Received: 1 February 2018 | Revised: 24 August 2018 | Accepted: 13 September 2018

DOI: 10.1002/sim.7994

RESEARCH ARTICLE

WILEY Statistics in Medicine

A shared-parameter continuous-time hidden Markov and survival model for longitudinal data with informative dropout

Francesco Bartolucci[1] | Alessio Farcomeni[2]

[1]Faculty of Economics, University of Perugia, Perugia, Italy

[2]Department of Public Health and Infectious Diseases, Sapienza University of Rome, Rome, Italy

Correspondence
Alessio Farcomeni, Department of Public Health and Infectious Diseases, Sapienza University of Rome, Piazzale Aldo Moro 5, 00186 Roma, Italy.
Email: alessio.farcomeni@uniroma1.it

A shared-parameter approach for jointly modeling longitudinal and survival data is proposed. With respect to available approaches, it allows for time-varying random effects that affect both the longitudinal and the survival processes. The distribution of these random effects is modeled according to a continuous-time hidden Markov chain so that transitions may occur at any time point. For maximum likelihood estimation, we propose an algorithm based on a discretization of time until censoring in an arbitrary number of time windows. The observed information matrix is used to obtain standard errors. We illustrate the approach by simulation, even with respect to the effect of the number of time windows on the precision of the estimates, and by an application to data about patients suffering from mildly dilated cardiomyopathy.

KEYWORDS

Baum-Welch recursions, expectation-maximization algorithm, latent class model, mildly dilated cardiomyopathy

1 | INTRODUCTION

Informative dropout in longitudinal studies is often modeled by linking a model for time to drop out and one for the longitudinal outcome. In shared parameter models,[1-4] this link is provided by the effect of a (possibly scaled) shared latent variable. This is often Gaussian, but more flexible discrete random-effects distributions are also possible,[5,6] together with semiparametric approaches.[7]

Many models devised for informative dropout assume that subject-specific parameters are time constant. This is a limitation as unobserved factors affecting the outcomes and the relationship between longitudinal and survival outcomes might evolve over time in an unpredictable way, especially when the follow-up is relatively long. We focus here on exceptions based on latent Markov processes.[6,8-11] While these approaches rely on different strategies to take into account informative dropout with time-varying subject-specific parameters, they have two common limitations: (*i*) dropout is summarized as occurring within a time interval, therefore ignoring precise follow-up time information (alternatively it is simply used as a covariate for modeling the latent process[10]); (*ii*) latent transitions are based on a discrete-time stochastic process and hence may only occur at visit times. In terms of interpretation, assuming that transitions may occur only at certain time occasions is rather unrealistic and the explicit use of a hazard function is preferable to that of a conditional logit model.

In order to overcome the above limitations, we propose a shared-parameter model characterized by the following features. First of all, the time-varying unobserved heterogeneity is accounted for by a continuous-time discrete-state hidden Markov (HM) model[12] parameterized by an initial probability vector and an infinitesimal transition matrix. In this

respect, our approach can be seen as a complete generalization of a previous work,[13] which is limited to $k = 2$ and to missing-at-random data. Second, for the survival time, we assume a Weibull model with hazard function depending on the (entire) trajectory of the continuous-time latent variable. A latent class model (with time-constant subject-specific parameters) is obtained whenever the infinitesimal transition matrix is constrained to be the zero matrix.

Latent continuous-time Markov chains have already been considered,[14,15] for example, for modeling multistate processes with random informative observation times[16] (but without the survival component of the model), and inference is similar to the case of discrete-time chains with inhomogeneous transition matrices.[17-20] In our case, the likelihood depends on the entire trajectory of the latent process and not only on its state at certain time points, making inference much more challenging. In this regard, the proposed approach is related to the mixed functional approach[21,22] that has a high degree of flexibility for longitudinal continuous responses, but it does not jointly consider longitudinal and survival processes. Additional related approaches[23,24] may be used with functional covariates.

For model fitting, we introduce a novel method based on a discretization of the time scale in a certain number of windows of arbitrary length and on an extension of the Baum-Welch recursions.[25,26] Our algorithm converges in an accurate and stable way and represents an advance also within the literature about estimation of continuous HM models, in general, in terms of computational demand, ease of implementation, and stability.[27,28]

The proposed approach is motivated by an original application about monitoring quality of life (QoL) of patients with mildly dilated cardiomyopathy (MDCM). Our data regard $n = 642$ (anonymized) patients who were followed for up to 25 years. At regular visits during the first 10 years of follow-up, the New York Hearth Association (NYHA) class, indicating occurrence of heart insufficiency signs limiting daily physical activity, was assessed. We use NYHA class as a proxy for QoL.[29,30] Obviously, we must take into account survival since patients with worse QoL are also expected to have a worse prognosis. Given the long follow-up, it is unlikely that unmeasured factors have remained constant over time. Furthermore, disease progression (ie, transition among latent states) could obviously occur at any time and not only at visit times, which often are several months apart.

The remainder of this paper is organized as follows: in the next section, we introduce our novel shared-parameter longitudinal and survival model, where latent variables follow a continuous-time HM process. In Section 3, we outline the novel inferential procedure. In Section 4, we describe a brief simulation study and, in Section 5, we apply our approach to the MDCM data set. Concluding remarks are given in Section 6. We implemented the proposed estimation method in `R`, with recursions and likelihood computation embedded within Fortran code. Our code is available at `https://github.com/afarcome/lmjm`.

2 | SHARED-PARAMETER CONTINUOUS-TIME LATENT MARKOV AND SURVIVAL MODELS

Consider a sample of n individuals and for individual i, with $i = 1, \dots, n$, let $T_i = \min(T_i^*, C_i)$ be the survival time taken as the minimum between the true event time T_i^* and the censoring time C_i. Furthermore, let Δ_i be the corresponding event indicator defined as $\Delta_i = I(T_i^* \leq C_i)$, where $I(\cdot)$ is the indicator function equal to 1 if its argument is true and to 0 otherwise. The outcome $Y_i(t)$, which has a natural exponential family distribution, is repeatedly observed at arbitrary time points $t_{ij}, j = 1, \dots, j_i$, where j_i is the number of observations and we also let $Y_{ij} = Y_i(t_{ij})$. We assume that the longitudinal process is associated with T_i^*, namely, with the true event time, but, as customary in survival analysis, is independent of the censoring time C_i. In general, realizations of random variables are denoted by small letters, so that, for instance, t_i is the observed value of T_i and δ_i is the observed value of Δ_i.

We denote by $\boldsymbol{w}_i$ a row vector of (time fixed) baseline covariates to be used in modeling the survival process. For the longitudinal process, we denote by $\boldsymbol{x}_i(t)$ a vector of predictors at time t and we also let $\boldsymbol{x}_{ij} = \boldsymbol{x}_i(t_{ij}), j = 1, \dots, j_i$. This is related to time-varying frailty models.[31]

The assumptions of the proposed model are illustrated in the following. Then, in order to clarify how the model assumptions affect possible latent trajectories and outcome distributions, we show how it is possible to jointly sample from the longitudinal and the survival process. This sampling scheme will be also used in the simulation study described in Section 4.

2.1 | Model assumptions

The model is based on two equations. Specifically, the model for the longitudinal outcomes is formulated along the usual lines as for mixed-effects models,[32] and the model for the time-to-event outcome is based on a subject-specific hazard

function as in Cox-type models.[33,34] More formally, we assume that

$$\begin{cases} g(\mu_{ij}) = \alpha_i(t_{ij}) + \boldsymbol{x}'_{ij}\boldsymbol{\beta}, \quad j = 1, \ldots, j_i, \\ h_i\left(t_i^*\right) = h_0\left(t_i^*\right) \exp\left[\alpha_i\left(t_i^*\right)\phi + \boldsymbol{w}'_i\boldsymbol{\psi}\right], \end{cases} \tag{1}$$

where $g(\cdot)$ is a link function[35] applied to the conditional expectation of Y_{ij} denoted by μ_{ij} and $h(\cdot)$ is the hazard function, with $h_0(\cdot)$ being a baseline hazard. In this paper, we will assume a Weibull parametric form for $h_0(\cdot)$, that is, $h_0(t) = \nu t^{\nu-1}$, resulting in an accelerated failure time model (AFT) for the survival part. Also, other parametric choices, or even a nonparametric specification, are possible. For the longitudinal part, nonparametric specifications could be accommodated through the rank transformation.[36-39] We assume that $\alpha_i(t)$ follows a time-continuous Markov process, whereas $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ are fixed parameter vectors for the covariates and ϕ is a parameter for the effect of the latent process on the survival process. Note that several generalizations, including the case of more than one parameter being time-dependent according to the latent process, are straightforward.

Regarding the distribution of Y_{ij}, our model has the same degree of flexibility as generalized linear models. We consider, in particular, two cases: the normal distribution for continuous response variables and the Bernoulli distribution for binary variables. In the first case, we assume that the conditional distribution of Y_{ij} is $\mathrm{N}(\mu_{ij}, \sigma^2)$, where the variance parameter σ^2 must be estimated together with the other parameters, and $g(\mu_{ij}) = \mu_{ij}$. In the binary case, we assume that Y_{ij} has conditional $\mathrm{Bern}(\mu_{ij})$ distribution and $g(\mu_{ij}) = \log[\mu_{ij}/(1-\mu_{ij})]$.

The second equation in (1) implies that the survival function for the time to event is

$$S_i\left(t_i^*\right) = \exp\left[-H_i\left(t_i^*\right)\right],$$

where

$$H_i\left(t_i^*\right) = \int_0^{t_i^*} h_0(t) \exp\left[\alpha_i(t)\phi + \boldsymbol{w}'_i\boldsymbol{\psi}\right] dt$$

is the cumulative hazard function. Moreover, we consider the density function $f(t_i) = h(t_i)^{\delta_i} S(t_i)$. All quantities above are *conditional* on the random effects. It is worth stressing that the hazard function depends on the *entire* trajectory of the random effect $\alpha_i(t)$ and not only on the $\alpha_i(t_{ij})$.

Unlike usual formulations, random intercepts are assumed to be time varying. This greatly enhances model flexibility. In particular, as already mentioned, we assume that the random effects follow a continuous-time (discrete-state) Markov chain,[12] with state-space $\{\xi_1, \ldots, \xi_k\}$ having k elements collected in the column vector $\boldsymbol{\xi}$. We assume that the transition function of the latent chain satisfies the Chapman-Kolmogorov equations and specify its $\boldsymbol{Q}$-matrix based on positive off-diagonal elements q_{uv} for $u, v = 1, \ldots, k$, with $v \neq u$. By definition, the diagonal elements are given by $-q_u$, with

$$q_u = \sum_{\substack{v=1 \\ v \neq u}}^{k} q_{uv}, \quad u = 1, \ldots, k.$$

For instance, one could fix $k = 3$,

$$\boldsymbol{Q} = \begin{bmatrix} -1 & 0.5 & 0.5 \\ 0.5 & -1 & 0.5 \\ 0.5 & 0.5 & -1 \end{bmatrix}, \tag{2}$$

and assuming that the initial state is the third, obtain a trajectory as in the uppermost panel of Figure 1.

The transition probabilities from time t to time $t + s$ are collected in the $k \times k$ matrix

$$\boldsymbol{\Pi} = e^{s\boldsymbol{Q}}$$

based on the *matrix exponential* operator, that is,

$$e^{s\boldsymbol{Q}} = \sum_{j=0}^{\infty} \frac{s^j \boldsymbol{Q}^j}{j!}.$$

This formula can be tackled numerically in most cases. Note that irregularly spaced time occasions are directly accommodated, also in the presence of noninformative dropout, simply by restricting to the first equation. We also define the jump matrix $\boldsymbol{R}$ as a matrix with off-diagonal elements $r_{uv} = q_{uv}/q_u$, and collect initial probabilities π_u in the column vector $\boldsymbol{\pi}$.

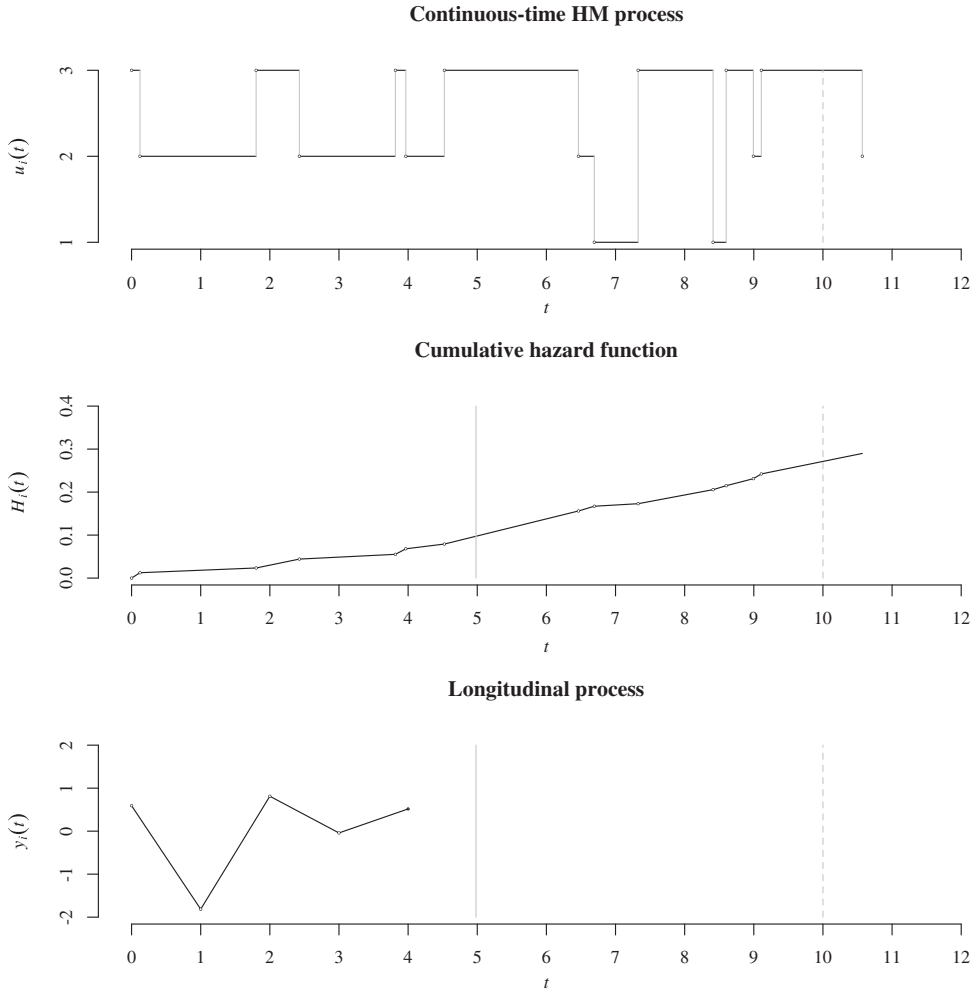

FIGURE 1 Illustration of the simulation process: continuous-time hidden Markov (HM) chain (indicized by latent states 1, 2, 3), cumulative hazard function, and longitudinal process; the simulated survival time is $t_i = 4.98$

For instance, with $\boldsymbol{Q}$ as in (2), $r_{uv} = 0.5/1 = 0.5$ for all $u \neq v$, that is, once there is a jump from state u, the probability is equally likely to move to any of the other states.

The latent process captures the time-varying unobserved heterogeneity linking the longitudinal and survival outcomes. The shared-parameter formulation is in the spirit of copula models.[40] In an equivalent formulation, the above model may be expressed as

$$\begin{cases} g(\mu_{ij}) = \xi_{u_{ij}} + \boldsymbol{x}_{ij}'\boldsymbol{\beta}, \quad j = 1, \ldots, j_i, \\ h_i\left(t_i^*\right) = h_0\left(t_i^*\right) \exp\left[\xi_{u_i(t_i^*)}\phi + \boldsymbol{w}_i'\boldsymbol{\psi}\right], \end{cases} \tag{3}$$

where $u_{ij} = u_i(t_{ij})$, with $U_i(t)$ following a continuous-time Markov chain with k states and the same initial probability vector $\boldsymbol{\pi}$ and infinitesimal transition matrix $\boldsymbol{Q}$. We stress that the possible values of $U_i(t)$ are $1, \ldots, k$ and indicate the latent state of individual i at time t. We also recall that, according to this process, the sojourn time in each state u has an exponential distribution with parameter q_u, denoted as $\text{Exp}(q_u)$, whereas the probability of moving at the end of the sojourn time to state v is equal to the suitable element of the jump matrix $\boldsymbol{R}$.

2.2 | Simulation from the model

For each individual i, we suppose that the censoring time c_i is known. Then, we sequentially generate: (i) the continuous Markov chain $u_i(t)$; (ii) the survival time t_i; and (iii) the longitudinal outcomes y_{ij}.

Regarding the generation of the Markov chain, we first draw the state at time $\widetilde{t}_{i1} = 0$, $\widetilde{u}_{i1} = u_i(0)$, from the initial distribution with mass probabilities $\pi_1, \ldots, \pi_k$. Then, the j-th jump is at time $\widetilde{t}_{i,j+1} = \widetilde{t}_{ij} + \widetilde{\tau}_{ij}$, where $\widetilde{\tau}_{ij}$ is sampled from

$\text{Exp}(q_{\widetilde{u}_{ij}})$ and the new state $\widetilde{u}_{i,j+1} = u_i(\widetilde{t}_{i,j+1})$ is randomly selected on the basis of the probabilities in the $\widetilde{u}_{ij}$-th row of $\boldsymbol{R}$. This process is iterated starting from $j = 1$ and the first time $\widetilde{t}_{i,j+1} > c_i$ the process is stopped considering $\widetilde{j}_i = j$ jumps.

In order to generate the survival time t_i, it is necessary to consider the cumulative hazard function $H_i(t)$ for individual i under the Weibull baseline hazard function $h_0(t) = \nu t^{\nu-1}$. Let $d_i = d_i(t)$ be the largest integer such that $\widetilde{t}_{id_i} \leq t$. Then, the cumulative hazard function at time t has expression

$$H_i(t) = \widetilde{H}_{id_i} + H_{id_i}\left(\widetilde{t}_{id_i}, t_i, u_{id_i}\right),$$

where in general

$$H_i(t_1, t_2, u) = \int_{t_1}^{t_2} h_i(t)dt = \exp\left(\xi_u \phi + \boldsymbol{w}_i'\boldsymbol{\psi}\right)\left(t_2^{\nu} - t_1^{\nu}\right) \tag{4}$$

and $\widetilde{H}_{i1} = 0$ for $d_i = 1$, while for $d_i > 1$, we have

$$\widetilde{H}_{id_i} = \sum_{j=2}^{d_i} H_i\left(\widetilde{t}_{i,j-1}, \widetilde{t}_{ij}, u_{i,j-1}\right). \tag{5}$$

Note that $H_i(\widetilde{t}_{ij}) = \widetilde{H}_{ij}$ for $j = 1, \ldots, \widetilde{j}_i$. We can easily realize that

$$t = \left[\widetilde{t}_{id_i}^{\nu} + \frac{H_i(t) - \widetilde{H}_{id_i}}{\exp\left(\xi_{u_{id_i}}\phi + \boldsymbol{w}_i'\boldsymbol{\psi}\right)}\right]^{1/\nu}. \tag{6}$$

Therefore, to generate t_i, we first generate $b = -\log(\widetilde{b})$, with $\widetilde{b}$ drawn from a uniform distribution between 0 and 1. Then, we find d_i as the largest integer such that $\widetilde{H}_{id_i} \leq b$ and t_i is taken as the minimum between the results of the application of formula (6) with $H_i(t) = b$ and the censoring time c_i.

Finally, regarding the longitudinal process, we recall that we need a set of observation times $t_{i1}, \ldots, t_{ij_i}$, with $t_{ij_i} < t_i$, and corresponding covariates $\boldsymbol{x}_{ij}$. Then, for each of these observations, we obtain u_{ij} as $u_{i\widetilde{j}}$, where $\widetilde{j}$ is the largest integer such that $\widetilde{t}_{i\widetilde{j}} \leq t_{ij}$. The response variables are generated according to the first equation in (3) depending on the type of variable.

The process is illustrated for a certain individual i in Figure 1 assuming that the censoring time is $c_i = 10$, the continuous-time Markov chain has $k = 3$ states, with initial probabilities $\boldsymbol{\pi} = (0.25, 0.50, 0.25)'$, and matrix $\boldsymbol{Q}$ having all off-diagonal elements equal to 0.5 as in (2). The parameters are fixed at arbitrary values, such as $\nu = 2$, and the covariates are randomly generated from a standard normal distribution. In this example, the generated survival time is $t_i = 4.98$ and $j_i = 5$ continuous responses are observed at the equally spaced time occasions $t_{ij} = j - 1, j = 1, \ldots, 5$.

3 | ESTIMATION

It is straightforward to check that the complete likelihood of the proposed model depends on the entire trajectory of the continuous-time latent process, through the integrals involved in the time-to-event part. This makes it hard to efficiently compute the observed likelihood (ie, classical Baum-Welch recursions are not directly available, even after their extension to continuous-time processes, due to lack of certain conditional independence statements). We have tried different routes, including a Monte Carlo expectation-maximization approach, which involves simulation from the latent process. The novel approach we propose has proved, in our experience, to be the most efficient and stable from a computational perspective.

We now outline our general inferential approach. We initially discuss the case of complete data and then the case of incomplete data. In the first case, latent transition times and the corresponding states are supposed to be known. Obviously this is not realistic, but it is useful to consider this case in order to illustrate the implementation of the estimation for the incomplete data case, where these data are missing.

3.1 | Complete data

The complete data for individual i consist of the following:

1. the time $\widetilde{t}_{ij}$ of any jump of the continuous HM and the corresponding state $\widetilde{u}_{ij}$, with $j = 1, \ldots, \widetilde{j}_i$;

2. the time t_{ij} of every longitudinal observation y_{ij} and the corresponding covariates $\boldsymbol{x}_{ij}$, with $j = 1, \ldots, j_i$;
3. the survival time t_i with indication if this is censored ($\delta_i = 0$) or not ($\delta_i = 1$).

The corresponding log-likelihood function has then three components corresponding to the sets of data described above and depending on sets of parameters that are disjoint, with the only exception of those in $\boldsymbol{\xi}$ that affect both the longitudinal and the survival process. In particular, this log-likelihood function may be expressed as

$$\widetilde{\ell}(\boldsymbol{\theta}) = \widetilde{\ell}_1(\boldsymbol{\pi}, \boldsymbol{Q}) + \widetilde{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2) + \widetilde{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi}),$$

where $\boldsymbol{\theta}$ is a shorthand notation for all parameters involved in the previous expression, and

$$\widetilde{\ell}_1(\boldsymbol{\pi}, \boldsymbol{Q}) = \sum_{i=1}^{n} \widetilde{\ell}_{i1}(\boldsymbol{\pi}, \boldsymbol{Q}),$$
$$\widetilde{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2) = \sum_{i=1}^{n} \widetilde{\ell}_{i2}(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2),$$
$$\widetilde{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi}) = \sum_{i=1}^{n} \widetilde{\ell}_{i3}(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi})$$

because individuals are assumed to be independent of each other and where parameter σ^2 disappears for the binary case.

The first log-likelihood function concerns the continuous-time HM process and is based on the individual components

$$\widetilde{\ell}_{i1}(\boldsymbol{\pi}, \boldsymbol{Q}) = \log \pi_{\widetilde{u}_{i1}} + \sum_{u=1}^{k} \sum_{\substack{v=1 \\ v \neq u}}^{k} \widetilde{n}_{iuv} \log q_{uv} - \sum_{u=1}^{k} \widetilde{s}_{iu} q_u,$$

where for individual i, $\widetilde{n}_{iuv}$ is the number of transitions from state u to state v and $\widetilde{s}_u$ is the time spent in state u. Therefore, we have

$$\widetilde{\ell}_1(\boldsymbol{\pi}, \boldsymbol{Q}) = \sum_{u=1}^{k} \widetilde{n}_u \log \pi_u + \sum_{u=1}^{k} \sum_{\substack{v=1 \\ v \neq u}}^{k} \widetilde{n}_{uv} \log q_{uv} - \sum_{u=1}^{k} \widetilde{s}_u q_u,$$

where $\widetilde{n}_u$ is the number of units in state u at initial time, $\widetilde{n}_{uv}$ is the number of transitions from state u to state v, and $\widetilde{s}_u$ is the time spent in state u overall. Maximization of $\widetilde{\ell}_1(\boldsymbol{\pi}, \boldsymbol{Q})$ with respect to the parameters in $\boldsymbol{\pi}$ and $\boldsymbol{Q}$ simply amounts to compute

$$\pi_u = \frac{\widetilde{n}_u}{n}, \quad u = 1, \ldots, k,$$
$$q_{uv} = \frac{\widetilde{n}_{uv}}{\widetilde{s}_u}, \quad u, v = 1, \ldots, k, \; u \neq v.$$

Regarding the second component for the longitudinal process, we have

$$\widetilde{\ell}_{i2}(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2) = \sum_{j=1}^{j_i} \log f\left(y_{ij}; \mu_{ij}, \sigma^2\right)$$

that, for the normal case and up to a constant term, specifies as

$$\widetilde{\ell}_{i2}(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2) = -\frac{1}{2} \sum_{j=1}^{j_i} \left[\log \sigma^2 + \frac{(y_{ij} - \mu_{ij})^2}{\sigma^2} \right]; \tag{7}$$

in the binary case, it becomes

$$\widetilde{\ell}_{i2}(\boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{j=1}^{j_i} \left[y_{ij} \log \mu_{ij} + (1 - y_{ij}) \log(1 - \mu_{ij}) \right]. \tag{8}$$

In the previous expressions, we have

$$\mu_{ij} = \xi_{u_{ij}} + \boldsymbol{x}_{ij}' \boldsymbol{\beta},$$

where $u_{ij} = \widetilde{u}_{i\widetilde{j}}$, with $\widetilde{j}$ being the largest integer such that $\widetilde{t}_{i\widetilde{j}} \leq t_{ij}$.

Regarding the third component, we have

$$\widetilde{\ell}_{i3}(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi}) = \delta_i \left[\log \nu + (\nu - 1) \log t_i + \xi_{\widetilde{u}_{id_i}} \phi + \boldsymbol{w}_i' \boldsymbol{\psi} \right] - \widetilde{H}_{id_i} - H_i \left(\widetilde{t}_{id_i}, t_i, \widetilde{u}_{id_i} \right), \tag{9}$$

where d_i is the largest integer such that $\widetilde{t}_{id_i} \leq t_i$ and $\widetilde{H}_{id_i}$ is defined in (5).

In practice, to estimate the parameters $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, ν, ϕ, and $\boldsymbol{\psi}$, we use a numerical maximization of $\widetilde{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2) + \widetilde{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi})$ with respect to these parameters, after reparameterizing ν to $\eta = \log \nu$ so as to include the constraint $\nu > 0$, with σ^2 fixed to an arbitrary value for the linear model based on the normal distribution. This maximization is based on the first derivatives of these two functions, which are reported in Appendix. For the linear model, we then estimate σ^2 by the explicit expression

$$\sigma^2 = \frac{1}{\sum_{i=1}^{n} j_i} \sum_{i=1}^{n} \sum_{j=1}^{j_i} (y_{ij} - \mu_{ij})^2.$$

3.2 | Incomplete data

We now consider the realistic case in which we do not observe the time $\widetilde{t}_{ij}$ of any jump of the continuous HM and the corresponding state $\widetilde{u}_{ij}$, which are conceived as missing data. In this case, the likelihood is defined as the integral, with respect to the missing data, of the complete data likelihood introduced in the previous section.

In order to deal with the incomplete data case, we build a sequence of equally spaced time windows corresponding to the fixed time points $\bar{t}_1, \ldots, \bar{t}_M$, with $\bar{t}_1 = 0$ and $\bar{t}_M$ equal to the largest time it is sensible to be considered. These time points are chosen so that each observation time t_{ij} corresponds to one of them and there is at most one observation in each time window. However, the extension to the case of ties, namely, more observations per individual that may be referred to the same time point $\bar{t}_m$, is straightforward. It may also happen that some observation times t_{ij} do not exactly correspond to any time $\bar{t}_m$. In this case, these observations are associated to specific time points $\bar{t}_m$ by an approximating rule. Moreover, for individual i, let m_i be defined so that t_{m_i} is the time of the last longitudinal observation; note that it is also the largest value of m such that $\bar{t}_m \leq t_i$. Finally, we let $\bar{y}_{im}$ denote the observation at time $\bar{t}_m$ for individual i, which may be missing for certain time occasions, let $\overline{\boldsymbol{x}}_{im}$ be the corresponding vector of covariates, and let $\overline{\boldsymbol{y}}_{i,\leq m}$ be vector of observations available until time $\bar{t}_m$.

In the framework described above, for estimation purposes, we approximate the proposed model by a discrete-time HM model that is similar to another joint discrete HM model[6] available in the literature, with the main difference that the survival process is here formulated in a different and more common way on the basis of a hazard function with time-varying latent variables. In this regard, we define latent variable $\overline{U}_{im}$ corresponding to time occasion $\bar{t}_m$, with the initial distribution characterized by the same probability vector $\boldsymbol{\pi}$ defined in Section 2.1 and the $k \times k$ transition matrix $\boldsymbol{\Pi}$, having elements denoted by $\pi_{v|u}$, which is obtained as e^{aQ} with $a = \bar{t}_{m+1} - \bar{t}_m$. Under these assumptions, the complete-data log-likelihood of the approximation model, which would be computable if we knew the latent variables $\overline{U}_{im}$, has again three components as that for the original model illustrated in the previous section. The first component is equal to

$$\overline{\ell}_1(\boldsymbol{\pi}, \boldsymbol{\Pi}) = \sum_{u=1}^{k} \overline{n}_u \log \pi_u + \sum_{u=1}^{k} \sum_{\substack{v=1 \\ v \neq u}}^{k} \overline{n}_{uv} \log \pi_{v|u},$$

with $\overline{n}_u$ being the number of individuals in state u at the beginning of the period of observation and $\overline{n}_{uv}$ being the number of individuals in state u at a certain occasion and in state v at the following occasion. For the normal case and up to a constant term, the second component is

$$\overline{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{m=1}^{m_i} \sum_{u=1}^{k} \overline{z}_{imu} \left\{ \log \sigma^2 + \frac{\left[\overline{y}_{im} - \left(\xi_u + \overline{\boldsymbol{x}}_{im}' \boldsymbol{\beta} \right) \right]^2}{\sigma^2} \right\},$$

which derives from (7), and in the binary case, it becomes

$$\overline{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{i=1}^{n} \sum_{m=1}^{m_i} \sum_{u=1}^{k} \overline{z}_{imu} \left\{ \overline{y}_{im} \left(\xi_u + \overline{\boldsymbol{x}}_{im}' \boldsymbol{\beta} \right) - \log \left[1 + \exp \left(\xi_u + \overline{\boldsymbol{x}}_{im}' \boldsymbol{\beta} \right) \right] \right\},$$

which derives from (8), where $\bar{z}_{imu}$ is an indicator variable equal to 1 if individual i is in latent state u at occasion $\bar{t}_m$. Finally, we have

$$\overline{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi}) = \sum_{i=1}^{n} \left\{ \sum_{u=1}^{k} \bar{z}_{im_i u} \delta_i \left[\log \nu + (\nu - 1) \log t_i + \xi_u \phi + \boldsymbol{w}_i' \boldsymbol{\psi} \right] - \sum_{m=2}^{m_i} \sum_{u=1}^{k} \bar{z}_{i,m-1,u} H_i(\bar{t}_{m-1}, \bar{t}_m, u) - \sum_{u=1}^{k} \bar{z}_{im_i u} H_i(\bar{t}_{m_i}, t_i, u) \right\},$$

which directly derives from (9) and is based on (4).

In order to make likelihood inference, we introduce the following forward recursion that is related to the well-known Baum-Welch recursion.[25,26] In general, the proposed recursions are related to those used for estimating the model-based time-discrete HM model.[6] Consider the joint density

$$f_{im}(u) = f_i(\overline{\boldsymbol{y}}_{i,\leq m} | T_i \geq \bar{t}_m, \overline{U}_{im} = u) p_i(T_i \geq \bar{t}_m, \overline{U}_{im} = u)$$

referred the observations available until time $\bar{t}_m$ for individual i, latent state at the same time occasion, and for the event that the individual survives time $\bar{t}_m$. We have that

$$f_{i1}(u) = \pi_u f_i(\overline{y}_{i1} | \overline{U}_{i1} = u), \quad u = 1, \ldots, k,$$

and

$$f_{im}(v) = f(\overline{y}_{im} | \overline{U}_{i1} = v) \sum_{u=1}^{k} \pi_{v|u} f_{i,m-1}(u) S_i(\bar{t}_{m-1}, \bar{t}_m, u), \quad m = 1, \ldots, m_i, \ v = 1, \ldots, k,$$

where, in general, $S_i(\bar{t}_1, \bar{t}_2, u) = \exp\{-H_i(t_1, t_2, u)\}$ and $f_i(\overline{y}_{im} | \overline{U}_{i1} = v)$ is set equal to 1 if the observation is not available at time $\bar{t}_m$. For individual i, the contribution to the likelihood is given by

$$f_i(\boldsymbol{y}_i, t_i, \delta_i) = \sum_{u=1}^{k} f_{im_i}(u) h(t_i)^{\delta_i} S_i(\bar{t}_{m_i}, t_i, u).$$

Regarding the transition probabilities $\pi_{v|u}$, note that these are the elements of $k \times k$ matrix $\boldsymbol{\Pi}$ introduced above and depending on $\boldsymbol{Q}$.

The log-likelihood function of the approximating model to be maximized is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_i(\boldsymbol{y}_i, t_i, \delta_i).$$

In order to maximize this function, we rely on an EM scheme[41] that some has a structure similar to that used for the corresponding discrete HM model[6] and requires a backward recursion to be used at the E-step. In particular, let

$$g_{im}(u) = f(\overline{\boldsymbol{y}}_{i,>m}, t_i, \delta_i | T_i > t_m, \overline{U}_{im} = u).$$

For $m = m_i$ have that

$$g_{im_i}(u) = h_i(t_i)^{\delta_i} S_i(\bar{t}_{m_i}, t_i, u),$$

and then for $m < m_i$, we have

$$g_{im}(u) = S_i(\bar{t}_m, \bar{t}_{m+1}, u) \sum_{v=1}^{k} \pi_{v|u} g_{i,m+1}(v) f(\overline{y}_{i,m+1}, \overline{U}_{i,m+1} = v), \quad m = 1, \ldots, m_i - 1.$$

From this recursion, we can obtain two posterior distributions used to update the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\Pi}$. In particular, we have that

$$p(\overline{U}_{im} = u | \boldsymbol{y}_i, t_i, \delta_i) = \frac{f_{im}(u) g_{im}(u)}{f_i(\boldsymbol{y}_i, t_i, d_i)}, \quad m = 1, \ldots, m_i, \ u = 1, \ldots, k,$$

and

$$p(\overline{U}_{im} = u, \overline{U}_{i,m+1} = v | \boldsymbol{y}_i, t_i, \delta_i) = \frac{f_{im}(u) S(u, \bar{t}_{m+1}) \pi_{v|u} f(\overline{y}_{i,m+1} \mid \overline{U}_{i,m+1} = v) g_{i,m+1}(v)}{f_i(\boldsymbol{y}_i, t_i, d_i)},$$
$$m = 1, \ldots, m_i - 1, \ u, v = 1, \ldots, k.$$

At the M-step, we update parameters π_u and $\pi_{v|u}$ as follows:

$$\pi_u = \frac{1}{n}\sum_{i=1}^{n} p(\overline{U}_{i1} = u|\boldsymbol{y}_i, t_i, \delta_i), \quad u = 1, \ldots, k,$$

and

$$\pi_{v|u} = \frac{\sum_{i=1}^{n}\sum_{m=1}^{m_i-1} p(\overline{U}_{im} = u, \overline{U}_{i,m+1} = v|\boldsymbol{y}_i, t_i, \delta_i)}{\sum_{i=1}^{n}\sum_{m=1}^{m_i-1}\sum_{\bar{v}=1}^{k} p(\overline{U}_{im} = u, \overline{U}_{i,m+1} = \bar{v}|\boldsymbol{y}_i, t_i, \delta_i)}, \quad u, v = 1, \ldots, k.$$

Then, the infinitesimal transition matrix $\boldsymbol{Q}$ is obtained from $\boldsymbol{\Pi}$ by inverting $e^{a\boldsymbol{Q}}$. To update the other parameters, we explicit the expected value of the complete log-likelihood of the other two components once again after reparameterizing $\eta = \log \nu$. The sum of these two expected values is maximized numerically on the basis of the derivatives that are reported in Appendix.

A crucial point concerns the initialization of the EM algorithm. In particular, we estimate the model starting from a deterministic initial solution and then adopt a multistart strategy to increase the chance of finding the global maximum of the log-likelihood function. Deterministic initial parameters are obtained as follows: a (biased) generalized linear regression model is estimated to obtain initial values for $\boldsymbol{\beta}$. When the outcome is Gaussian, σ^2 is initialized as the variance of the residuals of the latter model. Initial values for the latent intercepts are fixed as the centroids after k-means clustering of the residuals, whereas initial values for $\boldsymbol{\pi}$ are given by the appropriate cluster proportions. Off-diagonal elements of $\boldsymbol{\Pi}$ are initialized as $1/(k-1)$. Similarly, an AFT model is estimated to get initial values for $\boldsymbol{\psi}$ and ν. Coherently, the initial value for ϕ is set to zero. Usually, we repeat model fitting by initializing through additional 50 random starting values, which are obtained by either perturbing the deterministic initial solutions or the current best parameter values at convergence.

Regarding the score vector $\boldsymbol{s}(\boldsymbol{\theta})$, namely, the derivative of $\ell(\boldsymbol{\theta})$ with respect to the model parameters, it is well known that this corresponds to the first derivative of the expected complete log-likelihood.[42] This involves taking the first derivative of the complete data log-likelihood after replacing $\overline{n}_u$ with $\sum_{i=1}^{n} p(\overline{U}_{i1} = u|\boldsymbol{y}_i, t_i, \delta_i)$ and $\overline{n}_{uv}$ with $\sum_{i=1}^{n} p(\overline{U}_{im} = u, \overline{U}_{i,m+1} = v|\boldsymbol{y}_i, t_i, \delta_i)$; additionally, $\overline{z}_{imu}$ shall be replaced with $p(\overline{U}_{im} = u|\boldsymbol{y}_i, t_i, \delta_i)$. A closed-form expression is available for this derivative, which is reported in Appendix.

Note that the score vector $\boldsymbol{s}(\boldsymbol{\theta})$, obtained as above, may be used for different purposes. First of all it, allows us to check convergence of the EM algorithm in a precise way and implement a quasi-Newton algorithm to maximize $\ell(\boldsymbol{\theta})$ in a faster way, once the reliable starting values for the parameters have been found through some iterations of the EM algorithm. In this regard, the EM and the quasi-Newton algorithms can be combined to obtain, overall, an algorithm that is stable and fast at the same time. Secondly, by its numerical derivative, we can obtain the observed information matrix $\boldsymbol{J}(\boldsymbol{\theta})$ as in a related work[43]; this matrix may be used to obtain standard errors for the parameter estimates and check identifiability. Standard errors are obtained as the diagonal elements of the inverse of $\boldsymbol{J}(\widehat{\boldsymbol{\theta}})$, whereas a Delta method can be used to obtain the standard error for ν and other transformed parameters. Standard errors are then used for Wald tests and confidence intervals involving the parameters of the manifest distribution. A cautionary note is necessary for binomial outcomes, where, in certain cases of low information, the distribution of the estimator might be skewed. In this case, resampling (eg, bootstrap for confidence intervals and permutation for testing) might be more appropriate. For the parameters of the hidden distribution, a likelihood ratio test shall be used, where the null distribution in certain limiting cases is a chi-bar squared.[44] Regarding identifiability, we are referring to its local version[44-46] that is typically used for latent variable models and that is necessary for typical asymptotic properties of the maximum likelihood estimator. In particular, the model is locally identifiable if the observed information matrix is of full rank, which is also necessary to obtain the standard errors.

A final point concerns model selection in terms of the number of latent states (k); in this regard, we rely on the usual Bayesian information criterion (BIC)[47] based on minimization of the index

$$BIC = -2\ell(\widehat{\boldsymbol{\theta}}) + g\log(n),$$

where $\ell(\widehat{\boldsymbol{\theta}})$ is the maximum log-likelihood of the model of interest and g is the number of free parameters.

Once parameter estimates have been obtained, the latent trajectory of each subject can be predicted by maximizing $p(\overline{U}_{im} = u|\boldsymbol{y}_i, t_i, \delta_i)$ in u for each $m = 1, \ldots, m_i$. This is the so-called local decoding, which minimizes the misclassification error at each $\bar{t}_m$. In order to perform global decoding, that is, optimization of the subject-specific sequence of latent states, one could set up a Viterbi-type algorithm. We have found that a numerical approach which exploits the simulation

algorithm proposed in Section 2.2 is much simpler and more flexible. The latent process can be simulated and extrapolated for forecasting both the latent and manifest distribution, and the operation can be repeated several times therefore providing a distribution for the desired quantities. A simulation from the latent process, together with the covariates, is all that is needed in order to obtain a plug-in prediction/forecast of the longitudinal response at observed or future time occasions. On the other hand, the prediction of the survival probabilities is slightly more cumbersome and grounds for further work (see also the discussion in the last section).

4 | SIMULATION STUDY

We describe a brief simulation study of our approach that is based on 1000 samples simulated from the proposed model under different scenarios. Each scenario is characterized by a certain sample size (n), number of latent states (k), level of separation between these states, and type of distribution for the response variables. We always consider a censoring equal to 10 and we allow for 11 follow-up occasions at most.

The simulation design is based on two baseline covariates (collected in $\boldsymbol{w}_i$), which are generated from two independent standard normal distributions. The two covariates in $\boldsymbol{x}_{it}$ are generated from two independent AR(1) processes with autocorrelation coefficient 0.9 and variance of the Gaussian innovations equal to 0.19, so that the each covariate has the same standard normal distribution marginally. Regarding the model parameters, the latent support points are chosen as $\xi_u = \omega[u-(k+1)/2], u = 1, \dots, k$, for different values of the *separation parameter* $\omega = 1.5, 3.0$. We also have $\pi_u = 1/k$, $\boldsymbol{Q}$ with off-diagonal elements equal to $1/(k-1)$, $\boldsymbol{\beta} = (-1, 1)'$, $\phi = 0.5$, $\nu = 2$, and $\boldsymbol{\psi} = (-1, 1)'$. Data are generated according to the model described in Section 2.1, where, for the longitudinal component, we either use a Bernoulli distribution for binary outcomes or a normal distribution for continuous outcomes. The overall number of scenarios considered in the simulation study is 16 as we use every possible combination of $n = 250, 500$, $k = 2, 3$, $\omega = 1.5, 3$, binary or continuous longitudinal outcomes.

For each simulated data set, we estimated our joint continuous latent Markov (JCLM) model and, for comparison, two restricted models: a latent class (LC) model that rules out latent transitions (all elements of $\boldsymbol{Q}$ are constrained to 0) and a continuous-time latent Markov (CLM) model ignoring dependence between the longitudinal component and the survival

TABLE 1 Simulation study results averaged over 1000 replicates: root-mean-square error (RMSE), bias, and standard deviation of the estimates for estimation of $\boldsymbol{\beta}$ with the proposed model (JCLM), a joint latent class model (LC), and a model assuming ignorable dropout (CLM), for different values of n, k, separation among latent intercepts ω, and outcome distribution (either Bernoulli or normal). The last column reports the coverage of 95% confidence intervals using the JCLM model and estimated standard errors

				RMSE			Bias			Std. Dev.		Coverage	
n	k	ω	Y	JCLM	LC	CLM	JCLM	LC	CLM	JCLM	LC	CLM	JCLM
500	2	1.5	binary	0.152	0.106	0.163	−0.077	0.000	−0.087	0.131	0.106	0.137	0.947
500	2	1.5	norm	0.060	0.061	0.062	−0.002	0.000	−0.002	0.060	0.061	0.062	0.957
500	2	3.0	binary	0.154	0.309	0.179	0.006	0.000	−0.040	0.154	0.309	0.174	0.969
500	2	3.0	norm	0.073	0.079	0.084	−0.005	−0.003	−0.004	0.073	0.079	0.084	0.943
500	3	1.5	binary	0.203	0.210	0.291	−0.044	0.001	−0.117	0.199	0.210	0.267	0.941
500	3	1.5	norm	0.066	0.069	0.072	−0.001	−0.001	−0.001	0.066	0.069	0.072	0.936
500	3	3.0	binary	0.298	0.468	0.463	0.141	0.000	−0.034	0.262	0.468	0.462	0.960
500	3	3.0	norm	0.081	0.083	0.100	0.001	−0.002	0.002	0.081	0.083	0.100	0.947
1000	2	1.5	binary	0.099	0.101	0.106	−0.057	0.000	−0.062	0.081	0.101	0.086	0.936
1000	2	1.5	norm	0.043	0.043	0.045	−0.001	0.000	0.000	0.043	0.043	0.045	0.944
1000	2	3.0	binary	0.106	0.308	0.115	0.030	0.000	−0.005	0.102	0.308	0.115	0.953
1000	2	3.0	norm	0.051	0.054	0.058	−0.001	0.000	0.001	0.051	0.054	0.058	0.966
1000	3	1.5	binary	0.112	0.210	0.168	0.001	0.000	−0.046	0.112	0.210	0.161	0.943
1000	3	1.5	norm	0.048	0.048	0.053	0.002	0.000	−0.001	0.048	0.048	0.053	0.937
1000	3	3.0	binary	0.230	0.467	0.284	0.202	−0.001	0.138	0.110	0.467	0.249	0.942
1000	3	3.0	norm	0.055	0.057	0.065	−0.002	−0.002	−0.002	0.055	0.057	0.065	0.943

component (ie, assuming dropout to be ignorable). In our experiments, we have set $M = 45$ in order to show that, even with small values of M, we obtain a very good performance in terms of root-mean-square error (RMSE). More details about this point are given at the end of the section based on an additional simulation study.

In Table 1, we report average RMSE, bias, and standard deviation of the estimates for the $\boldsymbol{\beta}$ parameters. We do not report results for other parameters as not all models estimate those, and furthermore, our primary target is the accurate estimation of the effect of covariates for the longitudinal model. Moreover, in order to show the validity of the proposed method for estimation of the standard errors, in the last column of Table 1 we report the coverage of the 95% Wald confidence intervals based on our standard errors obtained by the procedure described in the previous section.

On the basis of the results, we conclude that the proposed estimation method performs adequately and that the proposed JCLM model is advantageous in terms of RMSE and standard error of the parameter estimates most of the times. Ignoring informative dropout might lead to clearly worse results while using a latent class model might have, in some cases, a better performance. This is related to (*i*) small sample sizes/lack of information, which might advantage more parsimonious models, and (*ii*) convergence to local optima for the JCLM in certain iterations. The second problem is easily faced through a multistart strategy, while, for computational reasons in our simulation study, we have used only one initial solution. Finally, the coverage of the 95% confidence intervals is always close to the nominal level.

We now report an additional simple simulation study which is aimed at illustrating the convergence properties of our algorithm as the tuning parameter M is changed. We generated two data sets, one with a binary and the other with a Gaussian outcome, according to the scheme above. We set $n = 1000$, $k = 2$, $\omega = 3$, and repeatedly estimated the model with $M = 16, 21, 26, \ldots, 106$. In Figure 2, we report, for each value of M, the log-likelihood at convergence and parameter estimates. Left panels refer to the Gaussian longitudinal outcome and right panels to the binary longitudinal outcome.

It can be seen that, even for small values of M, certain parameters are clearly well estimated, and their value at convergence slightly depends on M. In contrast, other parameters, especially with binary outcomes, are more dependent on the choice of the grid density. Even for these parameters, if M is above a certain threshold, stability (and, substantially, consistency) is obtained. An *a priori* assessment of what values of M are "large enough" is not possible in our experience. We recommend to perform a study similar to the one above in real applications to fix M.

5 | APPLICATION TO MDCM DATA

We now illustrate the proposed approach using an original study on a cohort of patients affected by MDCM,[48,49] a primary myocardial disease characterized by left ventricular systolic dysfunction and dilation.

Prognostic measurements were taken at basal time for $n = 642$ patients, who were followed-up until urgent heart transplant or death occurred. There were 212 events during the follow-up, which lasted up to 25 years. If censoring (administrative or due to the event) did not occur, measurement of longitudinal biomarkers was taken at visits scheduled at months 6, 12, 24, 48, 72, 96, and 120. Hence, each patient has a maximum of eight longitudinal measurements, with 79 patients having complete records.

The longitudinal outcome derives from the NYHA classification, a direct measure of discomfort caused by the disease. Specifically, for each subject at each follow-up occasion, an indicator of being in NYHA class III or IV was measured, indicating the presence of strong limitations to physical activity and/or the occurrence of dyspnea and discomfort during ordinary activities or even at rest.

For the longitudinal model, we parameterize probability of high NYHA class as a function of time (indicating medical treatment according to international guidelines), an indicator of history of heart disease in the family, and the left ventricular ejection fraction (EF). The latter is a measure of the proportion of blood that is pumped out of the left ventricle at each heart beat. Primary interests are in (*i*) relating predictors to QoL (as measured by NYHA class) and to risk of event, (*ii*) identifying subgroups of patients at higher risk, and (*iii*) summarizing for each group the overall risk of NYHA III or IV after beginning of medical treatment for the condition.

A continuous-time model should be more appropriate for the data at hand than any model assuming latent transitions occurring at visit times. In fact, latent states shall be interpreted as patients' frailty beyond that summarized by the predictors, and changes in disease status (and hence propensity to event and/or change of NYHA class) obviously can occur at any time point and not necessarily on the day of scheduled for the follow-up visit. Furthermore, a strong dependence between NYHA class and the event is expected, with patients in NYHA classes III or IV being at higher risk of death. For interpretability reasons, EF has been centered at 30 (which is believed to be a significant threshold for heart failure).

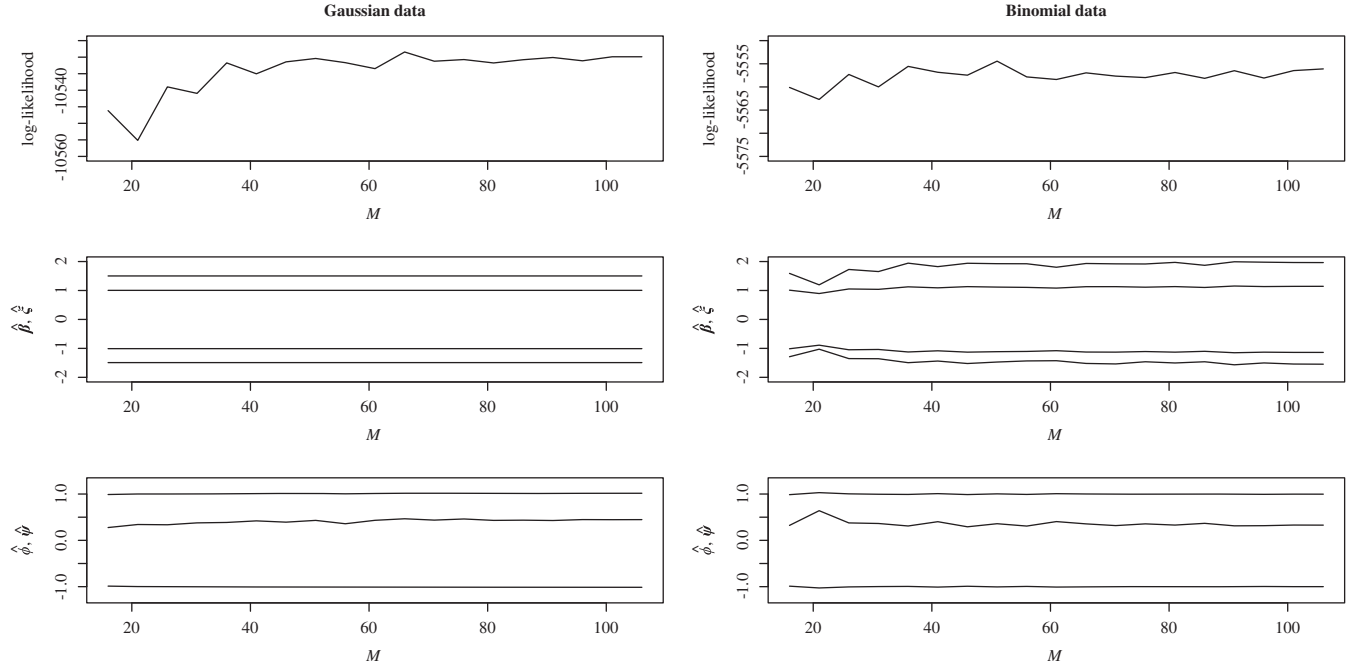

FIGURE 2 Simulated Gaussian and binomial data: log-likelihood at convergence and parameter estimates as a function of M

TABLE 2 Log-likelihood at convergence ($\ell(\widehat{\boldsymbol{\theta}})$), number of free parameters (g), Bayesian information criterion (*BIC*), computational time (in seconds) for the proposed model fit to MDCM data and obtain standard errors, for different values of k

k	$\ell(\widehat{\boldsymbol{\theta}})$	g	*BIC*	time
1	−2739.60	8	5537.38	0.88
2	−2509.49	12	5096.56	68.90
3	−2480.28	18	5076.93	110.72
4	−2468.07	26	5104.22	233.49

For our model fitting procedure, we evaluate several values for M. After a study similar to the one reported in the previous section, we end up fixing $M = 200$, which is well above values guaranteeing stability of estimates. In Table 2, we report the log-likelihood at convergence, number of parameters, BIC, and computational time (in seconds) for fitting the proposed model to the MDCM data with different values of k. We report, for each k, the largest log-likelihood obtained after fitting the model with 25 different initial solutions and running time in seconds for the reported solution (including computation of standard errors).

From the results in Table 2, we select $k = 3$. In order to estimate standard errors, we use the observed information as discussed above. We can confirm that the model is locally identifiable as $\boldsymbol{J}(\widehat{\boldsymbol{\theta}})$ is full rank. In Table 3, we report parameter estimates under this model for the manifest distribution, along with standard errors in parentheses. Furthermore, in Table 4, we report the other estimated parameters for the case $k = 3$. For $\boldsymbol{Q}$, we chose not to report standard errors but to provide a likelihood ratio test statistic for case $q_{cd} = 0, c = 1, \ldots, k, d = 1, \ldots, k$, which, in this case, is 122.60 for a difference of six free parameters. This provides clear evidence that latent transitions do occur during the observation period.

The estimate of $\boldsymbol{Q}$ is better understood after computation of the time-specific (inhomogeneous) transition matrices, the elements of which are represented in Figure 3.

The results indicate an important role of all predictors, with the exception of history of hearth disease for survival. Comparing $k = 1$ with $k > 1$, it is clear that taking into account unobserved heterogeneity leads to a clearer identification of the roles of EF and family history for NYHA classification. The effect of family history doubles when passing from $k = 1$ to $k = 3$, and the effect of each percentage point of EF is almost three times larger. Hence, based on our results, doctors

TABLE 3 Mildly dilated cardiomyopathy data: parameter estimates for the manifest distribution, different values of k. An asterisk indicates statistical significance at the 5% level

		k			
	Effect	**1**	**2**	**3**	**4**
Logit NYHA	ξ_1	−1.635 (0.169)	−4.745 (0.366)	−4.556 (0.764)	−6.473 (1.175)
	ξ_2	-	−0.164 (0.167)	−1.429 (0.323)	−2.300 (0.700)
	ξ_3	-	-	2.796 (1.447)	−1.475 (0.340)
	ξ_4	-	-	-	2.966 (0.614)
	$t > 0$	1.591 (0.160)	2.289 (0.216)	0.905 (0.356)	0.850 (0.343)
	history	0.926 (0.273)	0.724 (0.267)	1.171 (0.317)	1.129 (0.311)
	EF	0.109 0.0123	0.094 (0.011)	0.132 (0.024)	0.139 (0.016)
Survival	ϕ	0.000	−0.475 (0.089)	−0.320 (0.089)	−0.303 (0.064)
	history	−0.125 (0.218)	0.031 (0.191)	−0.022 (0.189)	−0.002 (0.201)
	EF	−0.058 (0.009)	−0.048 (0.008)	−0.049 (0.008)	−0.058 (0.009)
	ν	0.799 (0.050)	0.781 (0.068)	0.761 (0.067)	0.777 (0.072)

Abbreviations: EF, ejection fraction; NYHA, New York Hearth Association.

TABLE 4 Mildly dilated cardiomyopathy data: parameter estimates for the latent distribution when $k = 3$

π	Q		
0.282 (0.045)	−0.001	0.001	0.000
0.584 (0.054)	0.001	−0.050	0.049
0.133 (0.051)	0.005	0.002	−0.007

should probably pay more attention to EF and family history than expected when assessing prognosis to high NYHA classes. The estimate of ϕ is negative and significant in all cases indicating, as expected, that subjects with, for instance, dyspnea during ordinary activities are at higher risk of death than patients without clear signs of heart insufficiency. The fact that this holds also after adjusting for EF indicates that NYHA class *trajectory* might provide prognostic information beyond that linked to proximal heart failure.[50]

When $k = 3$, three clearly separate groups of patients are identified. Even when they have the same history, EF, and timing configuration, patients might be different due to unobserved factors. A group of patients (about 30% at baseline time) is at very low risk. From Figure 3, it can be seen that this group of patients is slightly stable, with low probability of transition to different states during the follow-up. The second group (about 60%) has a slightly larger propensity to high MDCM at baseline. These patients are very likely to change state across time, with many switches to a higher risk (especially in the period 15-40 months from the baseline) and the rest of switches to the low risk (first) latent state (possibly due to successful medical treatment). Finally, the third group of patients is at very high risk of high NYHA class at baseline time. Most of them remain at high risk during the follow-up, but a slight proportion switches to better propensity states, surprisingly more often to state 1 than to state 2. This might be due to the increased medical attention given to high risk patients.

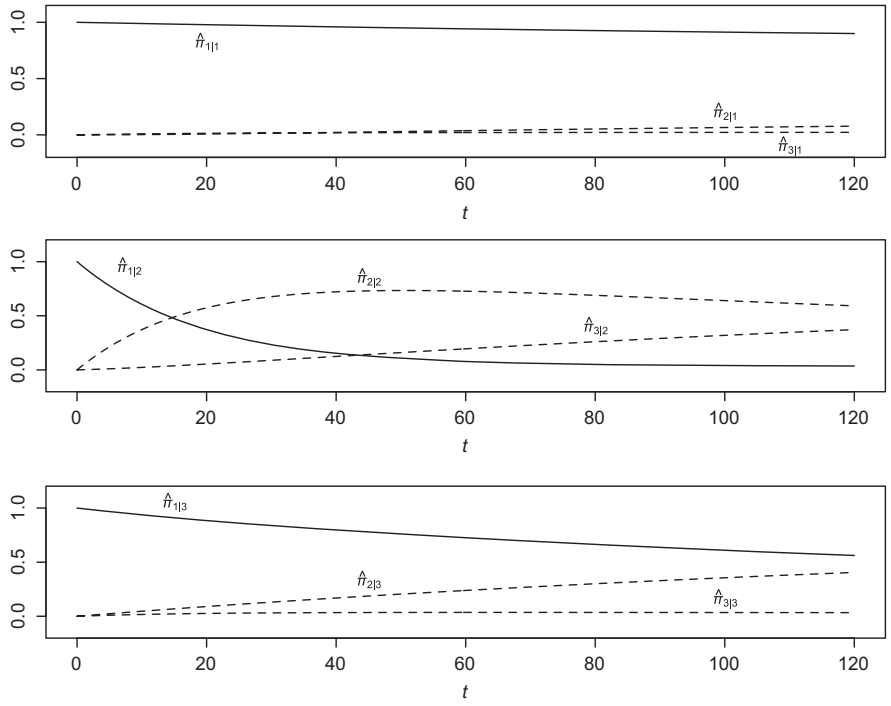

FIGURE 3 Mildly dilated cardiomyopathy data: representation of the estimated elements of the transition matrix across time when $k = 3$

6 | CONCLUSIONS

We have introduced a general latent Markov model based on a continuous-time latent distribution, allowing for informative dropout through a longitudinal-survival shared-parameter model. A particularly appealing feature of our approach is that it allows for latent transitions, which correspond to time-dependent changes in unobserved heterogeneity (eg, disease progression or improvement not explained by available covariates), to occur at any time during the follow-up and not only at follow-up times. Note that the advantage of a continuous-time latent process is mostly linked to the joint modeling with a time-to-event manifest process; otherwise, latent transitions can be still efficiently tackled through a discrete-time process (and inhomogeneous transition matrices). Nevertheless, continuous-time processes might more flexibly and parsimoniously accommodate nonequally spaced time intervals than inhomogeneous transition matrices in discrete time.[13]

A novel inferential procedure has been proposed to fit the model, which we believe might be of independent interest in the more general context of fitting continuous-time hidden Markov models. It shall be noted that with moderately small sample sizes (or high censoring proportion), the log-likelihood surface might be slightly flat for some parameters, leading to some risk of incurring in a local optimum. In our experience, the problem can be simply solved through repeated model fitting from different initial parameter values, but the obvious price is an increased computing time.

Estimation of the most likely hidden trajectory (jump times and states) for each subject is straightforward and might be of independent interest for solving problems in precision medicine. Relatedly, an open issue for further work is the development of methods for producing dynamic predictions of quantities of interest (eg, probability of death within a given time frame). The specification of joint models allows the researcher to update predictions by taking into account all measurements up to the most recent one.[51] There are also two specific assumptions that could be relaxed in further work: the first is that observation times within the longitudinal process are noninformative; the second is that censoring is independent of the other data generating mechanisms, including unobserved outcomes. The first assumption is clearly tenable as soon as observation times are scheduled in advance, and this happens in our application. The second assumption is a rather common one and is tenable in our example as censoring is administrative, that is, follow-up was ended for all patients still in the study on the date in which the study was closed. Finally, note that extension to more than one longitudinal outcomes is rather straightforward while more work would be needed to include additional or different kinds of time-to-event outcome (eg, repeated events, multistate processes, and competing risks).

ACKNOWLEDGEMENTS

The authors are grateful to the Associate Editor and one Referee for useful comments. The authors also thank the Cardiovascular Department of "Ospedali Riuniti," Trento, Italy and, in particular, Giulia Barbati for the permission to use the dilated cardiomyopathy data.

ORCID

Alessio Farcomeni http://orcid.org/0000-0002-7104-5826

REFERENCES

1. Wu MC, Carroll RJ. Estimation and comparison of changes in presence of informative right censoring by modelling the censoring process. *Biometrics*. 1988;44:175-188.
2. Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics*. 1995;51:151-168.
3. Wulfsohn M, Tsiatis A. A joint model for survival and longitudinal data measured with error. *Biometrics*. 1997;53:330-339.
4. Rizopoulos D. JM: an R package for the joint modelling of longitudinal and time-to-event data. *J Stat Softw*. 2010;35:1-33.
5. Roy J. Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*. 2003;59:829-836.
6. Bartolucci F, Farcomeni A. A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. *Biometrics*. 2015;71:80-89.
7. Tsonaka R, Verbeke G, Lesaffre E. A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics*. 2009;65:81-87.
8. Spagnoli A, Henderson R, Boys RJ, Houwing-Duistermaat JJ. A hidden Markov model for informative dropout in longitudinal response data with crisis states. *Stat Probab Lett*. 2011;81:730-738.
9. Maruotti A. Handling non-ignorable dropouts in longitudinal data: a conditional model based on a latent Markov heterogeneity structure. *TEST*. 2015;24:84-109.
10. Marino MF, Alfó M. Latent drop-out based transitions in linear quantile hidden Markov models for longitudinal responses with attrition. *Adv Data Anal Classif*. 2015;9:483-502.
11. Marino MF, Tzavidis N, Alfó M. Mixed hidden Markov quantile regression models for longitudinal data with possibly incomplete sequences. *Stat Methods Med Res*. 2018;27:2231-2246.
12. Liggett TM. *Continuous Time Markov Processes: An Introduction*. Providence, RI: American Mathematical Society; 2010.
13. Böckenholt U. A latent Markov model for the analysis of longitudinal data collected in continuous time: states, durations, and transitions. *Psychol Methods*. 2005;10:65-83.
14. Titman A, Sharples L. Semi-Markov models with phase-type sojourn distributions. *Biometrics*. 2010;66:742-752.
15. Lange JM, Minin VN. Fitting and interpreting continuous-time latent Markov models for panel data. *Statist Med*. 2013;32:4581-4595.
16. Lange JM, Hubbard RA, Inoue LYT, Minin VN. A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics*. 2015;71:90-101.
17. Bartolucci F, Lupparelli M, Montanari GE. Latent Markov models for longitudinal binary data: an application to the performance evaluation of nursing homes. *Ann Appl Stat*. 2009;3:611-636.
18. Bartolucci F, Farcomeni A, Pennoni F. *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: Chapman & Hall/CRC Press; 2012.
19. Bartolucci F, Farcomeni A, Pennoni F. Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *TEST*. 2014;23:433-465.
20. Farcomeni A. Generalized linear mixed models based on latent Markov heterogeneity structures. *Scand J Stat*. 2015;42:1127-1135.
21. Guo W. Functional mixed effects models. *Biometrics*. 2002;58:121-128.
22. Guo W. Functional data analysis in longitudinal settings using smoothing splines. *Stat Methods Med Res*. 2004;13:49-62.
23. James GM. Generalized linear models with functional predictors. *J Royal Stat Soc, Ser B*. 2002;64:411-432.
24. Zhang D, Lin X, Sowers M. Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome. *Biometrics*. 2007;63:351-362.
25. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann Math Stat*. 1970;41:164-171.
26. Welch LR. Hidden Markov models and the Baum-Welch algorithm. *IEEE Inf Theory Soc Newsl*. 2003;53:10-13.
27. Metzner P, Horenko I, Schütte C. Generator estimation of Markov jump processes based on incomplete observations nonequidistant in time. *Phys Rev E*. 2007;76:066702.
28. Liu YY, Li S, Li F, Song L, Rehg JM. Efficient learning of continuous-time hidden Markov models for disease progression. Paper presented at: 2015 Neural Information Processing Systems Conference; 2015; Montreal, Canada.
29. Coelho R, Ramos S, Prata J, Bettencourt P, Ferreira A, Cerqueira-Gomes M. Heart failure and health related quality of life. *Clin Pract Epidemiol Ment Health*. 2005;1:19.

30. Athanasopoulos LV, Dritsas A, Doll HA, Cokkinos DV. Comparative value of NYHA functional class and quality-of-life questionnaire scores in assessing heart failure. *J Cardiopulm Rehabil Prev*. 2010;30:101-105.
31. Unkel S, Farrington CP, Whitaker HJ, Pebody R. Time varying frailty models and the estimation of heterogeneities in transmission of infectious diseases. *J Royal Stat Soc, Ser C*. 2014;63:141-158.
32. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. New York, NY: Springer; 2000.
33. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc Ser B*. 1972;34:187-220.
34. Andersen P, Gill R. Cox's regression model for counting processes: a large sample study. *Ann Stat*. 1982;10:1100-1120.
35. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. London, UK: Chapman and Hall/CRC; 1989.
36. Conover WJ, Iman RL. Analysis of covariance using the rank transformation. *Biometrics*. 1982;38:715-724.
37. Pirone L, Bragonzi A, Farcomeni A, et al. *Burkholderia cenocepacia* strains isolated from cystic fibrosis patients are apparently more invasive and more virulent than rhizosphere strains. *Environ Micriobiology*. 2008;10:2773-2784.
38. Finos L, Farcomeni A. *k*-FWER control without *p*-value adjustment, with application to detection of genetic determinants of multiple sclerosis in Italian twins. *Biometrics*. 2011;67:174-181.
39. Farcomeni A, Finos L. FDR control with pseudo-gatekeeping based on a possibly data driven order of the hypotheses. *Biometrics*. 2013;69:606-613.
40. Rizopoulos D, Verbeke G, Molenberghs G. Shared parameter models under random effects misspecification. *Biometrika*. 2008;95:63-74.
41. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J R Stat Soc Ser B*. 1977;39:1-38.
42. Bartolucci F, Farcomeni A. Information matrix for hidden Markov models with covariates. *Stat Comput*. 2015;25:515-526.
43. Bartolucci F, Farcomeni A. A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *J Am Stat Assoc*. 2009;104:816-831.
44. Bartolucci F. Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *J R Stat Soc Ser B*. 2006;68:155-178.
45. McHugh RB. Efficient estimation and local identification in latent class analysis. *Psychometrika*. 1956;21:331-347.
46. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*. 1974;61:215-231.
47. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6:461-464.
48. Farcomeni A, Viviani S. Longitudinal quantile regression in presence of informative dropout through longitudinal-survival joint modeling. *Statist Med*. 2015;34:1199-1213.
49. Gigli M, Stolfo D, Merlo M, et al. Insights into mildly dilated cardiomyopathy: temporal evolution and long-term prognosis. *Eur J Heart Fail*. 2017;19:531-539.
50. Barbati G, Farcomeni A. Prognostic assessment of repeatedly measured time-dependent biomarkers, with application to dilated cardiomyopathy. *Stat Methods Appl*. 2018;27:545-557.
51. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*. 2011;67:819-829.

How to cite this article: Bartolucci F, Farcomeni A. A shared-parameter continuous-time hidden Markov and survival model for longitudinal data with informative dropout. *Statistics in Medicine*. 2018;1–19. https://doi.org/10.1002/sim.7994

APPENDIX

A general result that we will use concerns the derivative of the function defined in (4); we have that

$$\begin{aligned}
\frac{\partial H_i(t_1,t_2,u)}{\partial \xi_u} &= H_i(t_1,t_2,u)\phi,\\
\frac{\partial H_i(t_1,t_2,u)}{\partial \phi} &= H_i(t_1,t_2,u)\xi_u,\\
\frac{\partial H_i(t_1,t_2,u)}{\partial \boldsymbol{\psi}} &= H_i(t_1,t_2,u)\boldsymbol{w}_i,\\
\frac{\partial H_i(t_1,t_2,u)}{\partial \eta} &= \nu \exp\left(\xi_u\phi + \boldsymbol{w}_i'\boldsymbol{\psi}\right)\left(t_2^{\nu}\log t_2 - t_1^{\nu}\log t_1\right),
\end{aligned}$$

where we recall that $\eta = \log \nu$.

Derivatives used for estimation in the complete case

We report the derivatives of the complete log-likelihood, with respect to certain model parameters, which are used for the numerical maximization described in Section 3.1. For the first derivative of the second component, we have, in the normal case, that

$$\frac{\partial \widetilde{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{j=1}^{j_i} \frac{y_{ij} - \mu_{ij}}{\sigma^2} \boldsymbol{x}_{ij},$$

$$\frac{\partial \widetilde{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2)}{\partial \xi_u} = \sum_{i=1}^{n} \sum_{j=1}^{j_i} \frac{y_{ij} - \mu_{ij}}{\sigma^2} I(\widetilde{u}_{ij} = u), \quad u = 1, \ldots, k.$$

For the binary case, we have

$$\frac{\partial \widetilde{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{j=1}^{j_i} (y_{ij} - \mu_{ij}) \boldsymbol{x}_{ij},$$

$$\frac{\partial \widetilde{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \xi_u} = \sum_{i=1}^{n} \sum_{j=1}^{j_i} (y_{ij} - \mu_{ij}) I(\widetilde{u}_{ij} = u), \quad u = 1, \ldots, k.$$

Regarding the first derivative of the third component, we have

$$\frac{\partial \widetilde{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi})}{\partial \eta} = \sum_{i=1}^{n} \left[\delta_i (1 + \nu \log t_i) - \sum_{j=2}^{d_i} \frac{\partial H_i(\widetilde{t}_{i,j-1}, \widetilde{t}_{ij}, \widetilde{u}_{i,j-1})}{\partial \eta} - \frac{\partial H_i(\widetilde{t}_{id_i}, t_i, \widetilde{u}_{id_i})}{\partial \eta} \right]$$

and

$$\frac{\partial \widetilde{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi})}{\partial \xi_u} = \phi \sum_{i=1}^{n} \left[\delta_i I(\widetilde{u}_{id_i} = u) - \sum_{j=2}^{d_i} H_i(\widetilde{t}_{i,j-1}, \widetilde{t}_{ij}, \widetilde{u}_{i,j-1}) I(\widetilde{u}_{i,j-1} = u) - H_i(\widetilde{t}_{id_i}, t_i, \widetilde{u}_{id_i}) I(\widetilde{u}_{id_i} = u) \right], \quad u = 1, \ldots, k.$$

Finally, we have

$$\frac{\partial \widetilde{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi})}{\partial \phi} = \sum_{i=1}^{n} \left[\delta_i \xi_{\widetilde{u}_{id_i}} - \sum_{j=2}^{d_i} H_i(\widetilde{t}_{i,j-1}, \widetilde{t}_{ij}, \widetilde{u}_{i,j-1}) \xi_{\widetilde{u}_{i,j-1}} - H_i(\widetilde{t}_{id_i}, t_i, \widetilde{u}_{id_i}) \xi_{\widetilde{u}_{id_i}} \right],$$

and

$$\frac{\partial \widetilde{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \sum_{i=1}^{n} \left[\delta_i - \sum_{j=2}^{d_i} H_i(\widetilde{t}_{i,j-1}, \widetilde{t}_{ij}, \widetilde{u}_{i,j-1}) - H_i(\widetilde{t}_{id_i}, t_i, \widetilde{u}_{id_i}) \right] \boldsymbol{w}_i.$$

Derivatives used for estimation in the incomplete case

We report the derivatives of the complete log-likelihood with respect to the model parameters. Regarding the first component, we reparametrize vector $\boldsymbol{\pi}$ by the multinomial logits λ_u collected in the column vector $\boldsymbol{\lambda}$ e defined as

$$\lambda_u = \log \frac{\pi_{u+1}}{\pi_1}, \quad u = 1, \ldots, k-1;$$

we also reparametrize the off-diagonal elements of $\boldsymbol{\Pi}$ by the logits in each row defined as

$$\gamma_{uv} = \log \frac{\pi_{v|u}}{\pi_{u|u}}, \quad u, v = 1, \ldots, k, u \neq v.$$

The logits referred to the same row u are collected in the vector $\boldsymbol{\gamma}_u$. With respect to these parameters, we have

$$\frac{\partial \widetilde{\ell}_1(\boldsymbol{\pi}, \boldsymbol{Q})}{\boldsymbol{\lambda}} = \boldsymbol{G}_1'(\overline{\boldsymbol{n}} - n\boldsymbol{\pi}),$$

where $\overline{\boldsymbol{n}}$ is the column vector with elements $\overline{n}_u, u = 1, \ldots, k$, and, in general, $\boldsymbol{G}_u$ is obtained by removing the uth column from an identity matrix of size k. We also have

$$\frac{\partial \widetilde{\ell}_1(\boldsymbol{\pi}, \boldsymbol{\Pi})}{\partial \boldsymbol{\gamma}_u} = \boldsymbol{G}_u'(\overline{\boldsymbol{n}}_u - \overline{n}_{u+}\boldsymbol{\pi}_u), \quad u = 1, \ldots, k,$$

where $\overline{\boldsymbol{n}}_u$ contains the elements $\overline{n}_{uv}$ for $v = 1, \ldots, k$, $\overline{n}_{u+}$ is the sum of these elements, and $\boldsymbol{\pi}_u$ contains the probabilities in the uth row of $\boldsymbol{\Pi}$.

For the first derivative of the second component in the normal case, we have

$$\frac{\partial \overline{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{m=1}^{m_i} \sum_{u=1}^{k} \overline{z}_{imu} \frac{\overline{y}_{im} - \left(\xi_u + \overline{\boldsymbol{x}}'_{im}\boldsymbol{\beta}\right)}{\sigma^2} \boldsymbol{x}_{ij},$$

$$\frac{\partial \overline{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2)}{\partial \xi_u} = \sum_{i=1}^{n} \sum_{m=1}^{m_i} \overline{z}_{imu} \frac{\overline{y}_{im} - \left(\xi_u + \overline{\boldsymbol{x}}'_{im}\boldsymbol{\beta}\right)}{\sigma^2}, \quad u = 1, \ldots, k,$$

$$\frac{\partial \overline{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma^2)}{\partial \zeta} = -\frac{1}{2} \sum_{i=1}^{n} \sum_{m=1}^{m_i} \sum_{u=1}^{k} \overline{z}_{imu} \left\{ 1 - \frac{\left[\overline{y}_{im} - \left(\xi_u + \overline{\boldsymbol{x}}'_{im}\boldsymbol{\beta}\right)\right]^2}{\sigma^2} \right\},$$

where we use the reparametrization $\zeta = \log \sigma^2$. For the binary case, we have

$$\frac{\partial \overline{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{m=1}^{m_i} \sum_{u=1}^{k} \overline{z}_{imu} (y_{ij} - \overline{\mu}_{iu}) \boldsymbol{x}_{ij},$$

$$\frac{\partial \overline{\ell}_2(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \xi_u} = \sum_{i=1}^{n} \sum_{m=1}^{m_i} (y_{ij} - \overline{\mu}_{iu}), \quad u = 1, \ldots, k,$$

where $\overline{\mu}_{iu} = \exp(\xi_u + \overline{\boldsymbol{x}}'_{im}\boldsymbol{\beta})/[1 + \exp(\xi_u + \overline{\boldsymbol{x}}'_{im}\boldsymbol{\beta})]$.

Regarding the first derivative of the third component, we have

$$\frac{\partial \overline{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi})}{\partial \eta} = \sum_{i=1}^{n} \left[\sum_{u=1}^{k} \overline{z}_{imu} \delta_i (1 + \nu \log t_i) - \sum_{m=2}^{m_i} \sum_{u=1}^{k} \overline{z}_{i,m-1,u} \frac{\partial H_i(\bar{t}_{m-1}, \bar{t}_m, u)}{\partial \eta} - \sum_{u=1}^{k} \overline{z}_{im_i u} \frac{\partial H_i(t_{m_i}, t_i, u)}{\partial \eta} \right]$$

and

$$\frac{\partial \overline{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi})}{\partial \xi_u} = \phi \sum_{i=1}^{n} \left[\delta_i \phi - \sum_{m=2}^{m_i} \sum_{u=1}^{k} \overline{z}_{i,m-1,u} H_i(\bar{t}_{m-1}, \bar{t}_m, u) - \sum_{u=1}^{k} \overline{z}_{im_i u} H_{im_i}(t_{m_i}, t_i, u) \right], \quad u = 1, \ldots, k.$$

Finally, we have

$$\frac{\partial \overline{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi})}{\partial \phi} = \sum_{i=1}^{n} \left[\delta_i \sum_{u=1}^{k} \overline{z}_{i,m_i,u} \xi_u - \sum_{m=2}^{m_i} \sum_{u=1}^{k} \overline{z}_{i,m-1,u} H_i(\bar{t}_{m-1}, \bar{t}_m, u) \xi_u - \sum_{u=1}^{k} \overline{z}_{im_i u} H_{im_i}(\bar{t}_{m_i}, t_i, u) \xi_u \right]$$

and

$$\frac{\partial \overline{\ell}_3(\nu, \boldsymbol{\xi}, \phi, \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \sum_{i=1}^{n} \left[\delta_i - \sum_{m=2}^{m_i} \sum_{u=1}^{k} \overline{z}_{i,m-1,u} H_i(\bar{t}_{m-1}, \bar{t}_m, u) - \sum_{u=1}^{k} \overline{z}_{im_i u} H_i(\bar{t}_{m_i}, t_i, u) \right] \boldsymbol{w}_i.$$