In any case, this assumption can be easily relaxed when all outcomes are categorical (Bartolucci and Farcomeni, 2009), using a marginal parameterization based on logits and log-odds ratios. When all outcomes are continuous, one could simply model them with a multivariate normal distribution, possibly with structural assumptions on the covariance matrix. These extensions are straightforward and involve minor changes to the EM algorithm outlined in the following section.

In the dataset used to illustrate the proposed approach (see Section 4) there are $r = 2$ response variables. The first of these variables is continuous, modeled by a Normal distribution, and the second is binary, modeled by a Bernoulli distribution. Hence, for this application we choose $g_1(\cdot)$ as the identity function and $g_2(\cdot)$ as the logit link function; for the distribution of the first variable we also introduce a dispersion parameter indicated, in general, by σ_1^2. The model for each longitudinal outcome is a classical generalized linear mixed effects model, while the time to drop-out follows an inhomogeneous geometric distribution as in classical discrete time EH models (e.g., Steele, 2011).

We regard drop-out as a trial within each time interval. The resulting likelihood is that of a Bernoulli model; consequently, we have

$$\begin{aligned} &\Pr(J_i = j_i \mid U_i = u, V_{i1} = v_1, \ldots, V_{ij_i} = v_{j_i}) \\ &\quad = p_{ij_i}(u, v_{j_i}) \prod_{j=1}^{j_i - 1} \{1 - p_{ij}(u, v_j)\}, \end{aligned}$$

where, since $p_{is_i}(u,v) = 1$, the probability $p_{ij_i}(u, v_{j_i})$ disappears when there is no drop-out until the end of the study ($j_i = s_i$). This recovers the *truncated* inhomogeneous geometric distribution given that s_i is finite. In order to derive the results in Section 3 it is also important to note that

$$\begin{aligned} &\Pr(J_i > j \mid U_i = u, V_{i1} = v_1, \ldots, V_{ij} = v_j) \\ &\quad = \prod_{l=1}^{j} \{1 - p_{il}(u, v_l)\}, \quad j = 1, \ldots, s_i, \end{aligned}$$

and that

$$\begin{aligned} &\Pr(J_i = j_i \mid J_i > j, U_i = u, V_{i,j+1} = v_{j+1}, \ldots, V_{ij_i} = v_{j_i}) \\ &\quad = p_{ij_i}(u, v_{j_i}) \prod_{l=j+1}^{j_i - 1} \{1 - p_{il}(u, v_l)\} \end{aligned}$$

for $j = 1, \ldots, j_i - 1$. Note that the only information that is used in the model is the interval censored event time (i.e., that drop-out occurs between occasion j_i and j_{i+1}). This is a limitation of the proposed approach when exact event times are known, as some information is disregarded. The use of exact (right-censored) event times would anyway lead to a more complex inferential strategy.

When continuous-time durations are grouped into discrete intervals, a continuous-time hazard model (with constant hazard within each interval) would lead to the complementary log–log link: $g\{p_{ij}(u,v)\} = \log[-\log\{1 - p_{ij}(u,v)\}]$; see Kalbfleisch and Prentice (2002). In our application we use a logit link, which we find more convenient. Regarding the latent structure of the model, we assume that, for $i = 1, \ldots, n$, U_i is independent of the sequence $V_{i1}, \ldots, V_{is_i}$. In particular, the first variable has a distribution based on the mass probabilities $\lambda_u = \Pr(U_i = u)$, $u = 1, \ldots, k_1$, which are collected in the column vector $\boldsymbol{\lambda}$. Moreover, the sequence $V_{i1}, \ldots, V_{is_i}$ follows a Markov chain with initial probabilities $\pi_v = \Pr(V_{i1} = v)$, $v = 1, \ldots, k_2$, which are collected in the column vector $\boldsymbol{\pi}$, and time-homogeneous transition probabilities $\pi_{\overline{v}v} = \Pr(V_{ij} = v \mid V_{i,j-1} = \overline{v})$, $\overline{v}, v = 1, \ldots, k_2$, which are collected in the transition matrix $\boldsymbol{\Pi}$.

Non-homogeneous distributions can be considered by reparameterizing also the latent variables. A very general formulation is as follows:

$$\log \frac{\lambda_{iu}}{\lambda_{i1}} = \tau_{u,\lambda} + \boldsymbol{x}_{i1,\lambda}'\boldsymbol{\psi}_{u,\lambda}, \quad u = 2, \ldots, k_1, \tag{1}$$

$$\log \frac{\pi_{iv}}{\pi_{i1}} = \tau_{v,\pi} + \boldsymbol{x}_{i1,\pi}'\boldsymbol{\psi}_{v,\pi}, \quad v = 2, \ldots, k_2, \tag{2}$$

$$\log \frac{\pi_{i\overline{v}v}^{(t)}}{\pi_{i\overline{v}\overline{v}}^{(t)}} = \tau_{\overline{v}v,\Pi} + \boldsymbol{x}_{ij,\Pi}'\boldsymbol{\psi}_{\overline{v}v,\Pi}, \quad \overline{v} = 1, \ldots, k_2,\ v \neq \overline{v},\ t > 1, \tag{3}$$

where $\tau_{u,\lambda}$, $\tau_{v,\pi}$, and $\tau_{\overline{v}v,\Pi}$ are class and state specific intercepts, $\boldsymbol{\psi}_{u,\lambda}$, $\boldsymbol{\psi}_{v,\pi}$, and $\boldsymbol{\psi}_{\overline{v}v,\Pi}$ are vectors of parameters, and $\boldsymbol{x}_{i1,\lambda}$, $\boldsymbol{x}_{i1,\pi}$, and $\boldsymbol{x}_{ij,\Pi}$ are corresponding vectors of covariates. The parameterization above is particularly useful when (*i*)

measurement times are irregularly spaced, so that $\boldsymbol{x}_{ij,\Pi}$ can include the time difference between the jth and $(j-1)$th occasion (Bartolucci, Lupparelli, and Montanari, 2009) and (*ii*) when the hidden distribution is of main interest. In this second case it is not common to include covariates in the manifest distribution for ease of interpretation. On the other hand, in the proposed application we include covariates in the measurement model, that is, in the conditional distribution of the response variables given the latent variables; this is because we want to evaluate the direct effect of these covariates on the outcomes (see Section 4). Among these covariates we also include the time interval between consecutive occasions of observation, and then we rely on a time-homogeneous hidden Markov chain with initial distribution constrained to be equal to the stationary distribution. This gives rise to a more parsimonious model and estimation results which are easier to interpret with respect to the case of free initial probabilities.

The degree of dependence between the longitudinal and the survival processes is measured by the total variation of $\gamma_u^{(1)}$ and $\gamma_v^{(2)}$. Accordingly, one can perform a sensitivity analysis by comparing estimates for increasing values of k_1 and k_2, where larger values are approximately associated with more sensitivity to dropout within the model class. In this analysis we also consider MAR versions of the proposed approach, hence a direct comparison between results under NMAR and under MAR is obtained within the class of models. It shall be noted that suitable analyses are often performed to assess sensitivity to untestable assumptions (e.g., Creemers et al., 2010); this is slightly different with respect to our sensitivity analysis. However, in our experience when k_1 and k_2 are too large the likelihood becomes near flat, and hence $\gamma_u^{(1)}$ and $\gamma_v^{(2)}$ are weakly identifiable.

The link between the random effects in the different outcomes and the indicator for missingness is based on the assumption that they follow the same latent Markov chain. This is similar in spirit to situations in which a copula is used to model the dependence of random effects in the longitudinal and survival processes, as in Rizopoulos, Verbeke, and Molenberghs (2008).

3. Likelihood Inference

In this section, we illustrate how to perform likelihood inference for the proposed class of models. The presentation is for the case in which the hidden Markov chain is time homogeneous and with parameters common to all sample units. However, the extension to the case in which covariates also affect the time-varying and/or the time-constant latent distribution, possibly based on (1), (2), and (3), may be implemented in a rather simple way following the methods described in Bartolucci et al. (2013), Chapter 5.

We start from considering the observed likelihood $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ij_i}, j_i)$, where $\boldsymbol{\theta}$ is a short-hand notation for all the model parameters and $f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ij_i}, j_i)$ is the density or probability of the observed outcomes. Note that we omit, for ease of notation, to indicate the dependence of $f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ij_i}, j_i)$ on $\boldsymbol{\theta}$ and on the covariates which are considered as fixed. The joint distribution of the observed outcomes is given by

$$f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ij_i}, j_i) = \sum_{u=1}^{k_1} \lambda_u f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ij_i}, j_i \mid U_i = u), \tag{4}$$

$$\begin{aligned} f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ij_i}, j_i \mid U_i = u) = & \sum_{v_1=1}^{k_2} \cdots \sum_{v_{j_i}=1}^{k_2} \left(\pi_{v_1} \prod_{j=2}^{j_i} \pi_{v_{j-1} v_j} \right) \\ & \times \left\{ \prod_{h=1}^{r} \prod_{j=1}^{j_i} f(y_{hij} \mid u, v_j) \right\} \\ & \times \left[p_{ij_i}(u, v_{j_i}) \prod_{j=1}^{j_i - 1} \{1 - p_{ij}(u, v_j)\} \right], \end{aligned} \tag{5}$$

where $f(y_{hij} \mid u, v)$ refers to the conditional density or probability of Y_{hij} evaluated at y_{hij}, given $U_i = u$, $V_{ij} = v$, and the corresponding covariates, and

$$p_{ij}(u, v) = \frac{\exp(\gamma_u^{(1)} + \gamma_v^{(2)} + \boldsymbol{z}_{ij}' \boldsymbol{\delta})}{1 + \exp(\gamma_u^{(1)} + \gamma_v^{(2)} + \boldsymbol{z}_{ij}' \boldsymbol{\delta})}$$

for $j = 1, \ldots, j_i - 1$ with $p_{ij}(u, v) = 1$ for $j = j_i + 1, \ldots, s_i$. Similarly, we can also express the distribution of the missing outcomes given the observed outcomes when a subject drops out before the end of the study. More precisely, consider the case $j_i < s_i$ and, in order to stress the difference between observed and missing outcomes, let $\boldsymbol{y}_{ij}^o = \boldsymbol{y}_{ij}$ for $j = 1, \ldots, j_i$ (observed outcomes) and $\boldsymbol{y}_{ij}^m = \boldsymbol{y}_{ij}$ for $j = j_i + 1, \ldots, s_i$ (missing outcomes). We have that

$$p(\boldsymbol{y}_{ij}^m \mid \boldsymbol{y}_{i1}^o, \ldots, \boldsymbol{y}_{ij_i}^o, j_i) = \frac{f(\boldsymbol{y}_{i1}^o, \ldots, \boldsymbol{y}_{ij_i}^o, \boldsymbol{y}_{ij}^m, j_i)}{f(\boldsymbol{y}_{i1}^o, \ldots, \boldsymbol{y}_{ij_i}^o, j_i)}, \quad j = j_i + 1, \ldots, s_i,$$

where $f(\boldsymbol{y}_{i1}^o, \ldots, \boldsymbol{y}_{ij_i}^o, j_i)$ at the denominator is defined as in (4), whereas

$$\begin{aligned} & f(\boldsymbol{y}_{i1}^o, \ldots, \boldsymbol{y}_{ij_i}^o, \boldsymbol{y}_{ij}^m, j_i) \\ & \quad = \sum_{u=1}^{k_1} \lambda_u f(\boldsymbol{y}_{i1}^o, \ldots, \boldsymbol{y}_{ij_i}^o, \boldsymbol{y}_{ij}^m, j_i \mid U_i = u), \\ & f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ij_i}, \boldsymbol{y}_{ij}^m, j_i \mid U_i = u) \\ & \quad = \sum_{v_1=1}^{k_2} \cdots \sum_{v_j=1}^{k_2} \left(\pi_{v_1} \prod_{l=2}^{j} \pi_{v_{l-1} v_l} \right) \left\{ \prod_{h=1}^{r} \prod_{l=1}^{j_i} f(y_{hil} \mid u, v_l) \right\} \\ & \qquad \times \left\{ \prod_{h=1}^{r} f(y_{hij} \mid u, v_j) \right\} \left[p_{ij_i}(u, v_{j_i}) \prod_{l=1}^{j_i - 1} \{1 - p_{il}(u, v_l)\} \right]. \end{aligned}$$

This result confirms that the dependence between missing and observed outcomes is due to common latent variables, as already noted in Section 2.

Expression (5) can be efficiently computed by an extension of the *forward recursion* (Baum et al., 1970; Zucchini and MacDonald, 2009; Bartolucci et al., 2013) that we here propose. First of all, ruling out the trivial case in which $s_i = 1$, we consider the following density or probability for $j = 1, \ldots, j_i$ and $u = 1, \ldots, k$:

$$a_{ij}(v \mid u) = \begin{cases} f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ij}, J_i > j, V_{ij} = v \mid U_i = u) \text{ if } j < j_i, \\ f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ij}, J_i = j, V_{ij} = v \mid U_i = u) \text{ if } j = j_i. \end{cases}$$

Then, for $j = 1$ we have that

$$a_{i1}(v \mid u) = \begin{cases} \pi_v \left\{ \prod_{h=1}^{r} f(y_{hi1} \mid u, v) \right\} \{1 - p_{ij}(u, v)\} & \text{if } j_i > 1, \\ \pi_v \left\{ \prod_{h=1}^{r} f(y_{hi1} \mid u, v) \right\} p_{ij}(u, v) & \text{if } j_i = 1, \end{cases}$$

whereas, provided that $j_i > 1$, for $j = 2, \ldots, j_i$ we have

$$a_{ij}(v \mid u) = \begin{cases} \sum_{u=1}^{k} a_{i,j-1}(v \mid u) \pi_{\overline{v}v} \left\{ \prod_{h=1}^{r} f(y_{hij} \mid u, v) \right\} \{1 - p_{ij}(u, v)\} & \text{if } j < j_i, \\ \sum_{u=1}^{k} a_{i,j-1}(v \mid u) \pi_{\overline{v}v} \left\{ \prod_{h=1}^{r} f(y_{hij} \mid u, v) \right\} p_{ij}(u, v) & \text{if } j = j_i, \end{cases}$$

where $v = 1, \ldots, k$. At the end of the recursion ($j = j_i$), we have that

$$f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ij_i}, j_i \mid U_i = u) = \sum_{v=1}^{k} a_{ij_i}(v \mid u).$$

In order to maximize the likelihood $L(\boldsymbol{\theta})$, we use a version of the EM algorithm (Dempster et al., 1977), which is based on the *complete data likelihood*. Let $w_i(u)$ denote an indicator variable equal to 1 if the ith subject is in latent class u ($U_i = u$) and let $z_{ij}(v)$ denote an indicator variable equal to 1 if, at the jth occasion, the same subject is in latent state v ($V_{ij} = v$); also let $z_{ij}(\overline{v}, v) = z_{i,j-1}(\overline{v}) z_{ij}(v)$ be a dummy variable equal to 1 if there is a transition from latent state $\overline{v}$ to latent state v at occasion j. The logarithm of the complete likelihood has the following expression:

$$\begin{aligned} \ell_c(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \sum_{u=1}^{k_1} w_i(u) \{\log(\lambda_u) + m_i(\boldsymbol{\theta} \mid u)\}, \\ m_i(\boldsymbol{\theta} \mid u) &= \sum_{v=1}^{k} z_{i1}(v) \log \pi_v + \sum_{j=2}^{j_i} \sum_{\overline{v}=1}^{k_2} \sum_{v=1}^{k_2} z_{ij}(\overline{v}, v) \log \pi_{\overline{v}v} \\ &\quad + \sum_{h=1}^{r} \sum_{j=1}^{j_i} \sum_{v=1}^{k_2} z_{ij}(v) \log f(y_{hij} \mid u, v) \\ &\quad + \sum_{v=1}^{k_2} \left[\sum_{j=1}^{j_i - 1} z_{ij}(v) \log\{1 - p_{ij}(u, v)\} \right. \\ &\quad \left. + z_{ij_i}(v) \log p_{ij_i}(u, v) \right], \end{aligned}$$

where the second sum in $m_i(\boldsymbol{\theta} \mid u)$, which involves $z_{ij}(\overline{v}, v)$, disappears if $j_i = 1$.

The EM algorithm alternates two steps until convergence: first, the conditional expected value of the complete data log-likelihood is obtained (E-step). The resulting expression is then maximized with respect to $\boldsymbol{\theta}$ (M-step). These steps are illustrated in detail in the Supplementary Material to the present article. The EM algorithm is guaranteed to converge to a local optimum of the observed likelihood. In order to increase the chances of reaching the global maximum, we use a multistart strategy.

4. Application to Primary Biliary Cirrhosis Data

We illustrate the proposed approach through an application to a randomized study for the treatment of primary biliary cirrhosis. These data regard $n = 312$ patients and were previously analyzed by Rizopoulus, Verbeke, and Molenberghs (2010). These patients were randomized to a placebo or a treatment based on D-penicillamine. We are interested in evaluating the effect of treatment on a continuous outcome (*logarithm of serum Bilirubin in mg/dl*, Y_1) and a binary outcome (*presence of edema*, Y_2), after adjusting for certain baseline covariates (*drug*, *age*, *gender*, *albumin in gm/dl*, *logarithm of alkaline phosphatase in U/L*, and *logarithm of transaminase at first visit in U/ml*, the last one is abbreviated as SGOT) and drop-out. Among the covariates, we also include the *time* in terms of number of years between enrollment and the visit, together with its interaction with the treatment. The two outcomes summarize two aspects of the disease. Serum Bilirubin is linked to liver and spleen functionality, and it is likely to be connected with worsening conditions, hence making drop-out due to death more likely (this enforces the idea that drop-out must be treated as informative). Edema is a consequence of accumulation of toxic compounds and fluids. D-penicillamine, which is a chelant of copper, is now routinely used for the treatment of primary biliary cirrhosis.

The maximum number of follow-up time occasions is 16 for these data and only 1% of patients have a complete record; the median time to drop-out is 5. Table 1 reports the proportion of subjects having a certain number of observations, that is, $\sum_{i=1}^{n} I(j_i = j)/n$, $j = 1, \ldots, 16$, where $I(\cdot)$ is the indicator function. The table also shows the corresponding Kaplan–Meier estimates, and the mean of the two outcomes based on the number of survivors.

We begin the analysis of these data by trying different configurations in terms of number of support points of the time-constant latent variable (k_1) and number of states of the hidden Markov chain (k_2). For the reasons of model interpretability clarified in Section 2, we assume that the initial distribution corresponds to the stationary distribution of the transition matrix. Note that, though measurement times are not exactly regularly spaced, the variability of the time intervals between consecutive visits on the patient is rather small; hence we decided to use a homogeneous latent chain. In Table 2 we report the BIC for increasing values of k_1 and k_2 from 1 to 3.

Table 1

Observed mean of serum Bilirubin and proportion of subjects with edema by time for the primary biliary cirrhosis data, together with the proportion of survivors ($\sum_{i=1}^{n} I(j_j = j)/n$) and corresponding Kaplan–Meier estimates (KM_j); n_j is equal to the total number of subjects observed at time point j, for $j = 1, \ldots, 16$.

Time	$\sum_{i=1}^{n} I(j_i = j)/n$	KM_j	Mean (Y_1)	Mean (Y_2)	n_j
1	0.09	0.09	3.22	0.21	312
2	0.08	0.09	3.07	0.21	285
3	0.10	0.10	3.45	0.24	259
4	0.14	0.14	4.26	0.27	227
5	0.10	0.10	3.62	0.28	183
6	0.07	0.07	3.91	0.30	153
7	0.06	0.06	3.77	0.37	130
8	0.07	0.07	3.93	0.36	111
9	0.05	0.05	4.01	0.37	88
10	0.07	0.07	3.49	0.41	71
11	0.05	0.05	5.14	0.40	48
12	0.04	0.04	4.23	0.34	32
13	0.02	0.02	5.06	0.40	20
14	0.02	0.02	4.32	0.43	14
15	0.02	0.02	6.28	0.33	9
16	0.01	0.02	5.17	0.67	3

On the basis of the results in Table 2, we select $k_1 = 3$ and $k_2 = 3$, as the corresponding model has the lowest BIC of 5451.8; the corresponding log-likelihood value is -2588.1 with 48 free parameters. We also consider the model with $k_2 = 1$ and $k_1 = 10$ chosen according to BIC. For this model, which only includes time-fixed latent variables, we have a log-likelihood value of -2944.3 with 64 parameters, so that $BIC = 6256.2$, a much higher value than that of the proposed model, confirming that the inclusion of time-varying latent variables may considerably improve the fit. On the other hand, for the model constrained with constant elements in each row of the transition matrix, with $k_1 = 3$ and $k_2 = 3$, we have maximum log-likelihood equal to -2635.0 with 45 parameters and then a higher BIC value, equal to 5528.4. Consequently, we retain the model with unconstrained transition matrix and initial distribution equal to the stationary distribution.

It is important to note that in selecting the model, we limit k_1 and k_2 to 3 to avoid the uncertainty due to the multimodality of the model likelihood, despite a model with more classes would have probably achieved a smaller BIC. Note that models with larger values of k_1 and k_2 also result in groups of very close estimates for $\gamma_u^{(1)}$ and $\gamma_v^{(2)}$, indicating that the MNAR structure has been taken into account within the model class with the chosen values of k_1 and k_2.

We recall that the EM algorithm is initialized by a combination of a deterministic rule and a stochastic rule. Overall, in estimating a model with certain values of k_1 and k_2, we tried a number of random initializations equal to ten times the difference between the number of free parameters of this model and the number of parameters of the initial model without latent variables ($k_1 = k_2 = 1$). In Table 2 we also report the overall number of initializations of the EM algorithm and the number of times that this algorithm reaches the maximum log-likelihood level among these initializations. On the basis of these results, we consider the obtained solutions as reliable because, for each fitted model, the best point at convergence is obtained from a reasonable number of different random initializations. On the other hand, we noted that for larger values of k_1 and k_2 it is rather rare to obtain a repetition of the best solution, confirming the appropriateness of limiting these quantities to 3. Trying higher values of k_1 and k_2 also leads to a higher computing time and may lead to unstable support point estimates for the binary outcome.

Under the selected model with $k_1 = k_2 = 3$, we obtain the parameter estimates reported in Table 3. In order to perform a sensitivity analysis on k_1 and k_2, we also report the estimates obtained with $k_1 = k_2 = 1$ and $k_1 = k_2 = 2$. For comparison, we report in Table 4 parameter estimates obtained with a Missing at Random (MAR) version of our model in which the drop-out is not explicitly modeled for $k_1 = k_2 = 2$ and $k_1 = k_2 = 3$ and with the MNAR model with $k_1 = 10$ and $k_2 = 1$. The last is important as a comparison for absence of time-varying unobserved heterogeneity.

Table 2

BIC for different values of k_1 and k_2 for the primary biliary cirrhosis data, together with log-likelihood and number of parameters. The table also shows the number of initializations of the EM algorithm for each combination of k_1 and k_2, together with the number of times the log-likelihood at convergence is equal to the best solution.

	Log-lik.			# Param.			BIC		
k_1	$k_2 = 1$	$k_2 = 2$	$k_2 = 3$	$k_2 = 1$	$k_2 = 2$	$k_2 = 3$	$k_2 = 1$	$k_2 = 2$	$k_2 = 3$
1	−4176.1	−3356.7	−2990.0	28	33	40	8512.9	6903.0	6209.7
2	−3523.6	−3008.0	−2716.2	32	37	44	7230.9	6228.4	5685.1
3	−3323.0	−2855.7	−2588.1	36	41	48	6852.8	5946.9	5451.8
	# Initializations			# Best solutions					
k_1	$k_2 = 1$	$k_2 = 2$	$k_2 = 3$	$k_2 = 1$	$k_2 = 2$	$k_2 = 3$			
1	1	51	121	1	48	99			
2	41	91	161	40	7	5			
3	81	131	201	19	15	11			

Table 3

Parameter estimates for the latent Markov model with informative drop-out for the primary biliary cirrhosis data with $k_1 = k_2$ from 1 to 3

Outcome: *log(Serum Bilirubin)*									
	$k_1 = k_2 = 1$ (MNAR)			$k_1 = k_2 = 2$ (MNAR)			$k_1 = k_2 = 3$ (MNAR)		
Parameter	Est.	S.E.	p-value	est.	S.E.	p-value	Est.	S.E.	p-value
Intercept	−3.588	—	—	−1.562	—	—	−0.591	—	—
Treatment	−0.032	0.061	0.594	−0.112	0.064	0.079	−0.169	0.043	0.000
Age/10	−0.035	0.022	0.116	−0.006	0.023	0.801	0.013	0.013	0.311
Gender (F)	−0.431	0.066	0.000	−0.324	0.047	0.000	−0.485	0.031	0.000
Albumin	−0.477	0.058	0.000	−0.193	0.040	0.000	−0.143	0.031	0.000
Log-alkaline ph.	0.225	0.029	0.000	0.135	0.019	0.000	0.095	0.014	0.000
Log-SGOT	0.974	0.051	0.000	0.482	0.043	0.000	0.337	0.030	0.000
Time	0.033	0.010	0.001	0.051	0.006	0.000	0.033	0.005	0.000
Treatment.time	0.005	0.014	0.724	−0.001	0.009	0.902	0.010	0.007	0.173
Outcome: *Edema*									
	$k_1 = k_2 = 1$ (MNAR)			$k_1 = k_2 = 2$ (MNAR)			$k_1 = k_2 = 3$ (MNAR)		
Parameter	Est.	S.E.	p-value	Est.	S.E.	p-value	Est.	S.E.	p-value
Intercept	−1.043	—	—	−0.039	—	—	−1.327	—	—
Treatment	−0.165	0.161	0.305	−0.282	0.185	0.127	−0.062	0.260	0.812
Age/10	0.549	0.060	0.000	0.683	0.070	0.000	0.888	0.108	0.000
Gender (F)	0.533	0.176	0.003	0.737	0.195	0.000	0.269	0.299	0.368
Albumin	−1.107	0.149	0.000	−1.037	0.168	0.000	−0.615	0.227	0.007
Log-alkaline ph.	0.047	0.072	0.514	0.027	0.085	0.747	−0.006	0.120	0.962
Log-SGOT	0.591	0.131	0.000	0.291	0.157	0.063	0.176	0.240	0.464
Time	0.167	0.025	0.000	0.246	0.028	0.000	0.394	0.042	0.000
Treatment.time	−0.015	0.033	0.647	−0.044	0.037	0.237	−0.093	0.051	0.071
Drop-out									
	$k_1 = k_2 = 1$ (MNAR)			$k_1 = k_2 = 2$ (MNAR)			$k_1 = k_2 = 3$ (MNAR)		
Parameter	Est.	S.E.	p-value	Est.	S.E.	p-value	Est.	S.E.	p-value
Intercept	−1.815	—	—	1.713	—	—	2.916	—	—
Treatment	−0.132	0.251	0.600	−0.272	0.283	0.337	−0.257	0.282	0.362
Age/10	0.250	0.092	0.006	0.290	0.102	0.004	0.314	0.103	0.002
Gender (F)	−0.290	0.232	0.211	−0.362	0.263	0.170	−0.742	0.269	0.006
Albumin	−1.319	0.216	0.000	−1.197	0.247	0.000	−1.070	0.242	0.000
Log-alkaline ph.	−0.057	0.114	0.616	−0.199	0.132	0.130	−0.296	0.137	0.031
Log-SGOT	0.936	0.198	0.000	0.245	0.245	0.317	0.096	0.251	0.700
Time	0.150	0.038	0.000	0.197	0.043	0.000	0.175	0.045	0.000
Treatment.time	−0.001	0.052	0.992	−0.010	0.059	0.860	−0.014	0.061	0.816

From the results obtained under the chosen model, we conclude that the treatment is beneficial on serum Bilirubin, but it does not seem to be on the time to drop-out or on edema. Note that Rizopoulus et al. (2010) obtains similar results for the treatment effect on serum Bilirubin (as the MAR model does, see below). The treatment is a chelant, which binds copper and helps therefore the body to get rid of potentially toxic accumulation of this metal. An effect on edema could therefore be expected and the lack of significance suggests that there may be other biochemical or physiological pathways involved in the illness other than copper accumulation in tissues. Concerning the other predictors, it can be observed that, with respect to age, the probability of edema increases, that females have lower levels of Bilirubin, and that albumin is protective. The sensitivity analysis testifies that for these data there is a large sensitivity to drop-out, as treatment effect estimates are sensibly different for different values of k_1 and k_2.

We finally consider the estimated latent distribution parameters. In Table 5 we report the differences between the latent intercepts associated with the time-fixed component and their averages, and the probability mass estimates. In Table 6 we report the differences between the latent intercepts associated with the time-varying component and their averages, the initial parameter vector, and the transition matrix.

Table 4
Parameter estimates for the latent Markov model with informative drop-out for the primary biliary cirrhosis data with $k_1 = 10$, $k_2 = 1$ and with ignorable dropout with $k_1 = k_2 = 2$ and $k_1 = k_2 = 3$

Outcome: *log(Serum Bilirubin)*									
	$k_1 = 10, k_2 = 1$ (MNAR)			$k_1 = k_2 = 2$ (MAR)			$k_1 = k_2 = 3$ (MAR)		
Parameter	Est.	S.E.	p-value	Est.	S.E.	p-value	Est.	S.E.	p-value
Intercept	−1.747	—	—	−1.785	—	—	−0.605	—	—
Treatment	0.082	0.047	0.084	−0.076	0.054	0.163	−0.184	0.041	0.000
Age/10	0.029	0.019	0.123	−0.017	0.019	0.353	0.013	0.013	0.312
Gender (F)	−0.173	0.053	0.001	−0.349	0.049	0.000	−0.501	0.031	0.000
Albumin	−0.620	0.058	0.000	−0.152	0.039	0.000	−0.131	0.031	0.000
Log-alkaline ph.	0.171	0.024	0.000	0.119	0.025	0.000	0.095	0.014	0.000
Log-SGOT	0.717	0.041	0.000	0.520	0.044	0.000	0.336	0.028	0.000
Time	0.092	0.006	0.000	0.052	0.006	0.000	0.034	0.005	0.000
Treatment.time	0.011	0.008	0.206	−0.003	0.008	0.758	0.011	0.007	0.107
Outcome: *Edema*									
	$k_1 = 10, k_2 = 1$ (MNAR)			$k_1 = k_2 = 2$ (MAR)			$k_1 = k_2 = 3$ (MAR)		
Parameter	Est.	S.E.	p-value	Est.	S.E.	p-value	Est.	S.E.	p-value
Intercept	−4.846			0.242			−1.223		
Treatment	0.641	0.273	0.019	−0.242	0.185	0.190	−0.022	0.254	0.931
Age/10	1.281	0.118	0.000	0.667	0.070	0.000	0.840	0.104	0.000
Gender (F)	0.310	0.262	0.238	0.694	0.199	0.000	0.223	0.286	0.434
Albumin	−1.741	0.306	0.000	−1.031	0.178	0.000	−0.611	0.232	0.008
Log-alkaline ph.	0.051	0.138	0.712	−0.002	0.097	0.982	−0.024	0.119	0.842
log-SGOT	1.449	0.239	0.000	0.282	0.156	0.070	0.190	0.236	0.420
Time	0.611	0.051	0.000	0.237	0.031	0.000	0.390	0.042	0.000
Treatment.time	−0.094	0.056	0.091	−0.040	0.037	0.275	−0.099	0.051	0.052
	Drop − out								
	$k_1 = 10, k_2 = 1$ (MNAR)								
Parameter	Est.	S.E.	p-value						
Intercept	−0.824	—	—						
Treatment	0.048	0.282	0.864						
Age/10	0.423	0.104	0.000						
Gender (F)	−0.377	0.264	0.153						
Albumin	−1.647	0.266	0.000						
Log-alkaline ph.	−0.072	0.136	0.597						
Log-SGOT	0.822	0.243	0.001						
Time	0.302	0.046	0.000						
Treatment.time	0.009	0.058	0.873						

We observe that the groups are rather well separated both with respect to the time-constant and the time-varying latent variables, and that $\mathbf{\Pi}$ is far from a diagonal matrix, which corresponds to a situation of perfect persistence. We conclude that there are distinct subgroups of patients with different prognostic expectations (time-fixed unobserved heterogeneity), and that patients may move from one subgroup to another during the follow-up for unforeseen reasons (time-varying unobserved heterogeneity).

5. Discussion

We propose an event-history approach to Mixed LM models with informative drop-out. The model can be used in discrete time, that is, when it is only known that drop-out has occurred between two measurement occasions. It can also be applied in continuous time, but at the price of a loss of efficiency. The basic assumption is similar to that of classical shared-parameter models, that is, random effects capture all missing information due to drop-out and hence missing indicators and outcomes are conditionally independent given these random effects. We also relax the usual assumption that random effects are time constant by adding a second term distributed according to a hidden Markov chain. In order to derive inference we extend the usual forward-backward recursions to informative drop-out and consequently propose an extended version of the EM algorithm for maximum likelihood

Table 5

Time-fixed latent distribution parameter estimates for the latent Markov model with informative drop-out, with $k_1 = k_2 = 3$, for the primary biliary cirrhosis data (support points ordered according to the first dimension; standard errors in brackets)

u	$\widehat{\alpha}^{(1)}_{1u}$	$\widehat{\alpha}^{(2)}_{2u}$	$\widehat{\gamma}^{(1)}_{u}$	$\widehat{\lambda}_{u}$
1	−0.422	3.335	0.192	0.139
	(0.038)	(0.266)	(0.276)	(0.023)
2	−0.292	−2.249	−0.600	0.356
	(0.027)	(0.227)	(0.171)	(0.034)
3	0.322	0.664	0.369	0.505
	(0.027)	(0.163)	(0.117)	(0.036)

estimation in MLM models. The advantages of the proposed type of modeling are shown by an application based on a study about primary biliary cirrhosis in which there are two outcomes of interest, the first is continuous and the second is binary.

The high flexibility of the proposed model comes at the price of certain estimation complexity, which is mainly related to the multimodality of the likelihood function. Taking this into account, it is important to use a limited number of support points for the time-constant and for the time-varying latent variables, especially if the sample size is limited. Moreover, it is important to check if the best solution obtained from the EM estimation algorithm is also obtained starting from different points of the parameter space. In the application, we show how this task may be simply performed by using a random starting initialization rule for the EM algorithm.

6. Supplementary Materials

Web Appendix, Tables and code referenced in Sections 1 and 3 are available with this paper at the *Biometrics* website on Wiley Online library.

Acknowledgements

The authors are grateful to the Editor, two Referees, and an AE for kind comments. F. Bartolucci acknowledges the financial support from the grant RBFR12SHVV of the Italian Government (FIRB project "Mixture and latent variable models for causal inference and analysis of socio-economic data").

Table 6

Time-varying latent distribution parameter estimates for the latent Markov model with informative drop-out, with $k_1 = 3$ and $k_2 = 3$, for the primary biliary cirrhosis data (support points ordered according to the first dimension; standard errors in brackets)

v	$\widehat{\alpha}^{(2)}_{1v}$	$\widehat{\alpha}^{(2)}_{2v}$	$\widehat{\gamma}^{(2)}_{v}$	$\widehat{\pi}_{v}$	$\widehat{\pi}_{v1}$	$\widehat{\pi}_{v2}$	$\widehat{\pi}_{v3}$
1	−0.824	−1.353	−1.103	0.475	0.954	0.046	0.000
	(0.057)	(0.171)	(0.198)	(0.031)	(0.007)	(0.007)	(0.001)
2	0.121	0.622	0.011	0.269	0.018	0.851	0.130
	(0.054)	(0.185)	(0.239)	(0.025)	(0.011)	(0.017)	(0.016)
3	1.396	1.850	2.028	0.257	0.067	0.071	0.863
	(0.051)	(0.151)	(0.195)	(0.022)	(0.018)	(0.022)	(0.018)

References

Albert, P. S. (2000). A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics* **56**, 602–608.

Altman, R. M. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association* **102**, 201–210.

Bartolucci, F. and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association* **104**, 816–831.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: Chapman & Hall/CRC Press.

Bartolucci, F., Lupparelli, M., and Montanari, G. E. (2009). Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *Annals of Applied Statistics* **3**, 611–636.

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.

Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., and Kenward, M. G. (2010). A sensitivity analysis for shared-parameter models for incomplete longitudinal outcomes. *Biometrical Journal* **52**, 111–125.

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, Series B* **39**, 1–38.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics* **43**, 49–93.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley-Interscience.

Follmann, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151–168.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

Ip, E., Zhang, Q., Rejeski, J., Harris, T., and Kritchevsky, S. (2013). Partially ordered mixed hidden Markov model for the disablement process of older adults. *Journal of the American Statistical Association* **108**, 370–384.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Hoboken, NJ: Wiley.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90**, 1112–1121.

Little, R. J. A. and Wang, Y. (1996). Pattern-mixing models for multivariate incomplete data with covariates. *Biometrics* **52**, 98–111.

Maruotti, A. (2011). Mixed hidden Markov models for longitudinal data: An overview. *International Statistical Review* **79**, 427–454.

Molenberghs, G., Kenward, M. G., and Lesaffre, E. (1997). The analysis of longitudinal data with non-random dropout. *Biometrika* **84**, 33–44.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* **35**.

Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine* **30**, 1366–1380.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika* **95**, 63–74.

Rizopoulus, D., Verbeke, G., and Molenberghs, G. (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics* **66**, 20–29.

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics* **59**, 829–836.

Steele, F. (2011). Multilevel discrete-time event history models with applications to the analysis of recurrent employment transitions. *Australian & New Zealand Journal of Statistics* **53**, 1–20.

Ten Have, T. R., Kunselman, A. R., Pulkstenis, E. P., and Landis, J. R. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics* **54**, 367–383.

Viviani, S., Alfó, M., and Rizopoulos, D. (2014). Generalized linear mixed joint model for longitudinal and survival outcomes. *Statistics and Computing* **24**, 417–427.

Viviani, S., Rizopoulos, D., and Alfó, M. (2014). Local sensitivity to non-ignorability in joint models. *Statistical Modelling* **14**, 205–228.

Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter* **53**, 1–13.

Wu, M. and Carrol, R. (1988). Estimation and comparison of changes in presence of informative right censoring by modelling the censoring process. *Biometrics* **44**, 175–188.

Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics* **45**, 939–955.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. New York: Springer-Verlag.

Received November 2013. *Revised July* 2014. *Accepted July* 2014.