Received: 17 January 2022 | Revised: 4 January 2023 | Accepted: 10 January 2023

DOI: 10.1002/bimj.202200016

Biometrical Journal

RESEARCH ARTICLE

A hidden Markov model for continuous longitudinal data with missing responses and dropout

Silvia Pandolfi[1] | **Francesco Bartolucci**[1] | **Fulvia Pennoni**[2]

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

[2]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy

Correspondence
Silvia Pandolfi, Department of Economics, University of Perugia, Via A. Pascoli, 20 - 06123 Perugia, Italy.
Email: silvia.pandolfi@unipg.it

Funding information
Università degli Studi di Perugia, Grant/Award Number: RICBASE PROGETTURALE 2017-2019 PANDOLFI; Università degli Studi di Milano-Bicocca, Grant/Award Number: 2020-ATE-0615

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to the computational complexity of the results.

Abstract

We propose a hidden Markov model for multivariate continuous longitudinal responses with covariates that accounts for three different types of missing pattern: (I) partially missing outcomes at a given time occasion, (II) completely missing outcomes at a given time occasion (intermittent pattern), and (III) dropout before the end of the period of observation (monotone pattern). The missing-at-random (MAR) assumption is formulated to deal with the first two types of missingness, while to account for the informative dropout, we rely on an extra absorbing state. Estimation of the model parameters is based on the maximum likelihood method that is implemented by an expectation-maximization (EM) algorithm relying on suitable recursions. The proposal is illustrated by a Monte Carlo simulation study and an application based on historical data on primary biliary cholangitis.

KEYWORDS

expectation-maximization algorithm, forward–backward recursion, latent Markov model, missing values, prediction

1 | INTRODUCTION

It is well known that longitudinal data are typically affected by missing responses that may arise in different ways (Diggle et al., 2002; Fitzmaurice et al., 2004; Little & Rubin, 2020). The missing pattern is defined as monotone when individuals may drop out of the sample before the end of the study. In the medical field, this may be due to a terminal event, such as the death of a patient. Another type of missingness, also referred to as intermittent, is when individuals are still in the sample but, for any reason, they do not provide responses at one or more time occasions, as when a clinical visit is missed. The problem is particularly severe when the missing data mechanism is informative or not ignorable, that is, when the missingness depends on the unobserved responses (Albert, 2000; Albert et al., 2002; Diggle & Kenward, 1994; Little, 1995; Little & Rubin, 2020; Rubin, 1976; Rizopoulos, 2012; Yuan & Little, 2009). Additional reviews that are relevant for the present approach may be found in Verbeke et al. (2014) and Zhou et al. (2020).

In this paper, we propose a hidden Markov (HM) model (Bartolucci et al., 2013; Zucchini et al., 2016) for multivariate continuous longitudinal outcomes that can address both monotone and intermittent missing data patterns. The HM approach is of particular interest when dealing with longitudinal data (Bartolucci et al., 2014) as it models time dependence in a flexible way and allows us to perform a dynamic model-based clustering (Bouveyron et al., 2019). Within this approach, the same individual is allowed to move between clusters across time, and these dynamics are provided in terms of trajectories. This is because a sequence of discrete latent variables, rather than a single latent variable, is associated to every individual, giving rise to a hidden process assumed to follow a Markov chain, generally of first order. The states of this chain correspond to latent clusters or subpopulations of homogeneous individuals sharing the same latent characteristics. This framework may be formulated to deal with responses of different nature. In the proposed approach, we consider continuous outcomes so that, given the latent state, the response variables are assumed to follow a multivariate Gaussian distribution with state-specific parameters. On the other hand, with categorical responses, the conditional distribution of the response variables is freely parameterized on the basis of conditional response probabilities.

The proposed model may also include individual covariates that affect the distribution of the latent states and, in particular, the initial and the transition probabilities of the Markov chain (Bartolucci et al., 2014), so that it is possible to understand the effect of these covariates on the evolution of the latent states representing different levels of the latent trait of interest. For a general comparison between discrete latent variable approaches, as the present one, and corresponding continuous latent variable models, see Bartolucci et al. (2022).

In detail, the proposed HM model takes explicitly into account the following patterns of missing data: (I) partially missing outcomes at a given time occasion, (II) completely missing outcomes at one occasion without dropout from the sample of the individual, (III) dropout from the sample. The first two cases correspond to the intermittent missing responses defined above, whereas the third corresponds to a monotone missing pattern also defined as attrition. In particular, cases (I) and (II) are dealt with under the missing-at-random (MAR) assumption, according to which the missing pattern is independent of the missing responses given the observed data. In case (III), dropout is not ignorable, and specifying a model for the missing data mechanism is in order. Along with the proposal of Montanari and Pandolfi (2018), who applied a similar model in the context of longitudinal categorical response variables, we include a hidden (or latent) state, which is absorbing, in addition to the other states representing unobserved heterogeneous populations that may arise in the study; see also Spagnoli et al. (2011).

We rely on the maximum likelihood (ML) approach to estimate the proposed model, paying particular attention to the computational aspects. In more detail, we propose an extended expectation-maximization (EM; Baum et al., 1970; Dempster et al., 1977; Welch, 2003) algorithm based on suitable recursions and that relies on certain techniques used to estimate a finite mixture (FM) of Gaussian distributions (McLachlan & Peel, 2000) with MAR responses (Delalleau et al., 2012; Eirola et al., 2014). In fact, the proposed HM approach may be seen as an extension of the FM approach for cross-sectional data under the MAR assumption. The estimation algorithm is also employed when individual covariates are available. In this way, it is possible to understand the influence of these covariates on the dynamic allocation of the individuals between states over time.

The proposed approach directly compares with other approaches available in the literature that rely on an HM model for longitudinal data with missing responses; we refer, in particular, to Spagnoli et al. (2011), Maruotti (2015), Marino and Alfó (2015), Bartolucci and Farcomeni (2015), Marino et al. (2018), Montanari and Pandolfi (2018), and Bartolucci and Farcomeni (2019). The added value of the present proposal is that of unifying in the same model three different types of missing pattern, indicated by (I), (II), and (III) above. In particular, with respect to the shared-parameter approach illustrated in Bartolucci and Farcomeni (2015, 2019), we account for dropout relying on the absorbing state, without the need to specify a survival model component.

In order to illustrate our proposal, we rely on a series of simulations that allow us to assess the ML estimator in terms of finite sample properties. We also show an application based on historical data about primary biliary cholangitis (or cirrhosis; PBC) collected by the Mayo Clinic from January 1974 to May 1984 (Murtaugh et al., 1994). Data are referred to several biochemical measurements of the liver function scheduled for the patients according to the clinical protocol. Continuous and binary covariates related to the patients are also available such as age, gender, and medication use. The data are very sparse due to missing visits and the fact that some biochemical measurements were not collected at each visit. Moreover, several deaths were registered and this is considered a dropout indicator. In this applied context, as noted in Royston et al. (2006), it is important to dispose of risk groups of patients derived from the model, especially to make clinical decisions about therapy.

The codes used to perform ML estimation of the proposed HM model with missing values, both in its basic version and with the inclusion of individual covariates, are implemented by extending the functions of the `R` package `LMest` (Bartolucci et al., 2017) and are available as Supporting Information.

The remainder of the paper is organized as follows. In Section 2, we recall some preliminary statistical methods, and in Section 3, we illustrate the proposed HM formulation. In Section 4, we outline the inferential approach proposed for estimating model parameters together with aspects related to the computation of the standard errors, identifiability, model selection, and decoding. In Section 5, we present the simulation study. In Section 6, we illustrate the applicative example, whereas in Section 7, we provide some conclusions. The Appendix provides additional information on the simulation results, data, and results of the application.

2 | PRELIMINARIES

As mentioned in the previous section, the proposed HM model is an extension of the FM model of multivariate Gaussian distributions for cross-sectional data under the MAR assumption. Consequently, the latter model is briefly reviewed in this section, also in terms of estimation via the EM algorithm, in order to simplify the illustration of the proposed approach.

2.1 | Model formulation

We consider individual vectors $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{ir})'$ of r continuous response variables, with $i = 1, \dots, n$, referred to the same measurement occasions. It is well known that the FM model of Gaussian distributions assumes that there exist k components identified by the discrete latent variable U_i such that

$$\boldsymbol{Y}_i | U_i = u \sim N(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u), \qquad u = 1, \dots, k.$$

In the previous expression, $\boldsymbol{\mu}_u$ and $\boldsymbol{\Sigma}_u$ denote the mean vector and the variance–covariance matrix for the same mixture component u. Note that the variance–covariance matrices are typically assumed to be equal each other, that is, $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$. This assumption of homoscedasticity makes the model more parsimonious and avoids certain estimation problems because otherwise the likelihood is unbounded (McLachlan & Peel, 2000, chapter 3.8). More articulated constraints on the $\boldsymbol{\Sigma}_u$ matrices may be expressed, as proposed in Banfield and Raftery (1993), on the basis of suitable matrix decompositions.

With missing data, it is convenient to partition each response vector as $(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)'$, where $\boldsymbol{Y}_i^o$ is the (sub)vector of observed variables and $\boldsymbol{Y}_i^m$ is that of missing variables. Using a straightforward notation, the conditional mean vectors and variance–covariance matrix may be decomposed as

$$\boldsymbol{\mu}_u = \begin{pmatrix} \boldsymbol{\mu}_u^o \\ \boldsymbol{\mu}_u^m \end{pmatrix}, \quad \boldsymbol{\Sigma}_u = \begin{pmatrix} \boldsymbol{\Sigma}_u^{oo} & \boldsymbol{\Sigma}_u^{om} \\ \boldsymbol{\Sigma}_u^{mo} & \boldsymbol{\Sigma}_u^{mm} \end{pmatrix}, \tag{1}$$

where, for instance, $\boldsymbol{\Sigma}_u^{om}$ is the block of $\boldsymbol{\Sigma}_u$ containing the covariances between each observed and missing response. In this way, for the observed responses, we have that

$$\boldsymbol{Y}_i^o | U_i = u \sim N(\boldsymbol{\mu}_u^o, \boldsymbol{\Sigma}_u^{oo}), \quad u = 1, \dots, k.$$

Without individual covariates, each latent variable U_i has the same distribution based on the component weights $\pi_u = p(U_i = u)$, $u = 1, \dots, k$. Consequently, the manifest distribution of the observed responses is given by

$$f(\boldsymbol{y}_i^o) = \sum_{u=1}^{k} \pi_u \phi(\boldsymbol{y}_i^o; \boldsymbol{\mu}_u^o, \boldsymbol{\Sigma}_u^{oo}), \tag{2}$$

where $\boldsymbol{y}_i^o$ denotes a realization of $\boldsymbol{Y}_i^o$ and $\phi(\cdot;\cdot)$ denotes the multivariate Gaussian probability density function.

Individual covariates may be included in the model in different ways. In particular, we consider the case of component weights affected by these covariates. Let $\pi_{iu} = p(U_i = u | \boldsymbol{x}_i)$ denote the class weight for component u, which is now specific of individual i, and $\boldsymbol{x}_i$ denote the vector of individual covariates, with $i = 1, \dots, n$. We assume a multinomial logit model of type

$$\log \frac{\pi_{iu}}{\pi_{i1}} = \boldsymbol{x}_i' \boldsymbol{\beta}_u, \quad u = 2, \dots, k, \tag{3}$$

where $\boldsymbol{\beta}_u$ is a vector of regression parameters. Expression (2) for the manifest distribution is obviously extended as

$$f(\boldsymbol{y}_i^o) = \sum_{u=1}^{k} \pi_{iu} \phi(\boldsymbol{y}_i^o; \boldsymbol{\mu}_u^o, \boldsymbol{\Sigma}_u^{oo}).$$

2.2 | EM algorithm under the MAR assumption

In the presence of missing data, the observed log-likelihood can be written as

$$\ell(\theta) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i^o) = \sum_{i=1}^{n} \log \left[\sum_{u=1}^{k} \pi_u \phi(\boldsymbol{y}_i^o; \boldsymbol{\mu}_u^o, \boldsymbol{\Sigma}_u^{oo}) \right],$$

where θ is the vector of all model parameters. Its maximization for parameter estimation relies on the EM algorithm (Dempster et al., 1977). This algorithm is based on the *complete-data log-likelihood* that has expression

$$\ell^*(\theta) = \sum_{i=1}^{n} \sum_{u=1}^{k} z_{iu} \log[\pi_u \phi(\boldsymbol{y}_i; \boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u)],$$

where z_{iu} is a binary variable indicating whether or not unit i comes from component u.

As usual, the EM algorithm alternates two steps until convergence. The *E-step* consists in computing the conditional expectation of the indicator variables given the observed data and the current value of the parameters, that is,

$$\widehat{z}_{iu} = \frac{\pi_u \phi(\boldsymbol{y}_i^o; \boldsymbol{\mu}_u^o, \boldsymbol{\Sigma}_u^{oo})}{\sum_{v=1}^{k} \pi_v \phi(\boldsymbol{y}_i^o; \boldsymbol{\mu}_v^o, \boldsymbol{\Sigma}_v^{oo})}. \tag{4}$$

With individual covariates, this rule is modified by substituting every π_u with π_{iu}.

Following the proposals in Eirola et al. (2014) and Delalleau et al. (2012), in order to account for missing data, this step also includes the computation of the following conditional expectations:

$$\mathrm{E}(\boldsymbol{Y}_i | \boldsymbol{y}_i^o, u) = \begin{pmatrix} \boldsymbol{y}_i^o \\ \mathrm{E}(\boldsymbol{Y}_i^m | \boldsymbol{y}_i^o, u) \end{pmatrix}, \quad i = 1, \ldots, n, \tag{5}$$

where

$$\mathrm{E}(\boldsymbol{Y}_i^m | \boldsymbol{y}_i^o, u) = \boldsymbol{\mu}_u^m + \boldsymbol{\Sigma}_u^{mo} (\boldsymbol{\Sigma}_u^{oo})^{-1} (\boldsymbol{y}_i^o - \boldsymbol{\mu}_u^o),$$

and of the conditional variances

$$\mathrm{Var}(\boldsymbol{Y}_i | \boldsymbol{y}_i^o, u) = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}_u^{mm} - \boldsymbol{\Sigma}_u^{mo} (\boldsymbol{\Sigma}_u^{oo})^{-1} \boldsymbol{\Sigma}_u^{om} \end{pmatrix},$$

where $\boldsymbol{O}$ is a matrix of zeros of suitable dimension. The above expressions originate from the fact that the missing values' conditional distribution also follows a multivariate Gaussian distribution (Anderson, 2003).

The *M-step* consists in updating the model parameters by maximizing the expected value of $\ell^*(\theta)$ obtained at the E-step. This maximization leads to the following updating rules for the mean vector and the variance–covariance matrices:

$$\boldsymbol{\mu}_u = \frac{1}{\sum_{i=1}^{n} \widehat{z}_{iu}} \sum_{i=1}^{n} \widehat{z}_{iu} \mathrm{E}(\boldsymbol{Y}_i | \boldsymbol{y}_i^o, u), \tag{6}$$

$$\boldsymbol{\Sigma}_u = \frac{1}{\sum_{i=1}^{n} \widehat{z}_{iu}} \sum_{i=1}^{n} \widehat{z}_{iu} \left\{ \mathrm{Var}(\boldsymbol{Y}_i | \boldsymbol{y}_i^o, u) + [\mathrm{E}(\boldsymbol{Y}_i | \boldsymbol{y}_i^o, u) - \boldsymbol{\mu}_u][\mathrm{E}(\boldsymbol{Y}_i | \boldsymbol{y}_i^o, u) - \boldsymbol{\mu}_u]' \right\}. \tag{7}$$

The rule in (7) is suitably modified by summing the numerator for all u, and dividing the sum by n, under the constraint of homoscedasticity. The component weights are updated as

$$\pi_u = \frac{\sum_{i=1}^{n} \widehat{z}_{iu}}{n}, \qquad u = 1, \dots, k,$$

whereas with individual covariates, we maximize the complete log-likelihood component

$$\sum_{i=1}^{n} \sum_{u=1}^{k} \widehat{z}_{iu} \log \pi_{iu}$$

with respect to the parameter vectors $\boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_k$ defined in Expression (3) by a Newton–Raphson algorithm.

These two steps are repeated until convergence, which is checked on the basis of the relative log-likelihood difference, that is,

$$\frac{| \ell(\boldsymbol{\theta}^{(s)}) - \ell(\boldsymbol{\theta}^{(s-1)}) |}{| \ell(\boldsymbol{\theta}^{(s)}) |} \leq \epsilon,$$

where $\boldsymbol{\theta}^{(s)}$ is the vector of parameter estimates obtained at the end of the s-th M-step and ϵ is a suitable tolerance level (e.g., 10^{-8}). A crucial aspect is that of the initialization of the algorithm to deal with the likelihood multimodality. In this regard, it is important to rely on different strategies, even based on random rules, for choosing the starting values of the model parameters; see also Shireman et al. (2017).

2.3 | Model selection and clustering

Selection of the number of components is an important aspect when applying FM models (McLachlan & Peel, 2000, Chapter 6). Typically, this choice is based on information criteria such as the Akaike information criterion (AIC; Akaike, 1973) or the Bayesian information criterion (BIC; Schwarz, 1978), obtained through penalizations of the maximum log-likelihood. In particular, the AIC index is expressed as

$$AIC_k = -2\widehat{\ell}_k + 2\#par_k,$$

whereas the BIC index has expression

$$BIC_k = -2\widehat{\ell}_k + \log(n)\#par_k,$$

where $\widehat{\ell}_k$ denotes the maximum of the log-likelihood of the FM model with k components and $\#par_k$ denotes the corresponding number of free parameters. To select the optimal number of components, we estimate a series of models for increasing values of k, and we select the one corresponding to the minimum value of these indexes. The BIC is usually preferred to the AIC as the latter tends to overestimate the number of components (McLachlan & Peel, 2000, chapter 6.9).

An alternative procedure is based on the cross-validated likelihood criterion proposed for FM models by Smyth (2000), in which the data are repeatedly divided into two randomly chosen partitions: the training subset and the test subset. For every partition, the model is estimated on the training subset and the corresponding log-likelihood is evaluated on the test subset. The cross-validated log-likelihood is computed as the average, over all the partitions, of the log-likelihood on the testing data. The model with the highest value of the cross-validated log-likelihood is then selected as the best one.

Finally, on the basis of the estimation results, model-based clustering is performed with the maximum a posteriori (MAP) rule, consisting in assigning unit i to the latent component corresponding to the maximum over u of the posterior probability computed as in (4). The selected component for unit i is denoted by $\widehat{u}_i$. Note that, with missing data, it is possible to perform a sort of multiple imputation that allows us to predict the missing responses conditionally or unconditionally to the model component. In the first case, the predicted value is simply $\widehat{\boldsymbol{y}}_i = \mathrm{E}(\boldsymbol{Y}_i | \boldsymbol{y}_i^o, \widehat{u}_i)$, while for the unconditional case, it is computed as

$$\widetilde{\boldsymbol{y}}_i = \sum_{u=1}^{k} \widehat{z}_{iu} \mathrm{E}(\boldsymbol{Y}_i | \boldsymbol{y}_i^o, u),$$

where the expected value is defined in (5).

3 | PROPOSED HM MODEL WITH MISSING DATA

In the following, we illustrate the assumptions of the proposed HM approach to deal with the intermittent and monotone missing data structures indicated by (I), (II), and (III) in Section 1.

3.1 | Model formulation

We denote by $\boldsymbol{Y}_{it} = (Y_{i1t}, \dots, Y_{irt})'$ the vector of r continuous response variables measured at time t for individual i, with $i = 1, \dots, n$ and $t = 1, \dots, T_i$. Note that the number of time occasions T_i is specific for each individual i. In this way, we also conceive unbalanced panels that, for example, are due to a different number of scheduled visits for every individual in a medical study. Also note that the time between visits should be at least approximatively the same across the panel of patients. Clearly, in a longitudinal setting, it is also important to account for dropout that gives rise to monotone informative missing data. In this regard, we first introduce the indicator variable D_{it} for the dropout, which assumes a value equal to 0 if unit i is still in the panel at occasion t and equal to 1 if the same unit has dropped out. It is worth noting that the event of dropout at time t, so that $D_{it} = 1$, implies that $D_{i,t+1}, \dots, D_{iT_i} = 1$, where we use the convention that a comma is inserted between two indices when necessary to avoid misunderstandings. Obviously, even if $D_{it} = 0$, we may still have a missing observation at occasion t due to the intermittent missing data pattern with all or some of the outcomes that are not observed.

The general HM formulation assumes the existence of a latent process for each individual i, denoted by $\boldsymbol{U}_i = (U_{i1}, \dots, U_{iT_i})'$, which affects the distribution of the response variables and is assumed to follow a first-order Markov chain with a certain number of states equal to $k + 1$. This model may account for dropout by adding an extra latent state, the $(k + 1)$-th, defined as an absorbing state, in the sense that once it has been reached, then it is not possible to move away from it. This proposal has been previously introduced in Montanari and Pandolfi (2018) for the HM model with discrete response variables.

The HM model has two components: the measurement (sub)model, concerning the conditional distribution of the response variables given the latent process, and the latent (sub)model, concerning the distribution of the latent process. Regarding the first component, we assume that the response vectors are conditionally independent given the hidden state and that, as in the FM model presented in Section 2, for the first k states, the response vectors have a Gaussian distribution with a specific mean vector and variance–covariance matrix. More precisely, we assume that

$$\boldsymbol{Y}_{it} | U_{it} = u, D_{it} = 0 \sim N(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}), \qquad u = 1, \dots, k, \tag{8}$$

where the vectors $\boldsymbol{\mu}_u$ of length r, with r denoting the number of response variables, are specific of each state u, $u = 1, \dots, k$, and the variance–covariance matrix $\boldsymbol{\Sigma}$, of dimension $r \times r$, is assumed to be constant across states under the assumption of homoscedasticity. This choice is based on the same arguments clarified in Section 2.1 for FM models and allows us to avoid an excessive number of parameters and possible estimation problems. However, it may be suitably relaxed on the basis of proper matrix decompositions. In addition, we assume that

$$P(D_{it} = d | U_{it} = u) = \begin{cases} 1 & \text{with } d = 0 \text{ and } u = 1, \dots, k \text{ or } d = 1 \text{ and } u = k + 1, \\ 0 & \text{otherwise.} \end{cases}$$

In order to account for intermittent missing responses, we consider the partition $\boldsymbol{Y}_{it} = (\boldsymbol{Y}^o_{it}, \boldsymbol{Y}^m_{it})'$, where $\boldsymbol{Y}^o_{it}$ is the (sub)vector of the observed responses and $\boldsymbol{Y}^m_{it}$ is referred to the missing data. As for the model illustrated in Section 2, we consider a decomposition of the conditional mean vector and the variance–covariance matrix; see Expression (1). Consequently, we have that

$$\boldsymbol{Y}^o_{it} | U_{it} = u, D_{it} = 0 \sim N(\boldsymbol{\mu}^o_u, \boldsymbol{\Sigma}^{oo}), \qquad u = 1, \dots, k. \tag{9}$$

Note that the distribution of $\boldsymbol{Y}^o_{it}$ given $U_{it} = k + 1$ and $D_{it} = 1$ does not need to be defined.

Finally, the parameters of the latent model are the initial probabilities

$$\pi_u = p(U_{i1} = u), \qquad u = 1, \dots k,$$

and the transition probabilities

$$\pi_{u|\bar{u}} = p(U_{it} = u|U_{i,t-1} = \bar{u}), \qquad t = 2, \dots, T_i, \bar{u}, u = 1, \dots, k+1,$$

where u denotes a realization of U_{it} and $\bar{u}$ a realization of $U_{i,t-1}$. Note that the transition probabilities are assumed to be time homogeneous to reduce the number of free parameters, but even this assumption may be relaxed if necessary. Moreover, given the interpretation of the latent state referred to the dropout, the transition probabilities are suitably constrained as

$$\pi_{k+1|k+1} = 1, \quad \pi_{u|k+1} = 0, \qquad u = 1, \dots, k.$$

Therefore, once a subject reaches the $(k+1)$-th latent state, he/she yields missing values until the end of the study. Note also that the initial probabilities are defined for the first k states, as no unit is in the dropout state at the beginning of the study. The interest in modeling this extra state is evident when the model includes individual covariates, as shown in the following.

In the present formulation, the manifest distribution is expressed with reference to the observed data represented by $\mathcal{Y}_i^o = \{\boldsymbol{y}_{it}^o, t = 1, \dots, T_i : d_{it} = 0\}$, which is the set of vectors $\boldsymbol{y}_{it}^o$ observed when $d_{it} = 0$, for $i = 1, \dots, n$. Denoting by $\boldsymbol{d}_i$ the vector of the observed indicator variables D_{it} for individual i, we have that

$$\begin{aligned} f(\boldsymbol{d}_i, \mathcal{Y}_i^o) &= \sum_{\boldsymbol{u}_i} f(\mathcal{Y}_i^o|\boldsymbol{D}_i = \boldsymbol{d}_i, \boldsymbol{U}_i = \boldsymbol{u}_i) P(\boldsymbol{D}_i = \boldsymbol{d}_i|\boldsymbol{U}_i = \boldsymbol{u}_i) P(\boldsymbol{U}_i = \boldsymbol{u}_i) \\ &= \sum_{\boldsymbol{u}_i} \left[\prod_{t=1}^{T_i} f(\boldsymbol{y}_{it}^o|d_{it}, u_{it})\, p(d_{it}|u_{it}) \right] \left(\pi_{u_{i1}} \prod_{t=2}^{T_i} \pi_{u_{it}|u_{i,t-1}} \right), \end{aligned} \tag{10}$$

where the density $f(\boldsymbol{y}_{it}^o|d_{it}, u_{it})$ is based on assumption (9) for $d_{it} = 0$ and is let equal to 1 otherwise. As usual in dealing with HM models, to efficiently compute this distribution, we rely on a forward recursion (Baum et al., 1970; Welch, 2003).

The above formulation allows characterizing the process of generating informative dropout and the probability of transition from the latent states to the dropout state. More in detail, starting from the EM algorithm illustrated in Section 2.2, we develop an inferential approach to obtain exact ML estimates of model parameters under the MAR assumption for the intermittent missingness and with informative dropout. The resulting EM algorithm, illustrated in Section 4, also includes suitable forward–backward recursions to perform the E-step (Baum et al., 1970; Welch, 2003).

Finally note that a common problem in the literature of HM and FM models is the label switching problem, that is, the invariance of the likelihood with respect to permutations of the first k hidden (or latent) states. The dropout state is instead not exchangeable with the other states. However, although it has been known to be a challenging issue in the Bayesian context, this is not a problem when performing ML inference, as it is possible to arbitrarily order the hidden states after the estimation process to favor their interpretability.

3.2 | Inclusion of individual covariates

Longitudinal data allow for a precise assessment of the effect of individual covariates and this aspect is particularly relevant when these variables are related to a certain treatment as in the empirical illustration provided in Section 6. In the HM formulation, individual covariates may be included in the measurement model or in the latent model; for a general review, see Bartolucci et al. (2013) and Bartolucci et al. (2014). In the first case, the conditional distribution of the response variables given the latent states must be suitably parameterized. In such a situation, the latent variables account for the unobserved heterogeneity that is allowed to be time-varying (Bartolucci & Farcomeni, 2009). When covariates are included in the latent model, the interest is in modeling the effect of covariates on the distribution of the latent process (Vermunt et al., 1999). This formulation is relevant when the response variables measure an individual characteristic of interest represented by the latent variables.

In this work, we consider the second formulation, and we adopt a multinomial parameterization for the initial and transition Markov chain probabilities. More in detail, let $\boldsymbol{x}_{it}$ denote the vector of individual covariates available at the t-th time occasion for individual i. Now the initial and transition probabilities are individual specific and denoted by $\pi_{iu} = p(U_{i1} = u|\boldsymbol{x}_{i1})$, $u = 1, \dots, k$, and $\pi_{i,u|\bar{u}} = p(U_{it} = u|U_{i,t-1} = \bar{u}, \boldsymbol{x}_{it})$, $t = 2, \dots, T$, $\bar{u}, u = 1, \dots, k+1$, respectively. For

the initial probabilities, we rely on the following parameterization:

$$\log \frac{\pi_{iu}}{\pi_{i1}} = \beta_{0u} + \boldsymbol{x}'_{i1}\boldsymbol{\beta}_{1u}, \qquad u = 2, \dots, k, \tag{11}$$

whereas for the transition probabilities, we consider the following parameterization:

$$\log \frac{\pi_{i,u|\bar{u}}}{\pi_{i,\bar{u}|\bar{u}}} = \gamma_{0\bar{u}u} + \boldsymbol{x}'_{it}\boldsymbol{\gamma}_{1\bar{u}u}, \qquad \bar{u} = 1, \dots, k, u = 1, \dots, k+1, \bar{u} \neq u. \tag{12}$$

In the above expressions, $\boldsymbol{\beta}_u = (\beta_{0u}, \boldsymbol{\beta}'_{1u})'$ and $\boldsymbol{\gamma}_{\bar{u}u} = (\gamma_{0\bar{u}u}, \boldsymbol{\gamma}'_{1\bar{u}u})'$ are parameter vectors to be estimated that are collected in the matrices $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$, respectively. Parameters in $\boldsymbol{B}$ are not affected by the presence of the extra state since no unit is in the dropout state at the beginning of the study. On the other hand, parameters in $\boldsymbol{\Gamma}$ are properly constrained to avoid transitions from the latent absorbing state.

Finally, Expression (10) for the manifest distribution is extended as

$$f(\boldsymbol{d}_i, \mathcal{Y}_i^o) = \sum_{\boldsymbol{u}_i} \left[\prod_{t=1}^{T_i} f(\boldsymbol{y}_{it}^o | d_{it}, u_{it})\, p(d_{it}|u_{it}) \right] \left(\pi_{i,u_{i1}} \prod_{t=2}^{T_i} \pi_{i,u_{it}|u_{i,t-1}} \right). \tag{13}$$

4 | MODEL INFERENCE

In the following, we first illustrate the proposed inferential approach based on the maximization of the log-likelihood function. Then, we outline the strategy for the initialization of the estimation algorithm. Finally, we discuss issues related to the computation of the standard errors, model identifiability, selection of the number of states, and model-based dynamic clustering.

4.1 | Maximum log-likelihood estimation with missing responses

Assuming independence between sample units, the log-likelihood referred to the observed data is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{d}_i, \mathcal{Y}_i^o).$$

In the above expression, $\boldsymbol{\theta}$ is the vector of all model parameters and $f(\boldsymbol{d}_i, \mathcal{Y}_i^o)$ is the manifest distribution of the observed responses defined in (10) without covariates and in (13) with covariates. In order to estimate the parameters, we maximize $\ell(\boldsymbol{\theta})$ by an EM algorithm based on a *complete-data log-likelihood* that may be expressed as the sum of three components that are maximized separately

$$\ell^*(\boldsymbol{\theta}) = \ell_1^*(\boldsymbol{\theta}) + \ell_2^*(\boldsymbol{\theta}) + \ell_3^*(\boldsymbol{\theta}),$$

where

$$\ell_1^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{\substack{t=1 \\ (d_{it}=0)}}^{T_i} \sum_{u=1}^{k} z_{itu} \log f(\boldsymbol{y}_{it} | D_{it} = 0, u),$$

$$\ell_2^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{u=1}^{k} z_{i1u} \log \pi_u,$$

$$\ell_3^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{t=2}^{T_i} \sum_{\bar{u}=1}^{k+1} \sum_{u=1}^{k+1} z_{it\bar{u}u} \log \pi_{u|\bar{u}}.$$

In the above expressions, $f(\boldsymbol{y}_{it}|D_{it}=0,u)$ is based on assumption (8), $z_{itu}=I(u_{it}=u)$ is an indicator variable equal to 1 if individual i is in latent state u at time t, and $z_{it\bar{u}u}=z_{i,t-1,\bar{u}}\,z_{itu}$ is the indicator variable for the transition from state $\bar{u}$ to state u of individual i at time occasion t. When individual covariates are available, expressions for $\ell_2^*(\boldsymbol{\theta})$ and $\ell_3^*(\boldsymbol{\theta})$ are modified by substituting every π_u and $\pi_{u|\bar{u}}$ with π_{iu} and $\pi_{i,u|\bar{u}}$, respectively, which in turn are formulated as in Expressions (11) and (12). Also note that in $\ell_1^*(\boldsymbol{\theta})$, the sum over t is computed only when $d_{it}=0$, whereas in $\ell_3^*(\boldsymbol{\theta})$ the last two sums also involve the absorbing hidden state, $k+1$. More explicitly, the first component of the complete log-likelihood function may be written as

$$\log f(\boldsymbol{y}_{it}|D_{it}=0,u)=-\frac{1}{2}\log(|\boldsymbol{\Sigma}|)-\frac{1}{2}(\boldsymbol{y}_{it}-\boldsymbol{\mu}_u)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_{it}-\boldsymbol{\mu}_u).$$

Therefore, we have that

$$\ell_1^*(\boldsymbol{\theta})=\sum_{i=1}^{n}\left\{-\frac{T_i-d_{i+}}{2}\log(|\boldsymbol{\Sigma}|)-\frac{1}{2}\sum_{\substack{t=1\\(d_{it}=0)}}^{T_i}\sum_{u=1}^{k}z_{itu}\,\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_{it}-\boldsymbol{\mu}_u)(\boldsymbol{y}_{it}-\boldsymbol{\mu}_u)'\right]\right\},$$

where $d_{i+}=\sum_{t=1}^{T_i}d_{it}$.

At the *E-step* of the EM algorithm we compute the posterior expected value of the indicator variables given the observed data and the current value of the parameters. In particular, these expected values correspond to the following quantities:

$$\widehat{z}_{itu}=p(U_{it}=u|\boldsymbol{d}_i,\mathcal{Y}_i^o),\quad t=1,\ldots,T_i,u=1,\ldots,k+1, \tag{14}$$

$$\widehat{z}_{it\bar{u}u}=p(U_{it}=u,U_{i,t-1}=\bar{u}|\boldsymbol{d}_i,\mathcal{Y}_i^o),\quad t=2,\ldots,T_i,\bar{u},u=1,\ldots,k+1, \tag{15}$$

computed by means of forward–backward recursions of Baum et al. (1970) and Welch (2003); for an illustration, see Bartolucci et al. (2013). We stress that when $d_{it}=1$, that is, when unit i has dropped out at occasion t, we have $\widehat{z}_{isu}=0$, for $u=1,\ldots,k$ and $s=t,\ldots,T_i$, and $\widehat{z}_{is,k+1}=1$ for $s=t,\ldots,T_i$. With individual covariates, the posterior probabilities in (14) and (15) also take into account these covariates that affect the initial and transition probabilities. Furthermore, when $d_{it}=0$, the E-step also includes the computation of the following expected values resulting from the MAR assumption for the missing observations:

$$\mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o,u)=\begin{pmatrix}\boldsymbol{y}_{it}^o\\ \boldsymbol{\mu}_u^m+\boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}(\boldsymbol{y}_{it}^o-\boldsymbol{\mu}_u^o)\end{pmatrix}, \tag{16}$$

$$\begin{aligned}\mathrm{E}\left[(\boldsymbol{Y}_{it}-\boldsymbol{\mu}_u)(\boldsymbol{Y}_{it}-\boldsymbol{\mu}_u)'|\boldsymbol{y}_{it}^o,u\right]&=\mathrm{Var}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o)+\mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o,u)\mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o,u)'\\&\quad-\boldsymbol{\mu}_u\mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o,u)'-\mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o,u)\boldsymbol{\mu}_u'+\boldsymbol{\mu}_u\boldsymbol{\mu}_u'\\&=\mathrm{Var}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o)+[E(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o,u)-\boldsymbol{\mu}_u][E(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o,u)-\boldsymbol{\mu}_u]',\end{aligned}$$

where

$$\mathrm{Var}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o)=\begin{pmatrix}\boldsymbol{O}&\boldsymbol{O}\\ \boldsymbol{O}&\boldsymbol{\Sigma}^{mm}-\boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}\boldsymbol{\Sigma}^{om}\end{pmatrix}.$$

At the *M-step* of the EM algorithm, we update the model parameters by considering the closed-form solution for the means

$$\boldsymbol{\mu}_u=\frac{1}{\sum_{i=1}^{n}\sum_{\substack{t=1\\(d_{it}=0)}}^{T_i}\widehat{z}_{itu}}\sum_{i=1}^{n}\sum_{\substack{t=1\\(d_{it}=0)}}^{T_i}\widehat{z}_{itu}\mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o,u),\quad u=1,\ldots,k,$$

and we update $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \frac{1}{\sum_{i=1}^{n}(T_i - d_{i+})} \sum_{i=1}^{n} \sum_{\substack{t=1 \\ (d_{it}=0)}}^{T_i} \sum_{u=1}^{k} \widehat{z}_{itu} \left\{ \text{Var}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o) + [\text{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u) - \boldsymbol{\mu}_u][\text{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u) - \boldsymbol{\mu}_u]' \right\}$$

that directly compare with Expressions (6) and (7).

Finally, without individual covariates, the initial and transition probabilities may be updated as

$$\pi_u = \frac{\sum_{i=1}^{n} \widehat{z}_{i1u}}{n}, \qquad u = 1, \ldots, k,$$

$$\pi_{u|\bar{u}} = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \widehat{z}_{it\bar{u}u}}{\sum_{i=1}^{n} \sum_{t=2}^{T_i} \widehat{z}_{i,t-1,u}}, \qquad \bar{u}, u = 1, \ldots, k+1,$$

whereas, with individual covariates, in order to update the latent model parameters, we maximize the complete log-likelihood components $\ell_2^*(\boldsymbol{\theta})$ and $\ell_3^*(\boldsymbol{\theta})$, with respect to $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$, by a Newton–Raphson algorithm.

4.2 | Algorithm initialization

As already mentioned for FM models, the initialization of the EM algorithm plays a central role as the model log-likelihood is typically multimodal. This is a common problem in the estimation of discrete latent variable models implying that the EM algorithm may converge to one of the local modes that do not correspond to the global maximum. In such a situation, a multistart strategy, based both on a deterministic and a random starting rule, is necessary. More in detail, the deterministic rule consists in computing the starting values of the parameters of the measurement model, $\boldsymbol{\mu}_u$ and $\boldsymbol{\Sigma}$, on the basis of descriptive statistics (mean and covariance matrix) of the observed outcomes. In particular, the starting values for every $\boldsymbol{\mu}_u$, $u = 1, \ldots, k$, are diversified on the basis of the product between the observed standard deviation of the response variables and suitable quantiles of the standard Gaussian distribution that are added to the mean of the observed outcomes. The starting values for the initial probabilities π_u are chosen as $1/k$, for $u = 1, \ldots, k$, also including the constraint $\pi_{k+1} = 0$. For the transition probabilities, we use $\pi_{u|\bar{u}} = (h+1)/[h + (k+1)]$ when $u = \bar{u}$ and $\pi_{u|\bar{u}} = 1/[h + (k+1)]$ when $u \neq \bar{u}$, for $u = 1, \ldots, k+1$ and $\bar{u} = 1, \ldots, k$, where h is a suitable constant; for instance, in the application illustrated in Section 6, we use $h = 9$. We also constrain the last row of the transition matrix to have elements $\pi_{u|k+1} = 0$ for $u = 1, \ldots, k$ and $\pi_{k+1|k+1} = 1$. The random starting rule is based on values generated from a Gaussian distribution for the vectors $\boldsymbol{\mu}_u$, $u = 1, \ldots, k$, and on suitable normalized random numbers drawn from a uniform distribution between 0 and 1 for both initial and transition probabilities. The starting values for the variance–covariance matrix are again chosen according to the corresponding matrix computed for the observed outcomes. The same rules may be suitably adapted when individual covariates are included in the model. In this case, the initialization of the EM algorithm directly refers to the parameters in $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$.

Overall, for a given k, the inference is based on the solution corresponding to the largest value of the log-likelihood at convergence, which typically corresponds to the global maximum. The vector of parameter estimates obtained in this way is denoted by $\widehat{\boldsymbol{\theta}}$. For an illustration of alternative initialization strategies of HM models, we refer the reader to Maruotti and Punzo (2021).

4.3 | Standard errors, identifiability, model selection, and clustering

Once parameter estimates are computed for a given number of latent states k, and collected in $\widehat{\boldsymbol{\theta}}$, the corresponding standard errors may be obtained on the basis of different methods. In this work, considering the ease of implementation and robustness of the corresponding results, we mainly rely on a nonparametric bootstrap procedure (Davison & Hinkley, 1997). This is performed by repeatedly sampling with replacement data from the original sample, and fitting the proposed HM model with the selected number of states on these bootstrap samples. A drawback of this method is the high computational cost due to the need to fit the model for each resampled data set. A possible approach to reduce the computational burden is to use, as starting values for the EM algorithm, the parameter estimates obtained for the observed data.

Alternative methods to obtain standard errors may be based on a parametric bootstrap procedure, which relies on samples simulated from the estimated model, or on the computation of the square root of the diagonal elements of the inverse of the information matrix. In this regard, it is well known that the EM algorithm does not provide the observed information matrix directly. However, among the available approaches to obtain this matrix, one of the simplest is based on the numerical method proposed in Bartolucci and Farcomeni (2009), that is, as minus the numerical derivative of the score vector at convergence. The score vector, in turn, is obtained as the first derivative of the expected value of the complete data log-likelihood, as suggested by Oakes (1999).

The observed information matrix at the ML estimate may also be used to check identifiability of the model, which represents a fundamental issue when applying a statistical model. Given the complexity of the problem, no general rules are available in the literature of HM models. One possibility for assessing if the model is locally identifiable at $\widehat{\boldsymbol{\theta}}$ consists in checking if the observed information matrix is of full rank; see, among others, Goodman (1974). An alternative empirical approach is based on comparing the different solutions at convergence of the estimation algorithm starting from different points of the parameter space (Bartolucci et al., 2013). If there exist at least two different estimates leading to a value of the log-likelihood close enough to the maximum, we can assess that the model is not globally identifiable. Otherwise, provided that the number of initializations is large enough, we can be confident about the global identifiability of the model, apart from the invariance issue mentioned at the end of Section 3.1.

Concerning model selection, we rely on information criteria that are also common in the FM literature (McLachlan & Peel, 2000) and, in particular, on the BIC, which outperforms alternative information criteria as confirmed by Bacci et al. (2014). Based on this selection approach, we estimate a series of models for increasing values of k, and we select the number of latent states corresponding to the minimum value of the BIC index. However, as typically happens in applications to complex and high-dimensional data, this index may continue to decrease for each state added until a very large value of k. In such a situation, it is advisable to choose the value of k that represents a good compromise between goodness-of-fit and interpretability of the resulting latent states as implemented, among others, in Montanari and Pandolfi (2018). It is also useful to visually display the values of the index against increasing k so as to look for the "elbow," that is, a change of slope in the curve suggesting the optimal number of states (Nylund-Gibson & Choi, 2018). A similar approach has also been proposed by Pohle et al. (2017) in an application to ecological time-series data about animal movement, where a large number of states may not be easily interpretable. In such a context, the authors suggested a pragmatic solution that takes into account the standard information criteria as guidance, which are combined with an expert knowledge and model checking procedures to select the most suitable number of states for the application at hand, also taking its aim into account.

Finally, consider that also a cross-validated likelihood approach has been proposed by Celeux and Durand (2008) to select the number of states of an HM model. However, this approach turns out to be computationally demanding and we prefer to rely on the classical information criteria, whose behavior will be investigated in the section referred to the simulation study.

Once the number of states is selected, dynamic clustering is performed by assigning every unit to a latent state at each time occasion. The EM algorithm directly provides the estimated posterior probabilities of U_{it}, as defined in (14). These probabilities can be directly used to perform *local decoding* so as to predict the latent states of every unit i at each time occasion t. To obtain the prediction of the latent trajectories of a unit across time, that is, the a posteriori most likely sequence of hidden states, we also employ the so-called *global decoding*, which is based on an adaptation of the Viterbi algorithm (Viterbi, 1967); see also Juang and Rabiner (1991).

Even in this case, it is possible to perform imputation of the missing responses conditionally or unconditionally to the predicted latent state. As for FM models, in the conditional case, the predicted value is simply $\widehat{\boldsymbol{y}}_{it} = \mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, \widehat{u}_{it})$, where $\widehat{u}_{it}$ are the predicted states. The unconditional prediction of the missing responses is instead computed as

$$\widetilde{\boldsymbol{y}}_{it} = \sum_{u=1}^{k} \widehat{z}_{itu} \mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u),$$

where the expected value is defined as in (16).

5 | SIMULATION STUDY

In the following, we illustrate the simulation design carried out to assess the performance of the proposed approach. It is developed varying the sample size, the number of hidden states, and the assumed proportion of intermittent missing

responses and informative dropout observations. Here, we aim at evaluating whether the proposed inferential approach allows us to identify the correct data-generating process in terms of model parameters' recovery. We also assess the performance of the suggested information criteria in selecting the true number of hidden states.

5.1 | Simulation design

We randomly drew $B = 250$ samples of $n = 500, 1000$ individuals from an HM model with a number of hidden states equal to $k = 2, 3$. We also considered the same number of time occasions for each individual by setting $T_i = T = 5$, corresponding to balanced panel data. We finally considered $r = 3, 6$ continuous response variables, and a varying proportion of intermittent missing responses, that is, $p_{\text{miss}} = 0.01, 0.05, 0.10, 0.25$.

Regarding the measurement model, the following values for the conditional means are considered: $\boldsymbol{\mu}_1 = (-2, -2, 0)'$ and $\boldsymbol{\mu}_2 = (0, 2, 2)'$, with $k = 2$ states and $r = 3$ response variables; $\boldsymbol{\mu}_1 = (-2, -2, 0, -1, -1, 0)'$ and $\boldsymbol{\mu}_2 = (0, 2, 2, 0, 1, 1)'$, with $k = 2$ and $r = 6$. Moreover, with $k = 3$ latent states and $r = 3$ outcomes, we set $\boldsymbol{\mu}_1 = (-2, -2, 0)'$, $\boldsymbol{\mu}_2 = (0, 0, 0)'$, and $\boldsymbol{\mu}_3 = (0, 2, 2)'$, and with $r = 6$ variables, we set $\boldsymbol{\mu}_1 = (-2, -2, 0, -1, -1, 0)'$, $\boldsymbol{\mu}_2 = (0, 0, 0, 0, 0, 0)'$, and $\boldsymbol{\mu}_3 = (0, 2, 2, 0, 1, 1)'$. We also assumed a variance–covariance matrix constant across states, with all variances equal to 1 and covariances equal to 0.5. We considered equally likely hidden states at the first time period $\pi_u = 1/k$, with $u = 1, \dots, k$. Finally, we assumed a varying proportion of informative dropout, which was simulated by considering a transition matrix with increasing probabilities of moving toward the additional absorbing latent state, $k + 1$, corresponding to the dropout, that is, $p_{\text{drop}} = 0.01, 0.05, 0.10, 0.25$. This transition matrix is assumed to be time-homogeneous. Overall, we considered a total of 32 different scenarios. In the following, we report the estimation results obtained under the most sensible scenarios in order to assess the effect of the different design factors on the accuracy of the parameter estimates.

5.2 | Results

The simulation results are assessed in terms of bias and root mean square error (rmse) of the parameter estimates. In particular, given the true value of the vector of all free model parameters $\boldsymbol{\theta}_0$, the empirical bias is computed as

$$\text{bias} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}_0,$$

whereas the rmse is obtained as

$$\text{rmse} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\widehat{\boldsymbol{\theta}}_b - \boldsymbol{\theta}_0)^2},$$

where $\widehat{\boldsymbol{\theta}}_b$ is the vector of parameter estimates obtained under the simulated sample b, with $b = 1, \dots, B$, and the power and square root are applied element-wise. The estimation algorithm was initialized by means of a multistart strategy based on combining a deterministic initialization with 9 random starting values, as defined in Section 4.2.

The results reported in the following are based on a series of scenarios:

(1) The first scenario, which may be seen as a benchmark, is based on $n = 500$ individuals, $k = 2$ latent states, $r = 3$ response variables, and a low proportion of intermittent missing values and dropout, that is, $p_{\text{miss}} = p_{\text{drop}} = 0.01$.
(2) The second scenario, which is aimed at evaluating the performance of the proposed approach when the sample size increases, is based on $n = 1000$ individuals, whereas the other parameters of the simulation study are left unchanged with respect to the first scenario.
(3) The third scenario is aimed at evaluating the effect of an increase in the number of hidden states, by assuming $k = 3$ states, while letting $n = 500$, $r = 3$, and $p_{\text{miss}} = p_{\text{drop}} = 0.01$, as in the benchmark.
(4) The fourth scenario differs from the benchmark with respect to the number of response variables, $r = 6$.
(5) The fifth scenario allows us to assess the effect on the accuracy of the parameter estimates of an increased proportion of missing values, that is, when $p_{\text{miss}} = p_{\text{drop}} = 0.25$. Even in this case, the remaining parameters are left unchanged with respect to the benchmark.

TABLE 1 Average, over the latent states and response variables, of the bias (in absolute value) and rmse of the conditional mean vectors $\boldsymbol{\mu}_u$, $u = 1, \dots, k$; average, over pairs of response variables, of the bias (in absolute value) and rmse of the variance–covariance matrix $\boldsymbol{\Sigma}$; average, over the hidden states and pairs of hidden states, of the bias (in absolute value) and rmse of the initial and transition probabilities, π_u, $u = 1, \dots, k$, and $\pi_{u|\bar{u}}$, $u, \bar{u} = 1, \dots, k+1$, respectively. The benchmark scenario is based on $n = 500$ individuals, $r = 3$ response variables, $k = 2$ states, and $p_{\text{miss}} = p_{\text{drop}} = 0.01$.

		Scenario 1 Benchmark	Scenario 2 $n = 1000$	Scenario 3 $k = 3$	Scenario 4 $r = 6$	Scenario 5 $p_{\text{miss}} = p_{\text{drop}} = 0.25$
$\boldsymbol{\mu}_u$	|bias|	0.0013	0.0008	0.0016	0.0007	0.0023
	rmse	0.0297	0.0206	0.0371	0.0285	0.0433
$\boldsymbol{\Sigma}$	|bias|	0.0013	0.0006	0.0012	0.0011	0.0011
	rmse	0.0250	0.0176	0.0264	0.0245	0.0381
π_u	|bias|	0.0004	0.0002	0.0001	0.0011	0.0002
	rmse	0.0156	0.0115	0.0173	0.0156	0.0170
$\pi_{u\|\bar{u}}$	|bias|	0.0005	0.0004	0.0006	0.0004	0.0006
	rmse	0.0078	0.0055	0.0100	0.0082	0.0178

The estimation results obtained under the remaining scenarios are reported in Appendix A.1. Table 1 reports the average, over the latent states and response variables, of the bias (in absolute value) and rmse of the conditional mean vectors $\boldsymbol{\mu}_u$, $u = 1, \dots, k$, under the different scenarios. The table also reports the bias and rmse, computed as the average over pairs of response variables, for the variance–covariance matrix $\boldsymbol{\Sigma}$, and as the average over the hidden states and pairs of hidden states for the initial and transition probabilities, respectively.

Simulation results highlight the ability of our approach in recovering the true data-generating mechanism. In particular, we observe that, regarding the estimation of all model parameters, the average bias and rmse are relatively small under all scenarios. Moreover, as expected, they tend to decrease as the sample size increases (second scenario). Furthermore, the rmse tends to increase when considering a model with a higher number of hidden states (third scenario). When the number of response variables increases, leading to a larger amount of available information about clustering, we expect an improvement in the estimation results for the parameters of the latent process. This improvement is not so evident under the fourth scenario, where the bias and rmse are very close to those obtained under the benchmark scenario. This may be due to the fact that, according to the additional variables included in the model, the states are less separated, with higher uncertainty in the allocation of the units to the hidden states. This may lead to a lack of improvement in the estimation results. Finally, model parameters are estimated with good accuracy, even in the presence of missing data. However, the quality of results is negatively affected by the presence of higher rates of intermittent missing data and/or dropout, as observed in the last scenario with $p_{\text{miss}} = p_{\text{drop}} = 0.25$.

In order to evaluate how increasing frequencies of informative dropout affect the results, we also report in Table 2 the average, over the random samples, of the estimated probability of transition to the absorbing/dropout state, denoted as $drop$, under the scenario with $k = 3$ latent states and $r = 3$ response variables. From these results, we may observe that the probability of moving toward the dropout state is appropriately estimated as the dropout proportion increases, regardless of the sample size.

In evaluating the performance of the proposed approach, the computational cost is also of interest. We observe that, under all scenarios, the computing time required by the estimation algorithm to reach the convergence is, in average, of the order of a few minutes. The computational cost increases when moving from $r = 3$ to $r = 6$ response variables, but this increase is, in average, less than proportional to the increase in the number of response variables.

Finally, we consider the issue of selecting the number of latent states in connection with the proposed estimation approach. In this respect, referred to the above simulation scenarios with $k = 3$ states, we run the estimation algorithm for a varying number of states ranging from 1 to 5, and we selected the optimal k by identifying the model corresponding to the minimum of AIC and BIC. Results show that, regardless the simulated scenario, the suggested BIC is able to correctly identify the true number of hidden states in all of the $B = 250$ samples. On the other hand, as expected, the AIC tends to select a higher number of states, even if the true k is correctly identified in most samples. More in detail, among the different scenarios, the frequency of samples in which the number of states is correctly selected by the AIC index varies from a minimum of 0.7 to a maximum of 0.9.

TABLE 2 Averaged estimated probability of transition toward the dropout state under the HM model with $k = 3$ latent states and $r = 3$ response variables.

		$n = 500$	$n = 1000$
$p_{\text{miss}} = p_{\text{drop}} = 0.01$	$u = 1$	0.010	0.010
	$u = 2$	0.010	0.010
	$u = 3$	0.010	0.010
	$drop$	1.000	1.000
$p_{\text{miss}} = p_{\text{drop}} = 0.05$	$u = 1$	0.050	0.050
	$u = 2$	0.050	0.050
	$u = 3$	0.051	0.050
	$drop$	1.000	1.000
$p_{\text{miss}} = p_{\text{drop}} = 0.1$	$u = 1$	0.098	0.099
	$u = 2$	0.099	0.099
	$u = 3$	0.100	0.099
	$drop$	1.000	1.000
$p_{\text{miss}} = p_{\text{drop}} = 0.25$	$u = 1$	0.251	0.251
	$u = 2$	0.253	0.250
	$u = 3$	0.247	0.248
	$drop$	1.000	1.000

6 | APPLICATION

In the following, we describe the historical data on PBC, and then we illustrate the results obtained through the application of the proposed model on these data.

6.1 | Data description

The data come from biochemical measurements collected prospectively by the Mayo Clinic from January 1974 to May 1984 (Murtaugh et al., 1994). They are collected through a randomized control trial referred to the PBC, which is a liver disease implying inflammatory destruction of the bile ducts and eventually leads to cirrhosis of the liver (Dickson et al., 1989). It is a chronic disease of unknown causes with a prevalence of about 50-cases-per-million population. It is important to remark that this study is very popular since it is one of the last allowing the natural history of the disease for those patients treated with only supporting care or its equivalent (Fleming & Harrington, 1991).

The available data are referred to $n = 312$ patients, some of which (158) were randomized to D-penicillamine and some others (154) with placebo. The original clinical protocol for these patients specified visits at 6 months, and annually after that. We considered time occasions at 6 months from the baseline, thus accounting for missing observations, missing visits, and dropout in a period of 29 time occasions. As reported in Rizopoulos (2012, p. 2) by the end of the study 140 patients had died, 29 received a transplant, and 143 were still alive. These data have been frequently analyzed through the Cox hazard model (Cox, 1972) and, more recently, by joint models (Bartolucci & Farcomeni, 2015; Rizopoulos, 2012). In these previous works, despite the immunosuppressive properties of D-penicillamine, no relevant differences were observed between the distribution of treated and untreated patients' survival times. In this context, it is often of interest to account for multivariate analysis of the longitudinally collected measurements for the diagnosing of liver diseases. Moreover, as remarked in Rizopoulos (2012), physicians are interested in measuring the joint association of the levels of biomarkers with the risk of death. A recent study published in Mulcahy et al. (2022), using longitudinal data to study PBC, remarks the importance of a proper identification of subgroups with distinct disease trajectories especially to making clinical decisions about therapy.

In the present application, we considered the following biochemical variables: *bilirubin*, *cholesterol*, *albumin*, *platelets*, *prothrombin*, *alkaline*, and *transaminase*. To account for some extreme observations in the data, we chose to consider the natural logarithm of the biomarkers. In addition to drug use, we accounted for the effect of gender and age as covariates. Age is considered a continuous time-varying covariate and, along with gender and drug that are time-fixed binary

TABLE 3 Results from the fitting of the multivariate HM models for an increasing number of hidden states (k).

k	$\hat{\ell}_k$	$\#par_k$	BIC_k	AIC_k
1	−4618.66	36	9444.07	9309.33
2	−3679.84	47	7629.59	7453.67
3	−3242.58	60	6829.74	6605.16
4	−2879.04	75	6188.81	5908.09
5	−2561.74	92	5651.85	5307.49
6	−2404.29	111	5446.06	5030.58
7	−2314.55	132	5387.18	4893.11
8	−2215.85	155	5321.87	4741.70

TABLE 4 Estimated conditional means $\boldsymbol{\mu}_u$, $u = 1, \ldots, k$, of the biomarkers (in logarithm), under the HM model with $k = 5$ hidden states.

	u				
Responses	**1**	**2**	**3**	**4**	**5**
Bilirubin	−0.432	0.136	0.865	2.020	2.411
Cholesterol	5.508	5.783	5.496	6.146	5.415
Albumin	1.279	1.270	1.137	1.177	0.940
Platelets	5.477	5.556	4.776	5.525	5.010
Prothrombin	2.339	2.441	2.396	2.363	2.578
Alkaline	6.430	7.270	6.824	7.611	7.033
Transaminase	4.086	4.769	4.664	5.163	5.070

covariates, we investigate their association with the dropout risk. Overall, the research questions we try to address by analyzing these data are the following: (i) if and how it is possible to characterize distinct groups of patients on the basis of biomarkers, (ii) which is the most suitable number of these groups, (iii) how being classified in these groups is related to the risk of death, (iv) how individuals move between these groups according to the covariates with the possibility to predict individual-specific trajectories.

Descriptive statistics of the responses, dropout rates for treated and untreated patients, and additional details on covariates are provided in Section A.2 of the Appendix, along with the figures of the observed values for every patient on each response showing the missing pattern at every time occasion. We notice that the sample is not balanced according to gender since 88% are women. In the whole sample, the median age is 50: the youngest patient is 26 years old and the oldest 78.

6.2 | Results

The proposed HM model allows us to jointly account for the missing mechanism and the complex censoring mechanism (Rubin, 1976) involved in the PBC study. First, we estimated the HM model without covariates and with homogeneous transition probabilities in order to select a suitable number of groups of patients. The initialization strategy illustrated in Section 4.2 is adopted for the EM algorithm. In particular, after a deterministic initialization, a number of random initializations equal to $10 \times (k - 1)$ is used, with k denoting the number of latent states ranging from 1 to 8. Additional details on the computational aspects are reported in Appendix A.3. Table 3 shows that the decrease in the BIC index obtained with the model that has more than 5 latent states is relatively lower than that obtained with fewer states. In fact, the decrease of BIC_k from 5 to 6 latent states is of 3.64% to be compared with a decrease of 8.67% from 4 to 5 states and even larger decreases when moving from 1 to 2, from 2 to 3, and from 3 to 4 states. Therefore, as discussed in Section 4.3, for the parsimony principle, we selected the HM model with $k = 5$ latent states. Then, we estimated the HM model proposed in Section 3.2 including the available covariates and keeping the number of states fixed at $k = 5$. Table 4 shows the estimated conditional means $\boldsymbol{\mu}_u$, $u = 1, \ldots, 5$, of the biomarkers (in logarithm) obtained under this model. For a better understanding of these values, Appendix A.2 (see Table A6) reports the values of the averages for each latent state in the original scale. Note that it is always possible to order the states according to the informative content of the application.

TABLE 5 Estimated variance–covariances (lower part, in bold estimated variances) and estimated partial correlations (upper part) under the HM model with $k = 5$ hidden states.

Responses	*Biril*	*Chol*	*Albu*	*Plat*	*Proth*	*Alka*	*Tran*
Bilirubin	**0.270**	0.151	0.017	−0.102	0.097	0.043	0.255
Cholesterol	0.027	**0.087**	−0.011	0.115	−0.048	0.210	0.014
Albumin	0.000	−0.001	**0.017**	−0.026	−0.072	−0.066	0.015
Platelets	−0.016	0.014	−0.002	**0.118**	−0.072	0.150	−0.061
Prothrombin	0.005	−0.001	−0.001	−0.002	**0.008**	0.043	−0.020
Alkaline	0.039	0.038	−0.005	0.026	0.002	**0.246**	0.279
Transaminase	0.062	0.015	−0.000	−0.005	0.001	0.062	**0.161**

TABLE 6 Estimates of the logit regression parameters of the initial probability to belong to the other latent states with respect to the first state under the HM model with $k = 5$ hidden states (significant at †10%, *5%, **1%).

Effect	$\widehat{\beta}_{12}$	$\widehat{\beta}_{13}$	$\widehat{\beta}_{14}$	$\widehat{\beta}_{15}$
Intercept	4.403*	−0.624	4.048†	−9.103**
Drug	−0.341	0.203	−0.663	−0.638
Female	−1.204	−1.329	−1.616	5.430**
Age	−0.046*	0.023	−0.043†	0.047

In this context, we ordered the states according to increasing *bilirubin* levels, since it is the strongest univariate predictor of survival, and it distinguishes patients with good and poor prognosis. Patients in the first group show normal *bilirubin* levels; therefore, the illness is not really active because they also have the lowest levels of *platelets* and *alkaline*. The second and third states are those of patients in which the disease appears at the early stages. In particular, those in the third state have higher *bilirubin* levels but lower *platelets*, and this may be a group of patients with partial hypertension. We notice that the fourth and the fifth states include patients in the worst health conditions since they show the highest *bilirubin* and *alkaline* values. However, the fifth state is also characterized by the lowest average of *albumin*, the highest average of *prothrombin*, and by high values of *alkaline* and *transaminase*, and therefore this proves to be the worst group in terms of disease prognosis.

Table 5 shows the estimated variances and covariances and the partial correlations, from which we observe that *bilirubin* and *cholesterol* have a positive correlation (0.151) given all the remaining biochemical measurements, as well as *alkaline* and *transaminase* (0.279) and *alkaline* and *cholesterol* (0.210). We also observe a negative partial correlation between *albumin* and *prothrombin* (−0.072).

Table 6 provides the estimated regression parameters of the initial probabilities as in Expression (11), where the statistical significance of the coefficients is established according to the standard errors obtained with the nonparametric bootstrap and reported in Table A10 of Appendix A.3.

The estimated gender log-odds in Table 6 referred to the fifth state is positive and significant, indicating that the probability of being in the fifth state at the beginning of the study is higher for females with respect to males.

The estimated parameters in Table 7 refer to the coefficients affecting the transition from level $\bar{u}$ to level u of the latent process. We notice that the last column of the first panel of the table contains the parameter estimates measuring the influence of each covariate on the transition from the first state, corresponding to patients in relatively good health conditions, to the dropout state. The influence of gender is positive, indicating that this transition probability is higher for females than for males. This effect is confirmed also on the transition from the second state to dropout. Interestingly, treated patients in the first state show a higher probability of moving to the dropout state with respect to those untreated. The other estimated values refer to all possible transitions.

Another interpretation of the effects of covariates can be retrieved by looking at the averaged initial and transition probabilities defined by categories of patients. Table 8 reports these probabilities, computed with respect to a typical profile of the disease referred to a female with an average age between 48 and 52 years old at the baseline, who has not been treated with D-penicillamine. We recall that these probabilities are obtained according to the estimated regression parameters referred to Expressions (11) and (12) and are computed as average over patients according to the chosen profile, rounded to the third decimal digit. We observe that at the baseline, the second state is the most likely for this profile since 45% of

TABLE 7 Estimates of the logit regression parameters of the transition probabilities under the HM model with $k = 5$ hidden states (significant at †10%, *5%, **1%).

Effect	$\widehat{\gamma}_{12}$	$\widehat{\gamma}_{13}$	$\widehat{\gamma}_{14}$	$\widehat{\gamma}_{15}$	$\widehat{\gamma}_{1drop}$
Intercept	20.506	−17.921**	−31.507**	−20.653	−20.304**
Drug	−17.895**	8.795**	−4.017**	−1.745	7.761**
Female	−20.685**	6.639**	−3.181**	−6.747	5.205**
Age	−0.316	−0.044	0.176**	0.309	0.023
Effect	$\widehat{\gamma}_{21}$	$\widehat{\gamma}_{23}$	$\widehat{\gamma}_{24}$	$\widehat{\gamma}_{25}$	$\widehat{\gamma}_{2drop}$
Intercept	−5.569	−4.015*	−2.686	−20.658**	−10.749**
Drug	1.319	0.601	0.266	8.223*	1.193
Female	−0.420	−1.150†	−0.347	−10.402**	5.329**
Age	0.035	0.033	−0.015	0.165	−0.012
Effect	$\widehat{\gamma}_{31}$	$\widehat{\gamma}_{32}$	$\widehat{\gamma}_{34}$	$\widehat{\gamma}_{35}$	$\widehat{\gamma}_{3drop}$
Intercept	−20.780**	−36.680**	10.020**	−4.705	−7.047
Drug	−4.293**	1.207**	21.313**	−0.040	1.075
Female	2.015**	−6.604**	−25.621**	0.117	−1.430
Age	−0.017**	0.231**	−0.757**	0.034	0.045
Effect	$\widehat{\gamma}_{41}$	$\widehat{\gamma}_{42}$	$\widehat{\gamma}_{43}$	$\widehat{\gamma}_{45}$	$\widehat{\gamma}_{4drop}$
Intercept	−24.962	−7.228	−5.091	−3.817*	−8.306
Drug	4.426	−6.792**	2.355	−0.894	1.203
Female	5.484	4.579	9.356	−0.674	1.127
Age	0.173	−0.050	−0.227	0.044†	0.048
Effect	$\widehat{\gamma}_{51}$	$\widehat{\gamma}_{52}$	$\widehat{\gamma}_{53}$	$\widehat{\gamma}_{54}$	$\widehat{\gamma}_{5drop}$
Intercept	−5.866†	−6.887**	9.574**	3.689	−0.825
Drug	−1.389	−3.753**	−4.542	5.982	−0.595
Female	6.848**	4.468**	23.414**	−0.674	0.400
Age	−0.051	−0.289**	−0.892**	−0.261	−0.001

TABLE 8 Initial and transition probabilities under the HM model with $k = 5$ hidden states for a typical patient profile: untreated females with age between 48 and 52 years old.

	u					
	1	**2**	**3**	**4**	**5**	***drop***
$\widehat{\pi}_u$	**0.181**	**0.453**	**0.080**	**0.236**	**0.049**	**0.000**
$\widehat{\pi}_{u\|1}$	1.000	0.000	0.000	0.000	0.000	0.000
$\widehat{\pi}_{u\|2}$	0.017	0.927	0.035	0.019	0.000	0.002
$\widehat{\pi}_{u\|3}$	0.000	0.000	0.930	0.000	0.068	0.003
$\widehat{\pi}_{u\|4}$	0.000	0.004	0.000	0.866	0.120	0.010
$\widehat{\pi}_{u\|5}$	0.086	0.000	0.000	0.000	0.571	0.344
$\widehat{\pi}_{u\|drop}$	0.000	0.000	0.000	0.000	0.000	1.000

females are in this state and only 18% are in the first state and interestingly 24% are in the fourth state. The fifth state, which is referred to patients with worse health conditions, includes the 4.9% of patients of this profile. According to the transition probabilities, the most persistent state is the first, whereas the state with the highest probability toward dropout is the fifth ($\widehat{\pi}_{drop|5} = 0.344$) followed by the fourth state ($\widehat{\pi}_{drop|4} = 0.01$). Patients in the fourth latent state have a probability of moving to the fifth state equal to $\widehat{\pi}_{5|4} = 0.120$, which is the highest estimated probability out of the main diagonal of the transition matrix, excluding the probabilities referred to the dropout state. These results also show that higher values of *serum bilirubin, prothrombin*, and *alkaline*, characterizing the fourth state, are strongly related to a severe disease progression and the risk of death. See Appendix A.3 for additional comparisons across drug use, age, and gender.

An important aspect of this medical application is predicting the sequence of latent states to evaluate the time-varying patient risk of death. Figure 1 shows the relative frequency of the patients assigned to each state at every time occasion.

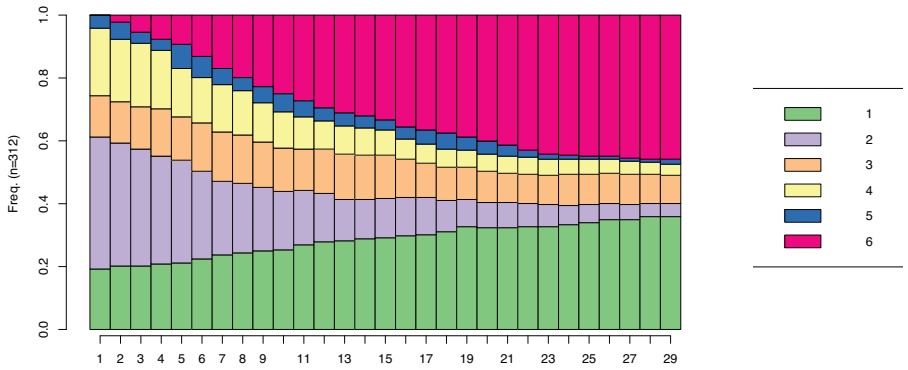

FIGURE 1 Relative frequency of the patients assigned to each group at each time occasion under the estimated HM model; the sixth state (in pink) is the dropout state.

Interestingly, from Figure 1, we notice that at the baseline the proportion of patients assigned to the second state is around 0.42, instead it is around 0.19 for the patients assigned to the first state and only 0.04 for those assigned to the fifth state. At the end of the period, instead, 35% are assigned to the first state, thus indicating recovered conditions, and 46% to the sixth state, corresponding to the dropout state. The frequencies of patients assigned to the second and fifth states are decreasing over time, while those of third state are almost the same over years.

7 | CONCLUSIONS

In this paper, we propose an HM model for continuous multivariate longitudinal data addressing three different missing response types. The first two are intermittent and correspond to completely or partially missing responses for a given occasion; these types are denoted by (I) and (II), respectively. The third type of missing data corresponds to dropout, indicated by type (III), which typically occurs due to the early termination from the trial, and we handle it by including an extra absorbing state in the model so that this type of missing is informative. Estimation is carried out by an extended EM algorithm implemented to account for missing values of types (I) and (II) on the basis of the MAR assumption and for the dropout, type (III), on the basis of the extra hidden state. This inferential approach is also developed to estimate the parameters of the model when individual covariates are included in the distribution of the latent process. In such a context, it may be of interest to evaluate the effect of these covariates on the transition toward the dropout state. We notice that in the presence of missing data, the proposed approach also allows us to perform a sort of multiple imputation so as to predict the missing responses conditionally or unconditionally to the assigned latent state.

The simulation study allowed us to conclude that the model parameters are properly estimated even with a relatively large proportion of missing responses and dropout. The application related to multivariate data about PBC referred to several biochemical measurements of liver function turned out to be particularly challenging. The data are very sparse due to missing visits of the patients, and some variables were not collected at each visit. Moreover, several deaths are registered, which is considered a dropout indicator. With the proposed approach, we identified five groups of patients differing in the severity of this rare disease and in their transitions across states and toward the dropout state over time. According to the available covariates, it was possible to predict the distributions of survival times for the groups of patients through the decoded states. Adopting a model-based approach like the HM model, which considers all possible missing patterns, allows us not only to infer the survival probability over time but also to disentangle disease severity based on biomarkers across relevant population subgroups of patients and to compare survival curves across these groups. Additionally, the model permits the prediction of individual-specific trajectories for those patients with intermitting missing values, and thus it allows us to identify patients who need to be prioritized with treatments. Estimating a model only accounting for missing responses and not for dropout provides different results with respect to those obtained with the proposed one. In particular, in our application, the average transition probability matrix differs from that reported in the paper since almost all individuals in the fourth state transit to the fifth group, that referred to patients with a worse prognosis.

Overall, we consider the proposal to be very flexible, given the possibility of accounting for different types of missingness at the same time, in addition to the possibility of dealing with multivariate continuous longitudinal data. It is important to stress that all continuous variables whose distribution cannot be approximated with a multivariate Gaussian one may be

properly transformed (with Box–Cox, log transformation, and so forth), so that the proposed model may still be effectively applied. Moreover, this approach may be easily extended to the case of categorical response variables, as already introduced in the work of Montanari and Pandolfi (2018) within a simpler model formulation than the one proposed here, where it is possible to rely on the conditional independence of the response variables given the latent variables. Finally, even if it is not always guaranteed that the selected states correspond to meaningful clusters, when adopting a discrete latent variable structure, latent states have in general a straightforward interpretation and may be used to cluster units in a meaningful way.

Clearly, in applying the approach, it is necessary to carefully evaluate how realistic are the model assumptions and, in particular, the assumptions about the missing pattern structures. This could be a limitation of the proposed modeling approach. In this regard, it is important to recall the difference in dealing with missing data of type (I) and (II), which are considered ignorable based on the MAR assumption, and of type (III), which are considered informative. Moreover, the way of accounting for the last type, based on using an extra absorbing state, distinguishes the proposed approach from alternative ones based on shared-parameter formulation and does not require to formulate a survival model. Finally, the proposed EM algorithm could also be adapted to estimate the partially HM model proposed in Bordes and Vandekerkhove (2005), which may be useful to analyze data if, after dropout, new patients are recruited during the study.

We also mention that, if more time occasions are available, it would be possible to dispose of the estimated stationary distribution of the underlying hidden process that could be represented as in van Beest et al. (2019). Other possible extensions of the proposed approach may concern the use with responses of a mixed nature, not all continuous. This amounts to assume that underlying every non-continuous response, there is an unobservable continuous response following a Gaussian distribution and affecting the former by a suitable link function while leaving unaltered the overall structure of the model. An interesting model to explore at this aim would also be the regime switching copula model to properly account for the correlation patterns between series, as proposed in other research fields (Chollete et al., 2009).

ACKNOWLEDGMENTS

We acknowledge Alessio Geruzzi M.D. of the Department of Medicine of the University of Milano-Bicocca, for helpful comments about the results of the illustrative example. F. Bartolucci and S. Pandolfi acknowledge the financial support from the grant "Inferential developments on statistical models for complex data" of the University of Perugia (RICBASE_PROGETTURALE_2017_2019_PANDOLFI). F. Pennoni acknowledges the financial support from the grant "Statistical models for repeated measures with missing data and dropout" of the University of Milano-Bicocca (2020-ATE-0615).

Open Access Funding provided by Universita degli Studi di Perugia within the CRUI-CARE Agreement.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the supporting information of this article.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to the computational complexity of the results.

ORCID

Silvia Pandolfi https://orcid.org/0000-0001-6019-8694

REFERENCES

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.). *Second international symposium on information theory* (pp. 267–281). Akademiai Kiado.

Albert, P., Follmann, D. A., Wang, S. A., & Suh, E. B. (2002). A latent autoregressive model for longitudinal binary data subject to informative missingness. *Biometrics*, *58*, 631–642.

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

Anderson, T. (2003). *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley.

Bacci, S., Pandolfi, S., & Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, *8*, 125–145.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*, 803–821.

Bartolucci, F., & 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. (2015). A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. *Biometrics*, *71*, 80–89.

Bartolucci, F., & Farcomeni, A. (2019). A shared-parameter continuous-time hidden Markov and survival model for longitudinal data with informative dropout. *Statistics in Medicine*, *38*, 1056–1073.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2013). *Latent Markov models for longitudinal data*. Chapman & Hall/CRC Press.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2014). Latent Markov models: A review of a general framework for the analysis of longitudinal data with covariates. *TEST*, *23*, 433–465.

Bartolucci, F., Pandolfi, S., & Pennoni, F. (2017). LMest: An R package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software*, *81*, 1–38.

Bartolucci, F., Pandolfi, S., & Pennoni, F. (2022). Discrete latent variable models. *Annual Review of Statistics and its Application*, *9*, 425–452.

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

Bordes, L., & Vandekerkhove, P. (2005). Statistical inference for partially hidden Markov models. *Communications in Statistics-Theory and Methods*, *34*, 1081–1104.

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science: With applications in R*. Cambridge, Cambridge University Press.

Celeux, G., & Durand, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics*, *23*, 541–564.

Chollete, L., Heinen, A., & Valdesogo, A. (2009). Modeling international financial returns with a multivariate regime-switching copula. *Journal of Financial Econometrics*, *7*, 437–480.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, *34*, 187–202.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Delalleau, O., Courville, A., & Bengio, Y. (2012). Efficient EM training of Gaussian mixtures with missing data. *arXiv preprint arXiv:1209.0521*.

Dempster, A. P., Laird, N. M., & 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.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., & Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, *10*, 1–7.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C*, *43*, 49–93.

Diggle, P. J., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data*. Oxford University Press.

Eirola, E., Lendasse, A., Vandewalle, V., & Biernacki, C. (2014). Mixture of Gaussians for distance estimation with missing data. *Neurocomputing*, *131*, 32–42.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Wiley-Interscience.

Fleming, T., & Harrington, D. (1991). *Counting processes and survival analysis*. John Wiley & Sons.

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

Juang, B., & Rabiner, L. (1991). Hidden Markov models for speech recognition. *Technometrics*, *33*, 251–272.

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., & Rubin, D. B. (2020). *Statistical analysis with missing data*. John Wiley Sons.

Marino, M. F., & Alfó, M. (2015). Latent drop-out based transitions in linear quantile hidden Markov models for longitudinal responses with attrition. *Advances in Data Analysis and Classification*, *9*, 483–502.

Marino, M. F., Tzavidis, N., & Alfó, M. (2018). Mixed hidden Markov quantile regression models for longitudinal data with possibly incomplete sequences. *Statistical Methods in Medical Research*, *27*, 2231–2246.

Maruotti, A. (2015). Handling non-ignorable dropouts in longitudinal data: A conditional model based on a latent Markov heterogeneity structure. *TEST*, *24*, 84–109.

Maruotti, A., & Punzo, A. (2021). Initialization of hidden Markov and semi-Markov models: A critical evaluation of several strategies. *International Statistical Review*, *89*, 447–480.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley.

Montanari, G. E., & Pandolfi, S. (2018). Evaluation of long-term health care services through a latent Markov model with covariates. *Statistical Methods & Applications*, *27*, 151–173.

Mulcahy, V., Rouanet, A., Gerussi, A., Duckworth, A., Flack, S., Carbone, M., Tom, B., & Mells, G. (2022). *Latent class mixed modelling for phenotypic stratification of primary biliary cholangitis patients on first line treatment*. https://arxiv.org/abs/2203.10508

Murtaugh, P. A., Dickson, E. R., Van Dam, G. M., Malinchoc, M., Grambsch, P. M., Langworthy, A. L., & Gips, C. H. (1994). Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. *Hepatology*, *20*, 126–134.

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, *4*, 440–461.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *61*, 479–482.

Pohle, J., Langrock, R., van Beest, F. M., & Schmidt, N. M. (2017). Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, *22*, 270–293.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press.

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, *25*, 127–141.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

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

Shireman, E., Steinley, D., & Brusco, M. J. (2017). Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior Research Methods*, *49*, 282–293.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, *10*, 63–72.

Spagnoli, A., Henderson, R., Boys, R., & Houwing-Duistermaat, J. (2011). A hidden Markov model for informative dropout in longitudinal response data with crisis states. *Statistics & Probability Letters*, *81*, 730–738.

Taavoni, M., Arashi, M., Wang, W. L., & Lin, T. I. (2020). Multivariate *t* semiparametric mixed-effects model for longitudinal data with multiple characteristics. *Journal of Statistical Computation and Simulation*, *92*, 260–281.

van Beest, F. M., Mews, S., Elkenkamp, S., Schuhmann, P., Tsolak, D., Wobbe, T., Bartolino, V., Bastardie, F., Dietz, R., von Dorrien, C., Galatius, A., Karlsson, O., McConnell, B., Nabe-Nielsen, J., Olsen, M. T., Teilmann, J., & Langrock, R. (2019). Classifying grey seal behaviour in relation to environmental variability and commercial fishing activity—A multivariate hidden Markov model. *Scientific Reports*, *9*, 1–14.

Verbeke, G., Fieuws, S., Molenberghs, G., & Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, *23*, 42–59.

Vermunt, J. K., Langeheine, R., & Böckenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, *24*, 179–207.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*, 260–269.

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

Yuan, Y., & Little, R. J. (2009). Mixed-effect hybrid models for longitudinal data with nonignorable dropout. *Biometrics*, *65*, 478–486.

Zhou, T., Daniels, M. J., & Müller, P. (2020). A semiparametric Bayesian approach to dropout in longitudinal studies with auxiliary covariates. *Journal of Computational and Graphical Statistics*, *29*, 1–12.

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov models for time series: An introduction using R*. CRC Press.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Pandolfi, S., Bartolucci, F., & Pennoni, F. (2023). A hidden Markov model for continuous longitudinal data with missing responses and dropout. *Biometrical Journal*, 2200016. https://doi.org/10.1002/bimj.202200016

APPENDIX: ADDITIONAL DETAILS

In the following Section A.1, we show some additional details on the results of the simulation study presented in Section 5, then in Section A.2, we report an accurate description of the historical data used for the applicative example shown in Section 6, and finally, in Section A.3, we illustrate supplementary results of the application of the HM model to these data.

A.1 | Additional simulation results

We report the estimation results obtained under all simulated scenarios. In particular, Table A1 reports the average, over the latent states and response variables, of the bias (in absolute value) and rmse of the conditional mean vectors $\boldsymbol{\mu}_u$, $u = 1, \dots, k$. Table A2 reports the same estimation results, computed as the average over pairs of response variables, of

TABLE A1 Average, over the latent states and response variables, of the bias (in absolute value) and rmse of the conditional mean vectors $\boldsymbol{\mu}_u$, $u = 1, \ldots, k$.

		$r = 3$				$r = 6$			
		$k = 2$		$k = 3$		$k = 2$		$k = 3$	
		$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
$p_{\text{miss}} = p_{\text{drop}} = 0.01$	| bias|	0.0013	0.0008	0.0016	0.0010	0.0007	0.0010	0.0013	0.0009
	rmse	0.0297	0.0206	0.0371	0.0266	0.0285	0.0195	0.0355	0.0253
$p_{\text{miss}} = p_{\text{drop}} = 0.05$	|bias|	0.0021	0.0010	0.0023	0.0016	0.0019	0.0013	0.0022	0.0013
	rmse	0.0324	0.0230	0.0400	0.0281	0.0310	0.0216	0.0381	0.0273
$p_{\text{miss}} = p_{\text{drop}} = 0.10$	| bias|	0.0025	0.0011	0.0021	0.0021	0.0017	0.0006	0.0027	0.0029
	rmse	0.0346	0.0240	0.0441	0.0313	0.0334	0.0233	0.0428	0.0297
$p_{\text{miss}} = p_{\text{drop}} = 0.25$	| bias |	0.0023	0.0015	0.0019	0.0018	0.0026	0.0017	0.0020	0.0032
	rmse	0.0433	0.0305	0.0600	0.0416	0.0421	0.0297	0.0543	0.0373

TABLE A2 Average, over pairs of response variables, of the bias (in absolute value) and rmse of the variance–covariance matrix $\boldsymbol{\Sigma}$.

		$r = 3$				$r = 6$			
		$k = 2$		$k = 3$		$k = 2$		$k = 3$	
		$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
$p_{\text{miss}} = p_{\text{drop}} = 0.01$	| bias|	0.0013	0.0006	0.0012	0.0012	0.0011	0.0009	0.0006	0.0005
	rmse	0.0250	0.0176	0.0264	0.0187	0.0245	0.0174	0.0245	0.0176
$p_{\text{miss}} = p_{\text{drop}} = 0.05$	| bias|	0.0008	0.0010	0.0019	0.0008	0.0007	0.0005	0.0015	0.0008
	rmse	0.0283	0.0203	0.0297	0.0206	0.0252	0.0181	0.0274	0.0194
$p_{\text{miss}} = p_{\text{drop}} = 0.10$	| bias|	0.0015	0.0014	0.0020	0.0017	0.0023	0.0007	0.0012	0.0009
	rmse	0.0304	0.0213	0.0313	0.0219	0.0282	0.0191	0.0287	0.0199
$p_{\text{miss}} = p_{\text{drop}} = 0.25$	| bias|	0.0011	0.0014	0.0019	0.0017	0.0029	0.0009	0.0017	0.0012
	rmse	0.0381	0.0280	0.0423	0.0293	0.0332	0.0254	0.0368	0.0257

TABLE A3 Average, over the latent states, of the bias (in absolute value) and rmse of the initial probabilities π_u, $u = 1, \ldots, k$.

		$r = 3$				$r = 6$			
		$k = 2$		$k = 3$		$k = 2$		$k = 3$	
		$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
$p_{\text{miss}} = p_{\text{drop}} = 0.01$	|bias|	0.0004	0.0002	0.0001	0.0004	0.0011	0.0010	0.0005	0.0006
	rmse	0.0156	0.0115	0.0173	0.0125	0.0156	0.0110	0.0171	0.0122
$p_{\text{miss}} = p_{\text{drop}} = 0.05$	|bias|	0.0005	0.0002	0.0006	0.0006	0.0004	0.0000	0.0014	0.0006
	rmse	0.0146	0.0106	0.0175	0.0125	0.0156	0.0105	0.0177	0.0122
$p_{\text{miss}} = p_{\text{drop}} = 0.10$	|bias|	0.0006	0.0000	0.0009	0.0006	0.0011	0.0004	0.0014	0.0007
	rmse	0.0161	0.0106	0.0176	0.0127	0.0163	0.0108	0.0159	0.0117
$p_{\text{miss}} = p_{\text{drop}} = 0.25$	|bias|	0.0002	0.0001	0.0010	0.0004	0.0019	0.0015	0.0021	0.0012
	rmse	0.0170	0.0116	0.0195	0.0143	0.0142	0.0110	0.0175	0.0128

the variance–covariance matrix $\boldsymbol{\Sigma}$. Finally, Tables A3 and A4 report the estimation results for the initial and transition probabilities, respectively.

A.2 | Data description

The applicative example, as illustrated in Section 6.1, concerns data referred to biochemical measurements related to a double-blinded randomized trial in PBC conducted by the Mayo Clinic from 1974 to 1984. In the following, we report some additional details and descriptions of these data.

PBC is a rare but fatal chronic liver disease of unknown cause, with a prevalence of about 50 cases per million inhabitants. The primary pathological event of this disease seems to be the destruction of intralobular bile ducts, mediated

TABLE A4 Average, over pairs of latent states, of the bias (in absolute value) and rmse of the transition probabilities $\pi_{u|\bar{u}}$, $u, \bar{u} = 1, \ldots, k+1$.

		$r=3$				$r=6$			
		$k=2$		$k=3$		$k=2$		$k=3$	
		$n=500$	$n=1000$	$n=500$	$n=1000$	$n=500$	$n=1000$	$n=500$	$n=1000$
$p_{\text{miss}} = p_{\text{drop}} = 0.01$	|bias|	0.0005	0.0004	0.0006	0.0004	0.0004	0.0003	0.0006	0.0006
	rmse	0.0078	0.0055	0.0100	0.0071	0.0082	0.0056	0.0098	0.0069
$p_{\text{miss}} = p_{\text{drop}} = 0.05$	|bias|	0.0006	0.0002	0.0009	0.0005	0.0005	0.0003	0.0004	0.0002
	rmse	0.0103	0.0072	0.0123	0.0088	0.0098	0.0070	0.0121	0.0083
$p_{\text{miss}} = p_{\text{drop}} = 0.10$	|bias|	0.0008	0.0005	0.0006	0.0007	0.0007	0.0008	0.0009	0.0005
	rmse	0.0120	0.0088	0.0147	0.0103	0.0121	0.0083	0.0141	0.0097
$p_{\text{miss}} = p_{\text{drop}} = 0.25$	|bias|	0.0006	0.0007	0.0014	0.0009	0.0008	0.0007	0.0008	0.0005
	rmse	0.0178	0.0120	0.0200	0.0144	0.0166	0.0120	0.0195	0.0137

by immunological mechanisms. To be eligible for these trials, patients had to meet well-established clinical, biochemical, serologic, and histologic criteria for PBC. These are historical data used in many studies since they come from a research that has been one of the most extensive controlled clinical trials on PBC for a long time. As remarked in Fleming and Harrington (1991), historical data are important since clinical trials are challenging to complete in rare diseases, and they permit the study of the natural history of the disease before the liver transplant was considered as a standard practice. Moreover, disposing of accurate and parsimonious models that can predict survival time based on inexpensive, noninvasive, and readily available measurements is a very relevant feature in clinical science.

The biochemical variables used as responses in the empirical illustration proposed in Section 6 of the paper are the following[1]:

(1) *Serum bilirubin* (Bilir in mg/dL) is a liver bile pigment; normal adult levels range between 0.3 and 1.0 (min log −1.2, max log 0). Values above 1.2 are synonymous with liver failure.
(2) *Serum cholesterol* (Chol in mg/dL) is a blood lipoprotein; normal adult levels range between 150 and 199 (min log 4.78, max log 5.39). Borderline values are considered between 200 and 239.
(3) *Serum albumin* (Albu in gm/dL) is a protein found in the blood; normal adult levels range between 3.5 and 5.4 (min log 1.5, max log 1.61). Low albumin values may result from liver malfunction.
(4) *Platelets* (Plat, counts per cubic mL/1000) are components of blood; normal adult levels range between 150 and 350 (min log 5.01, max log 6.11).
(5) *Prothrombin* (Prot) is a blood coagulation agent time and time is recorded until a blood sample begins coagulation in a certain laboratory test in seconds; normal adult levels range between 10 and 13 seconds (min log 2.17, max log 2.45).
(6) *Alkaline phosphatase* (Alka, in U/L) is an enzyme; normal adult levels range between 36 and 150 (min log 3, max log 4.94). High values of alkaline phosphatase can occur in the presence of liver disease.
(7) *Transaminase* (Tran or SGOT in U/mL) adult normal levels range between 0 and 35 (min log 2.08, max log 3.69). In liver damage, an increase of transaminase in blood concentration is observed up to 5- to 10-fold in the presence of severe damage.

In particular, *serum bilirubin* level is considered a strong indicator of disease progression and a strong predictor of survival (Fleming & Harrington, 1991; Taavoni et al., 2020). *Bilirubin* and *albumin* are two of the primary indicators to help evaluate and track the absence of liver diseases. A significantly higher level than bilirubin's standards excreted in bile usually indicates certain diseases. *Serum albumin* may be harmful to humans having too high or too low circulating levels. Typically, there exist some associations between the levels of *serum bilirubin* and *serum albumin*, and thus it is important to account for a joint analysis of the longitudinally collected biochemical markers for the diagnosing of liver diseases as that proposed in the paper through the multivariate HM model.

The original clinical protocol for the patients specified visits at 6 months, 1 year, and annually after that. In the analyses, we considered time occasions at 6 months from the baseline, thus accounting for missing observations, missing visits, and

[1] Values reported according to those provided at the following website accessed on April 2021: https://www.msdmanuals.com/professional/resources/normal-laboratory-values/normal-laboratory-values

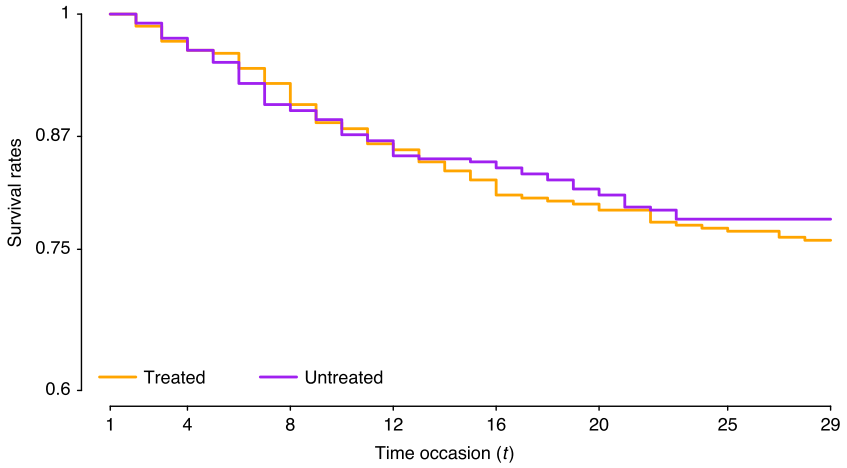

FIGURE A1 Observed survival rates for treated (with D-penicillamine) and untreated patients.

TABLE A5 Average and standard deviation (SD) of the responses (in logarithm): bilirubin (Bilir), cholesterol (Chol), albumin (Albu), platelets (Plat), prothrombin (Prot), alkaline (Alka), and transaminase (Tran) from the baseline to the end of the study.

t	Bilir	SD	Chol	SD	Albu	SD	Plat	SD	Prot	SD	Alka	SD	Tran	SD
1	0.57	1.03	5.80	0.43	1.25	0.13	5.50	0.40	2.37	0.08	7.27	0.72	4.71	0.45
2	0.47	1.07	6.05	0.27	1.25	0.15	5.42	0.43	2.38	0.12	7.10	0.64	4.72	0.57
3	0.53	1.04	5.76	0.36	1.24	0.15	5.42	0.42	2.36	0.07	7.10	0.64	4.73	0.54
4	0.48	1.01	5.78	0.34	1.16	0.21	5.27	0.43	2.39	0.07	6.73	0.75	4.55	0.58
5	0.70	1.14	5.69	0.36	1.22	0.15	5.40	0.48	2.38	0.13	7.07	0.67	4.72	0.56
6	0.93	1.47	5.56	0.35	1.36	1.20	5.34	0.36	2.43	0.16	6.96	0.84	4.62	0.58
7	0.59	1.11	5.72	0.36	1.23	0.13	5.39	0.48	2.37	0.08	6.97	0.61	4.65	0.56
8	0.66	1.25	5.72	0.40	1.16	0.15	5.31	0.48	2.42	0.12	6.91	0.66	4.62	0.52
9	0.63	1.16	5.67	0.48	1.19	0.17	5.32	0.45	2.39	0.12	6.92	0.61	4.63	0.56
10	0.40	0.92	5.50	0.29	1.23	0.15	5.25	0.47	2.38	0.10	6.93	0.62	4.59	0.43
11	0.68	1.18	5.62	0.38	1.18	0.19	5.32	0.48	2.39	0.11	6.82	0.60	4.58	0.60
12	0.75	1.14	5.61	0.30	1.15	0.16	5.23	0.45	2.41	0.10	6.92	0.57	4.60	0.53
13	0.73	1.17	5.61	0.31	1.17	0.15	5.27	0.47	2.40	0.09	6.73	0.58	4.59	0.65
14	0.46	1.13	5.48	0.22	1.18	0.14	5.08	0.55	2.43	0.10	6.77	0.46	4.46	0.50
15	0.74	1.07	5.60	0.32	1.15	0.16	5.15	0.45	2.42	0.10	6.72	0.50	4.58	0.60
16	0.54	1.15	5.54	0.35	1.13	0.26	5.23	0.49	2.40	0.07	6.77	0.52	4.46	0.55
17	0.61	1.11	5.59	0.23	1.13	0.17	5.15	0.42	2.43	0.08	6.72	0.50	4.55	0.67
18	0.41	1.23	5.53	0.31	1.12	0.18	5.31	0.43	2.43	0.07	6.68	0.44	4.44	0.66
19	0.62	1.15	5.58	0.32	1.12	0.20	5.17	0.46	2.48	0.21	6.66	0.59	4.51	0.71
20	1.22	1.43	5.64	0.21	0.99	0.25	5.28	0.39	2.51	0.15	6.89	0.41	4.63	0.42
21	0.77	1.22	5.61	0.29	1.18	0.12	5.23	0.43	2.46	0.08	6.76	0.43	4.56	0.59
22	0.21	0.37	5.78	0.06	1.26	0.04	5.14	0.50	2.38	0.02	6.82	0.25	4.32	0.25
23	0.57	1.29	5.46	0.30	1.11	0.19	5.22	0.48	2.46	0.05	6.73	0.45	4.59	0.63
24	0.51	0.82	5.60	0.17	1.20	0.17	5.01	0.41	2.44	0.11	6.59	0.26	4.40	0.37
25	0.42	1.03	5.57	0.20	1.71	1.89	5.13	0.45	2.47	0.06	6.73	0.53	4.38	0.59
26	1.55	2.19	5.74	0.12	1.31	0.02	5.24	0.01	2.46	0.01	6.44	0.04	5.04	0.77
27	0.44	1.25	5.38	0.16	1.20	0.12	5.10	0.46	2.47	0.05	6.60	0.45	4.30	0.54
28	1.22	1.60	5.80	0.12	1.09	0.06	4.85	0.84	2.49	0.01	7.40	1.01	4.77	1.14
29	0.98	1.41	5.51	0.19	1.12	0.24	5.12	0.30	2.46	0.07	6.97	0.71	4.61	0.68

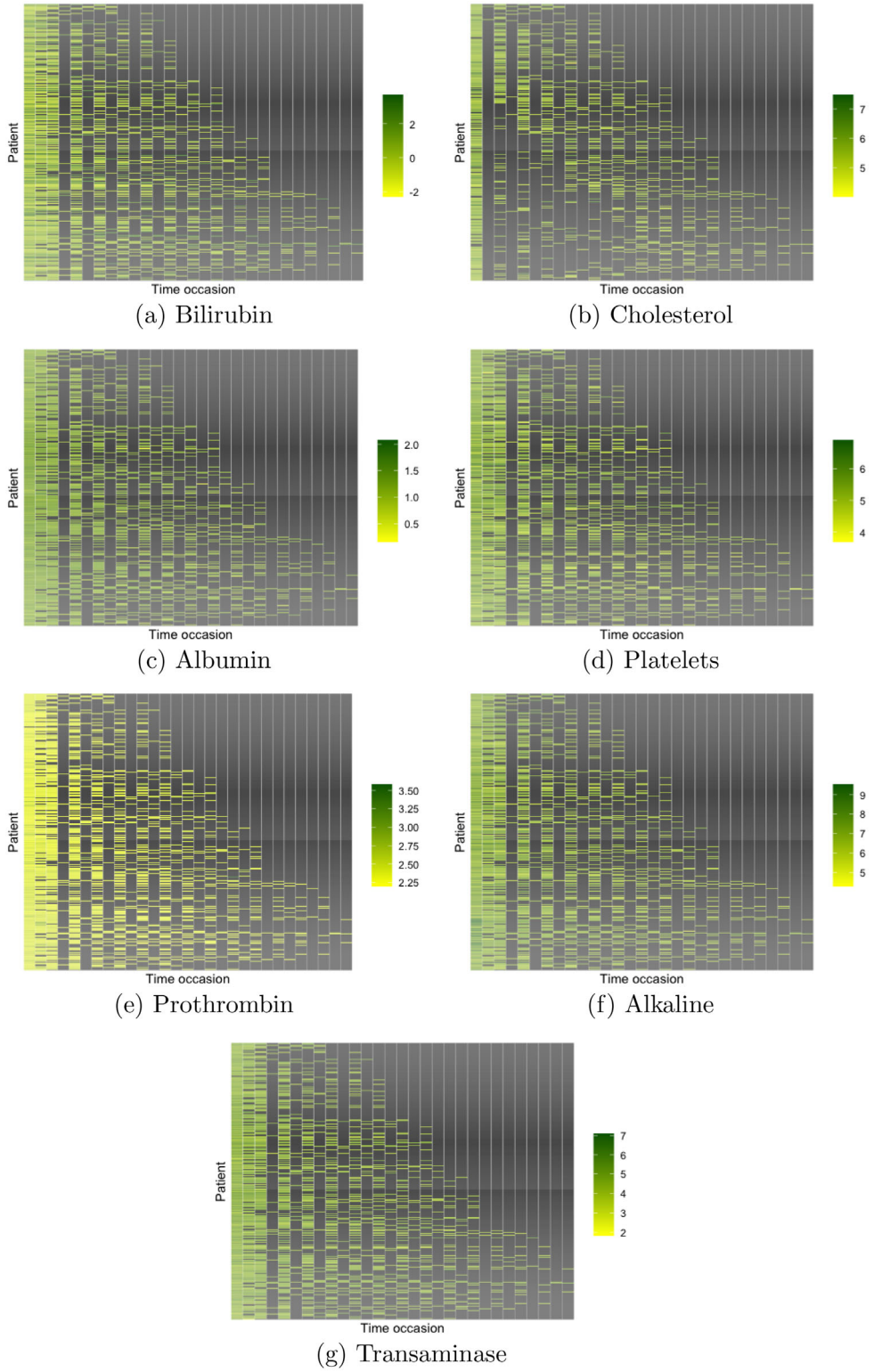

FIGURE A2 Observed values for every patient and response variable over time occasions; missing values are depicted in black.

dropout in a period of 29 time occasions. Despite the immunosuppressive properties of the D-penicillamine, which was the main treatment at that time, no detectable differences were observed between survival time distributions for placebo and treated groups. At the end of the study, 140 patients died, 29 had liver transplantation, and 143 were still alive. Figure A1 shows the survival rates for treated and untreated patients over time.

TABLE A6 Estimated conditional means μ_u, $u = 1, \ldots, k$, of the biomarkers (original scale), under the HM model with $k = 5$ hidden states.

	1	2	3	4	5
Bilirubin	0.650	1.150	2.370	7.540	11.140
Cholesterol	246.580	324.760	243.750	467.010	224.640
Albumin	3.590	3.560	3.120	3.240	2.560
Platelets	239.010	258.790	118.590	250.910	149.880
Prothrombin	10.620	10.370	11.480	10.970	13.170
Alkaline	620.450	1,435.860	919.530	2,020.120	1,133.750
Transaminase	59.510	117.770	106.090	174.660	159.250

TABLE A7 Averaged initial and transition probabilities with respect to the drug use (upper panel) or not use (bottom panel).

	u					
	1	**2**	**3**	**4**	**5**	***drop***
$\hat{\pi}_u$	**0.219**	**0.419**	**0.155**	**0.169**	**0.038**	**0.000**
$\hat{\pi}_{u\|1}$	0.955	0.000	0.006	0.000	0.037	0.002
$\hat{\pi}_{u\|2}$	0.064	0.811	0.073	0.022	0.024	0.005
$\hat{\pi}_{u\|3}$	0.000	0.000	0.909	0.009	0.069	0.013
$\hat{\pi}_{u\|4}$	0.021	0.000	0.014	0.865	0.065	0.035
$\hat{\pi}_{u\|5}$	0.024	0.000	0.003	0.028	0.719	0.226
$\hat{\pi}_{u\|drop}$	0.000	0.000	0.000	0.000	0.000	1.000
	1	**2**	**3**	**4**	**5**	***drop***
$\hat{\pi}_u$	**0.162**	**0.458**	**0.086**	**0.248**	**0.047**	**0.000**
$\hat{\pi}_{u\|1}$	0.918	0.044	0.000	0.000	0.038	0.000
$\hat{\pi}_{u\|2}$	0.018	0.917	0.043	0.020	0.000	0.002
$\hat{\pi}_{u\|3}$	0.000	0.000	0.929	0.000	0.067	0.004
$\hat{\pi}_{u\|4}$	0.000	0.004	0.002	0.852	0.132	0.010
$\hat{\pi}_{u\|5}$	0.087	0.000	0.027	0.000	0.561	0.325
$\hat{\pi}_{u\|drop}$	0.000	0.000	0.000	0.000	0.000	1.000

Descriptive statistics of the responses for each time occasion are provided in Table A5. Due to some atypical observations or outliers in the data, as also recently proposed in Taavoni et al. (2020), we considered the natural logarithm of the markers. Figure A2 shows the observed values (in logarithm) for each patient. We observe that the majority of patients had a record for each variable at the first three visits with the exception of the response referred to *cholesterol*. The number of patients with recorded values after the 11th visit decreased very fast. Patients made on average 6.2 visits, with an average variation around the mean of 3.8 visits. We observe that *bilirubin* was recorded on all patients at the baseline except one, and after the fourth visit, it was recorded on only 7% of patients; *cholesterol* was recorded on 90% of the patients at the baseline and on only 3% of the patients (on average) at the subsequent visits.

Concerning the available covariates referred to the drug use, gender, and age, we notice that females were 88% and the patient treated with D-penicillamine were 51%. The median age is 50 years, and the minimum and the maximum age are 26.5 and 78.5, respectively.

A.3 | Additional results of the application

In the following, we provide additional results of the application of the proposed methodology as illustrated in Section 6.2.

The computing time required by the model selection strategy obviously depends on the complexity of the data at hand. On a standard personal computer, the overall fitting procedure, based on $10 \times (k - 1)$ random starting values, may require up to several hours to reach the convergence. Note also that the adopted method, which consists in selecting the number of latent states for a simplified model without covariates and then introducing the covariates, has a reduced computational burden with respect to directly introducing these covariates. This strategy is quite common when estimating discrete

TABLE A8 Averaged initial and transition probabilities with respect to gender: Female (upper panel) or male (bottom panel).

	u					
	1	**2**	**3**	**4**	**5**	***drop***
$\hat{\pi}_u$	**0.205**	**0.440**	**0.109**	**0.198**	**0.048**	**0.000**
$\hat{\pi}_{u\|1}$	0.993	0.000	0.004	0.000	0.002	0.001
$\hat{\pi}_{u\|2}$	0.040	0.887	0.048	0.021	0.000	0.004
$\hat{\pi}_{u\|3}$	0.000	0.000	0.927	0.000	0.068	0.005
$\hat{\pi}_{u\|4}$	0.012	0.002	0.009	0.867	0.086	0.024
$\hat{\pi}_{u\|5}$	0.062	0.000	0.017	0.013	0.626	0.282
$\hat{\pi}_{u\|drop}$	0.000	0.000	0.000	0.000	0.000	1.000
	1	**2**	**3**	**4**	**5**	***drop***
$\hat{\pi}_u$	**0.086**	**0.425**	**0.209**	**0.281**	**0.000**	**0.000**
$\hat{\pi}_{u\|1}$	0.507	0.186	0.000	0.000	0.307	0.000
$\hat{\pi}_{u\|2}$	0.055	0.679	0.137	0.022	0.107	0.000
$\hat{\pi}_{u\|3}$	0.000	0.000	0.857	0.039	0.073	0.031
$\hat{\pi}_{u\|4}$	0.000	0.000	0.000	0.799	0.190	0.011
$\hat{\pi}_{u\|5}$	0.000	0.000	0.000	0.023	0.760	0.217
$\hat{\pi}_{u\|drop}$	0.000	0.000	0.000	0.000	0.000	1.000

TABLE A9 Averaged initial and transition probabilities with respect to age: People aged 50 years or older (upper panel) and aged less than 50 years (bottom panel).

	u					
	1	**2**	**3**	**4**	**5**	***drop***
$\hat{\pi}_u$	**0.224**	**0.363**	**0.170**	**0.180**	**0.062**	**0.000**
$\hat{\pi}_{u\|1}$	0.897	0.025	0.002	0.000	0.074	0.001
$\hat{\pi}_{u\|2}$	0.054	0.825	0.076	0.018	0.024	0.003
$\hat{\pi}_{u\|3}$	0.000	0.000	0.901	0.000	0.086	0.013
$\hat{\pi}_{u\|4}$	0.020	0.001	0.000	0.820	0.129	0.030
$\hat{\pi}_{u\|5}$	0.032	0.000	0.000	0.000	0.688	0.280
$\hat{\pi}_{u\|drop}$	0.000	0.000	0.000	0.000	0.000	1.000
	1	**2**	**3**	**4**	**5**	***drop***
$\hat{\pi}_u$	**0.158**	**0.513**	**0.071**	**0.236**	**0.022**	**0.000**
$\hat{\pi}_{u\|1}$	0.977	0.018	0.004	0.000	0.000	0.001
$\hat{\pi}_{u\|2}$	0.029	0.901	0.040	0.024	0.001	0.004
$\hat{\pi}_{u\|3}$	0.000	0.000	0.936	0.009	0.050	0.004
$\hat{\pi}_{u\|4}$	0.001	0.003	0.016	0.899	0.067	0.015
$\hat{\pi}_{u\|5}$	0.078	0.000	0.030	0.028	0.594	0.270
$\hat{\pi}_{u\|drop}$	0.000	0.000	0.000	0.000	0.000	1.000

latent variable models. The fitting procedure based on $5 \times (k-1)$ random starting values, with $k = 5$, when the model is estimated with covariates requires about 10 hours.

We show the estimated conditional means reported in Table A6 in the original scale according to which the states have been ordered.

In the following, we compare the averaged initial and transition probabilities to evaluate disease evolution according to drug use, gender, and individuals with age above or below the median; see Tables A7, A8, and A9, respectively. We observe that treated patients have a slightly lower probability of dropping out. For males, the dropout probability is almost equal to that of females. Still, males have a lower persistence probability in the first state with respect to females since around 31% of males are estimated to move toward the fifth state, which is the worst in terms of disease progression. Older patients are less persistent in each state, except the fifth, than younger patients.

TABLE A10 Standard errors for the estimated effects of the covariates on the initial and transition probabilities of the multivariate HM model with $k = 5$ hidden states obtained with the nonparametric bootstrap with 299 bootstrap samples.

	Initial probabilities			
Effect	$se(\widehat{\beta}_{12})$	$se(\widehat{\beta}_{13})$	$se(\widehat{\beta}_{14})$	$se(\widehat{\beta}_{15})$
Intercept	1.966	2.349	2.208	3.395
Drug	0.398	0.630	0.429	1.336
Female	1.457	1.564	1.746	0.890
Age	0.022	0.029	0.025	0.055
	Transition probabilities			
Effect	$se(\widehat{\gamma}_{12})$	$se(\widehat{\gamma}_{13})$	$se(\widehat{\gamma}_{14})$	$se(\widehat{\gamma}_{15})$	$se(\widehat{\gamma}_{1drop})$
Intercept	21.631	1.878	0.000	36.705	2.988
Drug	5.575	2.506	0.000	8.866	2.022
Female	5.870	1.872	0.000	15.563	1.929
Age	0.358	0.098	0.004	0.606	0.069
Effect	$se(\widehat{\gamma}_{21})$	$se(\widehat{\gamma}_{23})$	$se(\widehat{\gamma}_{24})$	$se(\widehat{\gamma}_{25})$	$se(\widehat{\gamma}_{2drop})$
Intercept	3.878	1.814	4.525	5.374	4.051
Drug	1.279	0.591	0.861	3.208	3.618
Female	2.753	0.692	3.447	1.287	1.594
Age	0.034	0.030	0.048	0.169	0.093
Effect	$se(\widehat{\gamma}_{31})$	$se(\widehat{\gamma}_{32})$	$se(\widehat{\gamma}_{34})$	$se(\widehat{\gamma}_{35})$	$se(\widehat{\gamma}_{3drop})$
Intercept	0.000	0.459	2.346	4.465	8.205
Drug	0.000	0.304	2.340	0.571	3.793
Female	0.000	0.001	0.950	0.799	3.457
Age	0.004	0.026	0.122	0.066	0.113
Effect	$se(\widehat{\gamma}_{41})$	$se(\widehat{\gamma}_{42})$	$se(\widehat{\gamma}_{43})$	$se(\widehat{\gamma}_{45})$	$se(\widehat{\gamma}_{4drop})$
Intercept	15.196	6.365	11.783	1.838	7.922
Drug	4.064	1.500	4.881	1.345	3.492
Female	3.628	3.998	6.371	1.307	3.489
Age	0.299	0.186	0.437	0.025	0.122
Effect	$se(\widehat{\gamma}_{51})$	$se(\widehat{\gamma}_{52})$	$se(\widehat{\gamma}_{53})$	$se(\widehat{\gamma}_{54})$	$se(\widehat{\gamma}_{5drop})$
Intercept	3.511	0.647	3.612	13.279	1.791
Drug	3.534	0.415	4.459	10.440	0.673
Female	2.222	0.651	3.612	6.989	1.268
Age	0.092	0.038	0.180	0.492	0.059

We finally show the standard errors for the estimated parameters referred to the effects of the covariates affecting the initial and transition probabilities of the Markov chain as in Expressions (11) and (12) of the paper. The following standard errors are referred to the estimated coefficients of the HM model with $k = 5$ hidden states reported in Tables 6 and 7. These are obtained (Table A10) by the nonparametric bootstrap based on 299 samples.