



UNIVERSITÀ DEGLI STUDI
DI PERUGIA

ISSN 2385-2275

Working papers of the Department of Economics University of Perugia (IT)

**A joint model for longitudinal
and survival data based
on an AR(1) latent process**

Silvia Bacci

Francesco Bartolucci

Silvia Pandolfi

Working paper No. 14

October 2015

A joint model for longitudinal and survival data based on an AR(1) latent process

Silvia Bacci^{*†} Francesco Bartolucci^{*‡} Silvia Pandolfi^{*§}

October 5, 2015

Abstract

A critical problem in repeated measurement studies is the occurrence of non-ignorable missing observations. A common approach to deal with this problem is joint modeling the longitudinal and survival processes for each individual on the basis of a random effect that is usually assumed to be time constant. We relax this hypothesis by introducing time-varying subject-specific random effects that follow a first-order autoregressive process, AR(1). We also adopt a generalized linear model formulation to accommodate for different types of longitudinal response (i.e., continuous, binary, count) and we consider some extended cases, such as counts with excess of zeros and multivariate outcomes at each time occasion. Estimation of the parameters of the resulting joint model is based on maximization of the likelihood computed by a recursion developed in the hidden Markov literature. The maximization is performed on the basis of a quasi-Newton algorithm that also provides the information matrix and then standard errors for the parameter estimates. The proposed approach is illustrated through a Monte Carlo simulation study and through the analysis of certain medical datasets.

KEYWORDS: GENERALIZED LINEAR MODELS; INFORMATIVE DROPOUT; NONIGNORABLE MISSING MECHANISM; SEQUENTIAL QUADRATURE; SHARED-PARAMETER MODELS

^{*}Department of Economics, University of Perugia, Via A. Pascoli, 20, 06123 Perugia.

[†]*email:* silvia.bacci@unipg.it

[‡]*email:* francesco.bartolucci@unipg.it

[§]*email:* silvia.pandolfi@unipg.it

1 Introduction

A relevant problem in the analysis of longitudinal data is due to missing observations, in particular when the missing mechanism is nonignorable (Missing Not At Random, MNAR; Little and Rubin, 2002). In the statistical literature there exist different approaches to model an MNAR mechanism. Among the best known, we recall the *selection approach* (Diggle and Kenward, 1994), in which a model is specified for the marginal distribution of the complete data and the conditional distribution of the missing indicators, given these data. On the other hand, according to the *pattern-mixture approach* (Little, 1993), a model is formulated for the marginal distribution of the missing indicators and the conditional distribution of the complete data, given these indicators.

Here we focus on the *shared-parameter approach* for longitudinal data (Wu and Carroll, 1988; Wu and Bailey, 1988, 1989; Follmann and Wu, 1995; Hogan and Laird, 1997, 1998), which introduces random effects to capture the association between the sequence of measurements and the missing process. An example of shared-parameter approach is represented by Joint Models (JMs; Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Tsiatis and Davidian, 2004; Rizopoulos, 2012).

In the standard formulation, a JM is characterized by a Generalized Linear Mixed Model (GLMM; McCulloch and Searle, 2001) for the longitudinal process, with normally distributed random effects, and by a proportional hazard Cox’s model (Cox, 2007) for the survival process, where the risk of the event of interest at a given time depends on the expected value of the longitudinal response at the same time. Such an event typically corresponds to the death of a patient.

Potentially, the misspecification of the standard assumption of normality of the random effects may be problematic in JMs, as these effects are used to capture both the correlation between the repeated measurements in the longitudinal process and the association between the longitudinal and the survival process; besides, the nonrandom dropout caused by the occurrence of the event complicates matters. However, simulation studies found that parameter estimates and standard errors have a certain degree of robustness with respect to misspecification (Rizopoulos et al., 2008; Huang et al., 2009). In this

regard, Song et al. (2002) propose a flexible semi-parametric approach based on a class of smooth densities for the random effects.

More problematic is the assumption of time constancy of the subject-specific random effects. This assumption may be restrictive; relaxing such an assumption by allowing subject-specific random effects to be time-varying amounts to introduce a latent process. For instance, Henderson et al. (2000) and Xu and Zeger (2001) introduce a latent Gaussian process shared by both the longitudinal and the survival processes, whereas Taylor (1994), Lavalley and DeGruttula (1996), and Wang and Taylor (2001) specify an integrated Ornstein-Uhlenbeck process, which is the continuous time analogue of a discrete-time first-order autoregressive, AR(1), process. These approaches are limited to the case of continuous longitudinal responses. The only exception is Xu and Zeger (2001) that formulate a model in terms of a Generalized Linear Model (GLM), but the implementation of the estimation algorithm is still limited to the case of a continuous response. Overall, the implementation of the mentioned model formulations, in which repeated measurement and time-to-event data are jointly modeled with time-varying random effects, may lead to computational difficulties. Recent proposals that try to compound the computational difficulties and the assumption of time-varying random effects come from Barrett et al. (2015) and Bartolucci and Farcomeni (2015), which are illustrated in more detail in the following.

Barrett et al. (2015) propose an approach based on the discretization of the time-to-event measurements, which can be made arbitrarily fine, so as to reach a good compromise between a feasible computational effort and a good approximation of its continuous time limit. For the survival process, Barrett et al. (2015) adopt a sequential probit approach, instead of a Cox's model, so as to model the probability of surviving a time interval conditional on having survived all the previous intervals. They also consider a continuous timescale for the repeated measurements and they adopt a linear mixed Gaussian model for the longitudinal process. This approach allows for different specifications of the random-effects structure, which include models based on a random intercept only, a random intercept and slope, a stationary Gaussian process, and a combination of the last with random intercept and slope. A drawback of this approach is that its use is limited

to normally distributed longitudinal outcomes as maximum likelihood estimation is based on results specific of this distribution and based on the extended skew normal family of distributions (Azzalini, 2005).

A discrete time-to-event history approach for the nonignorable missing process is also adopted by Bartolucci and Farcomeni (2015). However, their proposal differs from Barrett et al. (2015) in the way the assumption of time constancy of the random effects is relaxed. In particular, they consider random effects that follow a first-order Markov chain with a finite number of states. In addition, the authors also account for a time-constant unobserved heterogeneity represented by a latent variable having a discrete distribution. Moreover, a generalized linear mixed effects parameterization is adopted so that multivariate longitudinal data of different types are allowed. A drawback of this approach is the instability of parameter estimates due to the presence of local maxima of the log-likelihood function which is typical of complex statistical models based on latent variables having a discrete distribution. Moreover, this approach requires the choice of the number of latent states in a suitable way.

In this article, we generalize the approach of Barrett et al. (2015) based on a stationary Gaussian process to different longitudinal outcomes, by adopting a GLM formulation. In such a way, different types of longitudinal response may be considered. Besides, we adopt a sequential logit parameterization for the survival process. Moreover, we show that our proposal presents a high level of flexibility, as it may be easily adapted to longitudinal outcomes which are not generated from a distribution belonging to the exponential family (e.g., the Zero-Inflated Poisson), and we also deal with multivariate longitudinal response processes.

Our proposal is based on a latent AR(1) process instead of a discrete one, so that the resulting model is more parsimonious and stable than that of Bartolucci and Farcomeni (2015). As illustrated in the following, we adopt a discretization of the timescale for events, and each repeated observation which is measured in continuous time is considered as lying in a certain “time window”. As exact likelihood inference is no longer applicable for the proposed model, we rely on the quadrature method illustrated in Bartolucci et al. (2014) in a longitudinal context without missing data and on recursions developed in the

hidden Markov literature (Baum et al., 1970) to obtain estimates of model parameters and perform likelihood inference. The main feature of the estimation method is that its complexity increases linearly, rather than exponentially, with the number of time occasions; see also Zucchini and MacDonald (2009) and Bartolucci et al. (2013). In the end, the proposed model combines the main advantages of the models of Barrett et al. (2015) and Bartolucci and Farcomeni (2015), being more general with respect to the first, as longitudinal responses of a different nature may be modeled by a GLM parametrization, and more stable with respect to the second, due to a more regular likelihood function. The proposed estimation method is implemented by means of a set of R functions that we make available to the reader upon request.

The remainder of the paper is organized as follows. In Section 2 we illustrate the proposed JM and its extensions, whereas in Section 3 we develop likelihood inference for this model. Monte Carlo simulations to assess the performance of the proposed method are described in Section 4, whereas in Section 5 we illustrate three applications based on data coming from medical studies in which the longitudinal outcomes of interest are of different type. Some concluding remarks are presented in Section 6.

2 The proposed model

As in the standard formulation (Rizopoulos, 2012), the proposed JM is based on two sub-models, one for the longitudinal process and the other one for the survival time, which are linked through random effects and a specific association parameter accounting for the nonignorable missing mechanism. In the following, we clarify the assumptions of the proposed approach and then we introduce the resulting likelihood function.

2.1 Assumptions

Regarding the longitudinal process, we assume that for every subject i , with $i = 1, \dots, n$, m_i measurements are scheduled at time occasions t_{ij} , $j = 1, \dots, m_i$. We then let $y_{ij} = y_i(t_{ij})$ be the observed response of subject i at time t_{ij} which is seen as a realization of the random variable Y_{ij} ; any type of response is admitted, such as continuous, binary, and count. Let also $\mathbf{x}_{ij} = \mathbf{x}_i(t_{ij})$ be a vector of time-varying covariates, which may include

t_{ij} itself or a function of this quantity. The key-point of the proposed approach is that each observation j for subject i is taken at a time falling in a certain “time window” $s_{ij} = s(t_{ij})$, with $s_{ij} = 1, \dots, v_i$ and $v_i = s(t_{im_i})$. Note that, in general, $v_i \neq m_i$, as an individual may have a varying number of measurements in each time window, but also $v_i > m_i$ is possible when there are “sparse” measurements.

The sub-model for the longitudinal process is formulated in terms of a random intercept GLM, as follows:

$$g(\mu_{ij}) = \alpha_{is_{ij}} + \mathbf{x}'_{ij}\boldsymbol{\beta}, \quad (1)$$

with $g(\cdot)$ denoting a suitable link function and μ_{ij} denoting the conditional expected value of Y_{ij} . Moreover, $\boldsymbol{\beta}$ is the vector of regression coefficients and $\alpha_{is_{ij}}$ is the time-varying and subject-specific random intercept. The sequence of these random intercepts is assumed to depend on time according to the AR(1) process

$$\alpha_{is} = \alpha_{i,s-1}\rho + \eta_{is}\sqrt{1-\rho^2}, \quad s = 1, \dots, v_i,$$

where $\alpha_{i1} = \eta_{i1}$, the error terms η_{is} are independently distributed as $N(0, \sigma_\eta^2)$, and $\rho = \text{cor}(\alpha_{is}, \alpha_{i,s-1})$ is the autocorrelation parameter. A special case of the model is obtained for $\rho = 1$, which corresponds to the situation of time-constant subject-specific random effects; in fact, $\rho = 1$ implies that $\alpha_{is} = \alpha_{i,s-1}$ with probability 1 for $s = 2, \dots, v_i$.

Equation (1) specifies different types of GLM according to the chosen link function $g(\cdot)$. In the case of continuous responses, $g(\cdot)$ is the identity function, so this equation may be expressed as

$$Y_{ij} = \alpha_{is_{ij}} + \mathbf{x}'_{ij}\boldsymbol{\beta} + \varepsilon_{ij}, \quad (2)$$

with ε_{ij} being independent error terms with distribution $N(0, \sigma_\varepsilon^2)$, or equivalently

$$\text{E}(Y_{ij}|\alpha_{is_{ij}}, \mathbf{x}_{ij}) = \alpha_{is_{ij}} + \mathbf{x}'_{ij}\boldsymbol{\beta}.$$

In the case of binary responses, the longitudinal sub-model is based on a logit link of the type

$$\log \frac{p(Y_{ij} = 1|\alpha_{is_{ij}}, \mathbf{x}_{ij})}{p(Y_{ij} = 0|\alpha_{is_{ij}}, \mathbf{x}_{ij})} = \alpha_{is_{ij}} + \mathbf{x}'_{ij}\boldsymbol{\beta}, \quad (3)$$

whereas, in the case of count data, the following log-linear formulation results:

$$\log \text{E}(Y_{ij}|\alpha_{is_{ij}}, \mathbf{x}_{ij}) = \alpha_{is_{ij}} + \mathbf{x}'_{ij}\boldsymbol{\beta}. \quad (4)$$

As usual, under the formulation based on equation (3) or (4), the response variables are assumed to follow a Bernoulli or a Poisson distribution, respectively.

Concerning the sub-model for the survival process, and similarly to Barrett et al. (2015), we adopt a sequential logit formulation based on the following assumption for the random variable S_i corresponding to the number of periods that subject i survives:

$$\log \frac{p(S_i > s | S_i \geq s, \alpha_{is}, \mathbf{w}_{is})}{1 - p(S_i > s | S_i \geq s, \alpha_{is}, \mathbf{w}_{is})} = \alpha_{is}\gamma + \mathbf{w}_{is}'\boldsymbol{\delta}, \quad s = 1, \dots, v_i - 1, \quad (5)$$

where the conditioning argument $S_i \geq s$ vanishes for $s = 0$. In the previous expression, \mathbf{w}_{is} denotes the vector of covariates that are operative at time s and whose effect on the survival is measured by the regression coefficients in vector $\boldsymbol{\delta}$; covariates in \mathbf{w}_{is} may be the same as those in \mathbf{x}_{ij} , for $s = s(t_{ij})$. Finally, parameter γ provides a measurement of association between the longitudinal and the survival process. The case $\gamma = 0$ corresponds to a model incorporating the assumption of ignorable missingness or Missing At Random (MAR) data (Little and Rubin, 2002).

In summary, model based on assumptions (1) and (5) generalizes, to generic (i.e., continuous, binary, count) longitudinal outcomes, the proposal of Barrett et al. (2015) based on a stationary Gaussian process. Moreover, as in that approach, dropout implies that the scheduled observations y_{ij} are missing for each j such that $s(t_{ij}) > s_i$, where s_i denotes the value assumed by S_i . The number of available longitudinal observations is then denoted by $j_i \leq m_i$. It is also worth noting that, although we adopt a discretization of the timescale for event, we can also obtain survival curves representing the probability of survival of a group of individuals in each $s(t_{ij})$, as illustrated in detail in Section 5.3. Finally, it is important to remind that s_i may be censored so that we introduce the indicator variable d_i for the final status of subject i . In a typical medical application, d_i is equal to 1 if subject i is alive at the end of period s_i (censored data) and to 0 otherwise (uncensored data).

2.2 Extended models

We now show how the proposed model, relying on a GLM framework, may be extended to account for different types of longitudinal outcome, which are not generated from

a distribution belonging to the exponential family. Motivated by specific applications, which will be illustrated in Sections 5.2 and 5.3, we focus in particular on the Zero-Inflated Poisson (ZIP) distribution and we illustrate the extension of the proposed approach to deal with multivariate longitudinal data. In the latter case, more outcomes, and of a different nature, can be considered at each period of observation. It is also worth noting that the proposed approach may be easily extended to longitudinal outcomes of a different nature, such as categorical or ordinal, using a parametrization of the type adopted by Bartolucci et al. (2014) for the longitudinal sub-model.

2.2.1 Longitudinal zero-inflated count outcomes

For longitudinal count data, the problem of the excess of zero values often occurs in medical and sociological applications. The ZIP model (Lambert, 1992) can be effectively used to deal with this problem. The model assumes that data come from a mixture of a regular count distribution, such as the Poisson distribution, and a degenerate distribution at zero:

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability } \tau, \\ \text{Poisson}(\lambda_{ij}) & \text{with probability } 1 - \tau. \end{cases}$$

The proposed model formulation described in Section 2.1 may be extended to account for excessive zeros by allowing the longitudinal process to follow the previous ZIP distribution, instead of adopting a GLM parameterization, as follows:

$$\begin{aligned} p(Y_{ij} = 0 | \alpha_{is_{ij}}, \mathbf{x}_{ij}) &= \tau + (1 - \tau)e^{-\lambda_{ij}}, \\ p(Y_{ij} = y | \alpha_{is_{ij}}, \mathbf{x}_{ij}) &= (1 - \tau) \frac{\lambda_{ij}^y e^{-\lambda_{ij}}}{y!}, \quad y > 0, \end{aligned} \tag{6}$$

where

$$\lambda_{ij} = \exp(\alpha_{is_{ij}} + \mathbf{x}_{ij}'\boldsymbol{\beta}).$$

Concerning the sub-model for the survival process, we again adopt a sequential logit formulation as in (5). Also note that the probability τ could be modeled so that it depends on individual covariates by means of a logit parameterization; see, among others, Min and Agresti (2005).

2.2.2 Multivariate longitudinal outcomes

In order to extend the proposed approach to multivariate longitudinal outcomes, we first note that a formulation equivalent to that in Section 2.1 may be based on a standardized AR(1) process based on the assumption

$$\alpha_{is}^* = \alpha_{i,s-1}^* \rho + \eta_{is}^* \sqrt{1 - \rho^2}, \quad s = 1, \dots, v_i, \quad (7)$$

with $\alpha_{i1}^* = \eta_{i1}^*$ and $\eta_{is}^* \sim N(0, 1)$. In this case, assumption (1) is replaced by

$$g(\mu_{ij}) = \alpha_{is_{ij}}^* \phi + \mathbf{x}_{ij}' \boldsymbol{\beta} \quad (8)$$

and assumption (5) is replaced by

$$\log \frac{p(S_i > s | S_i \geq s, \alpha_{is}, \mathbf{w}_{is})}{1 - p(S_i > s | S_i \geq s, \alpha_{is}, \mathbf{w}_{is})} = \alpha_{is}^* \psi + \mathbf{w}_{is}' \boldsymbol{\delta}. \quad (9)$$

In practice, parameter ϕ in (8) corresponds to the square root of σ_η^2 , which is the stationary variance of the latent AR(1) in the original formulation, whereas parameter ψ in (9) corresponds to the product between this square root and parameter γ used in the initial survival model (5).

In the multivariate case, we observe the response variables Y_{hij} , $h = 1, \dots, r$, for each unit i and occasion j . These variables may be of a different nature; for instance, in Section 5.3 we deal with a case of two response variables, the first of which is continuous and the second is binary. To deal with this case, we extend the above formulation based on a standardized AR(1) process by assuming that

$$g_h(\mu_{hij}) = \alpha_{is_{ij}}^* \phi_h + \mathbf{x}_{ij}' \boldsymbol{\beta}_h, \quad h = 1, \dots, r, \quad (10)$$

with $g_h(\cdot)$ denoting a suitable link function and μ_{hij} denoting the conditional expected value of Y_{hij} . Moreover, $\boldsymbol{\beta}_h$ is the vector of regression coefficients for the h -th response variable, and ϕ_h measures the association between the longitudinal sub-model referred to variable Y_{hij} and the random effects. Finally, assumption (9) is retained, exactly in the same form, for the survival process.

It is worth noting that a more flexible formulation could be based on adopting a specific latent AR(1) process for each longitudinal response sequence, allowing all these processes

to affect the survival time through the use of specific coefficients. We considered also this extension but we verified that, at least using the maximum likelihood inference tools that will be illustrated in the following, it is impossible to deal with more than two response variables at the same time and even for the bivariate case there are severe numerical problems. Consequently, in this paper we focus on the formulation based on a single latent AR(1) process which is illustrated above. This process has the role of summarizing all unobservable factors that affect the response variables and, at the same time, the survival process. As shown in Section 5.3, the resulting model is competitive in terms of goodness-of-fit with less parsimonious models, while having a simpler interpretation and providing more stable results in terms of parameter estimates.

2.3 Model likelihood

Assuming independence between the n sample units, the model likelihood function has components corresponding to the manifest distribution $p(\mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i)$, for $i = 1, \dots, n$, where $\mathbf{y}_i = (y_{i1}, \dots, y_{ij_i})'$ is the observed vector of longitudinal responses and \mathbf{X}_i and \mathbf{W}_i are matrices of covariates with columns \mathbf{x}_{ij} and \mathbf{w}_{is} , respectively, for $j = 1, \dots, j_i$ and $s = 1, \dots, s_i$. The model log-likelihood is then

$$\ell(\boldsymbol{\theta}) = \sum_i \log p(\mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i), \quad (11)$$

where $\boldsymbol{\theta}$ is the vector of all parameters.

The previous distribution is defined as

$$p(\mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i) = \int p(\mathbf{y}_i, s_i, d_i | \boldsymbol{\alpha}_i, \mathbf{X}_i, \mathbf{W}_i) f(\boldsymbol{\alpha}_i) d\boldsymbol{\alpha}_i, \quad (12)$$

based on suitably marginalizing out $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{is_i})'$ from

$$p(\mathbf{y}_i, s_i, d_i | \boldsymbol{\alpha}_i, \mathbf{X}_i, \mathbf{W}_i) = p(\mathbf{y}_i | \boldsymbol{\alpha}_i, \mathbf{X}_i) p(s_i, d_i | \boldsymbol{\alpha}_i, \mathbf{W}_i).$$

The distributions involved in the previous expression are defined as follows:

$$\begin{aligned} p(\mathbf{y}_i | \boldsymbol{\alpha}_i, \mathbf{X}_i) &= \prod_{j=1}^{j_i} p(y_{ij} | \alpha_{is_{ij}}, \mathbf{x}_{ij}), \\ p(s_i, d_i | \boldsymbol{\alpha}_i, \mathbf{W}_i) &= \prod_{s=1}^{s_i-1} p(S_i > s | S_i \geq s, \alpha_{is}, \mathbf{w}_{is}) \\ &\times p(S_i > s_i | S_i \geq s_i, \alpha_{is_i}, \mathbf{w}_{is_i})^{d_i} p(S_i = s_i | S_i \geq s_i, \alpha_{is_i}, \mathbf{w}_{is_i})^{1-d_i}, \end{aligned}$$

with $p(y_{ij}|\alpha_{is_{ij}}, \mathbf{x}_{ij})$ defined according to the assumed model for the longitudinal responses, see equation (1), and the probabilities referred to S_i defined according to (5). The above expressions may be simply adapted to deal with the multivariate formulation illustrated in Section 2.2.2, considering that at each time occasion we observe a vector of outcomes rather than a single outcome.

In general, an explicit expression for the s_i -dimensional integral in (12) is not available. In the following, we propose a method to solve this integral and then performing maximum likelihood estimation of the model parameters.

3 Proposed estimation approach

In order to compute the likelihood function of any model in the proposed class, we rely on a quadrature method based on an equally spaced grid of points and on a recursion developed in the hidden Markov literature (see Baum et al., 1970); see also Heiss (2008) for a related sequential quadrature method and Bartolucci et al. (2014) for a related application to the estimation of models for longitudinal data.

3.1 Sequential quadrature method

The proposed method to compute the integral in (12) is based on expressing this integral as

$$p(\mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i) = \int q_{s_i}(\alpha_{is_i}, \mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i) d\alpha_{is_i},$$

where

$$\begin{aligned} q_1(\alpha_{i1}, \mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i) &= f(\alpha_{i1}) \left[\prod_{j:s(t_{ij})=1} p(y_{ij} | \alpha_{i1}, \mathbf{x}_{ij}) \right] \\ &\times \begin{cases} p(S_i > 1 | \alpha_{i1}, \mathbf{w}_{i1})^{d_i} p(S_i = 1 | \alpha_{i1}, \mathbf{w}_{i1})^{1-d_i} & \text{if } s_i = 1, \\ p(S_i > 1 | \alpha_{i1}, \mathbf{w}_{i1}) & \text{if } s_i > 1, \end{cases} \end{aligned}$$

and

$$\begin{aligned} q_s(\alpha_{is}, \mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i) &= \int q_{s-1}(\alpha_{i,s-1}, \mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i) f(\alpha_{is} | \alpha_{i,s-1}) d\alpha_{i,s-1} \left[\prod_{j:s(t_{ij})=s} p(y_{ij} | \alpha_{is}, \mathbf{x}_{ij}) \right] \\ &\times \begin{cases} p(S_i > s | S_i \geq s, \alpha_{is}, \mathbf{w}_{is})^{d_i} p(S_i = s | S_i \geq s, \alpha_{is}, \mathbf{w}_{is})^{1-d_i} & \text{if } s_i = s, \\ p(S_i > s | S_i \geq s, \alpha_{is}, \mathbf{w}_{is}) & \text{if } s_i > s, \end{cases} \end{aligned}$$

for $s = 2, \dots, s_i$.

In the above expression, $f(\alpha_{i1})$ is the density of a normal distribution centered on zero and with variance σ_η^2 , whereas $f(\alpha_{is}|\alpha_{i,s-1})$ is the density of the conditional distribution $N(\alpha_{i,s-1}\rho, (1-\rho^2)\sigma_\eta^2)$. Moreover, we compute the above integrals on the basis of uniform quadrature method based on a set of k nodes, denoted by a_1, \dots, a_k , in a certain interval. In our applications, we initially use $k = 51$ points in the grid from -5 to 5. The weights corresponding to the first time occasion are obtained as

$$\nu_u = \frac{f(a_u)}{\sum_{l=1}^k f(a_l)}, \quad u = 1, \dots, k, \quad (13)$$

whereas the weights for the following time windows are obtained as

$$\pi_{\bar{u}u} = \frac{f(a_u|a_{\bar{u}})}{\sum_{l=1}^k f(a_l|a_{\bar{u}})}, \quad \bar{u}, u = 1, \dots, k. \quad (14)$$

In practice, the proposed approach to obtain the model likelihood amounts to perform the following recursion for every sample unit $i = 1, \dots, n$:

1. compute

$$\begin{aligned} q_{i1u} &= \nu_u \left[\prod_{j:s(t_{ij})=1} p(y_{ij}|\alpha_{i1} = a_u, \mathbf{x}_{ij}) \right] \\ &\times \begin{cases} p(S_i > 1|\alpha_{i1} = a_u, \mathbf{w}_{is})^{d_i} p(S_i = 1|\alpha_{i1} = a_u, \mathbf{w}_{is})^{1-d_i} & \text{if } s_i = 1, \\ p(S_i > 1|\alpha_{i1} = a_u, \mathbf{w}_{is}) & \text{if } s_i > 1; \end{cases} \end{aligned}$$

2. compute

$$\begin{aligned} q_{isu} &= \left[\sum_{\bar{u}=1}^k q_{i,s-1,\bar{u}} \pi_{\bar{u}u} \right] \left[\prod_{j:s(t_{ij})=s} p(y_{ij}|\alpha_{is} = a_u, \mathbf{x}_{ij}) \right] \\ &\times \begin{cases} p(S_i > s|S_i \geq s, \alpha_{is} = a_u, \mathbf{w}_{is})^{d_i} p(S_i = s|S_i \geq s, \alpha_{is} = a_u, \mathbf{w}_{is})^{1-d_i} & \text{if } s_i = s, \\ p(S_i > s|S_i \geq s, \alpha_{is} = a_u, \mathbf{w}_{is}) & \text{if } s_i > s, \end{cases} \end{aligned}$$

for $u = 1, \dots, k$ and $s = 2, \dots, s_i$.

3. obtain the manifest distribution of \mathbf{y}_i , s_i , and d_i as

$$p(\mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i) = \sum_{u=1}^k q_{is_i u}.$$

On the basis of the recursion defined above, we obtain the log-likelihood function $\ell(\boldsymbol{\theta})$ as defined in (11) with a computational complexity that linearly increases with the number of time occasions. How to use this function for likelihood inference on the model parameters is clarified in the following. For the moment, it is worth noting that the proposed quadrature method amounts to approximate the latent variable distribution, which is continuous, with a discrete distribution. In other terms, we are approximating the proposed model by a hidden Markov model (Zucchini and MacDonald, 2009; Bartolucci et al., 2013) based on initial probabilities ν_u and transition probabilities $\pi_{\bar{u}u}$ which depend on the parameters σ_η^2 and ρ through definitions (13) and (14), respectively, and we are using the forward recursion for computing these models proposed by Baum et al. (1970); see also Welch (2003). Note that the sequential quadrature approach proposed by Heiss (2008) for dealing with simpler models based on an AR(1) latent process is similar, but the integral at each step are computed in a different way. As we experimented, the recursion we proposed here is more stable as it relies on normalized weights which sum up to 1, as it is already clarified in Bartolucci et al. (2014).

3.2 Likelihood inference

We use the log-likelihood function $\ell(\boldsymbol{\theta})$ computed as above to estimate the model parameters collected in $\boldsymbol{\theta}$. For this aim we use a numerical maximizer of quasi-Newton type and, to make the maximization faster, we also provide the maximizer with the score function, which is equal to

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^n \frac{1}{p(\mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i)} \frac{\partial p(\mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i)}{\partial \boldsymbol{\theta}}.$$

The derivative of $p(\mathbf{y}_i, s_i, d_i | \mathbf{X}_i, \mathbf{W}_i)$, used in the expression above, is computed by a recursive method that follows the same steps of the recursion illustrated above and recalls that introduced by Lystig and Hughes (2002). The quasi-Newton maximizer also provides the observed information matrix, computed as minus the numerical derivative of the score vector; this is used to compute the standard errors for the parameter estimates in the usual way.

Regarding the initialization of the estimation algorithm we use both a determinis-

tic and a random rule, the latter repeated a given number of times, so as to check for the presence of local maxima. More in detail, the estimation strategy based on a deterministic initialization of the model parameters consists in adopting as starting values the coefficients obtained by estimating separate naive models on the same data: a GLM (without random effects) for the longitudinal outcome, where the link function is specified in a suitable way according to the nature of the response variables, and a binary logit model for the individual status. In order to obtain random starting rules, these coefficients are perturbed by adding random values generated from a normal distribution with mean zero and standard deviation equal to the corresponding standard error. Under both initialization rules, to balance the computational time and the precision of estimates at convergence, the maximization process is performed in two steps: we first use $k = 51$ nodes until convergence and, then, we run again the maximization algorithm with $k = 101$ nodes using, as starting values, the estimates obtained at the previous step. We make our R implementation available to the reader upon request.

It is worth recalling that the method of Barrett et al. (2015), based on exact likelihood inference, is applicable only to longitudinal outcomes belonging to the family of skew normal distributions; their approach also requires a probit parameterization for the survival process. On the other hand, our estimation method, based on a sequential quadrature, allows for longitudinal outcomes of a different nature, including the normal case as a special one. A comparison between the two approaches is performed in Section 5.1 on the basis of a dataset with a continuous response.

4 Simulation study

In the following, we illustrate a Monte Carlo simulation study aimed at assessing the performance of the proposed method for different types of longitudinal outcome.

4.1 Design

We implemented a simulation design recalling that of Barrett et al. (2015). In particular, we simulated univariate longitudinal data with dropout for $n = 1000, 2000$ individuals and different longitudinal sub-models (with continuous, binary, and count outcomes). In

this design, the longitudinal measurements are randomly distributed over $v_i = 5, 10$ time windows, with a maximum of $m_i = 10, 20$ repeated measurements per individual. It is important stressing that each individual may have a varying number of visits in each time windows. A uniform distribution is adopted for the visit time and the dropout may occur during any time interval.

The simulation design considers continuous, binary, and count response variables. Moreover, following Barrett et al. (2015), it considers two covariates, in both longitudinal and survival models, supposed to be: *age*, with initial values generated uniformly from the interval $(10, 30)$, and a binary covariate (*sex*), which assumes value 1 with probability 0.5. Two different true values are considered for the association parameter, $\gamma = 0.05, 0.5$, and for the autocorrelation coefficient, $\rho = 0.7, 0.9$. For the variance of the random effects, it is assumed $\sigma_\eta^2 = 1, 4$, in the case of continuous and binary data, and $\sigma_\eta^2 = 0.25, 1$, for count data. In this case, $\sigma_\eta^2 = 1$ is a value quite large for the variance since the mean of the outcome variable is an exponential function of the combination of the fixed and the random effects (see Xu et al., 2007, for a similar setting). Finally, for continuous outcomes, we have $\sigma_\varepsilon^2 = 1, 4$, with σ_ε^2 denoting the variance of the repeated measures.

Overall, we considered a total of 22 different scenarios. Here, we focus on the results concerning 12 scenarios (i.e., four scenarios for each type of longitudinal outcome); the results of the remaining scenarios are included in the Appendix.

The first type of scenario (Scenario 1) assumes a parameter setting similar to that considered in Barrett et al. (2015). More in detail, we have $n = 1000$, $v_i = 5$, and a maximum of 10 repeated observations per individual. For the survival process, we assume $\boldsymbol{\delta} = (2, 0.01, 0.01, 0.1)'$, and $\gamma = 0.05$. The variance of the random effects is assumed to be $\sigma_\eta^2 = 1$ for continuous and binary data and $\sigma_\eta^2 = 0.25$ for count data, whereas the autocorrelation coefficient is set equal to $\rho = 0.7$. The parameters for the longitudinal process are set equal to $\boldsymbol{\beta} = (90, -1.7, -1.7, 2)'$ for continuous data, to $\boldsymbol{\beta} = (4, -0.2, -0.2, 2)'$ for binary data, and to $\boldsymbol{\beta} = (2, -0.02, -0.02, 0.5)'$ for count data. Moreover, for continuous outcomes, the first scenario considers a variance of the repeated measure equal to $\sigma_\varepsilon^2 = 1$. In such a context of continuous data, given the equality on the parameters setting, we can consider the results of Barrett et al. (2015), denoted in the

following as *SGP II*, as benchmark design. The second type of scenario (Scenario 2) is based on a larger variance of the random effects, with $\sigma_\eta^2 = 4$ for continuous and binary data and $\sigma_\eta^2 = 1$ for count data, whereas Scenario 3 is based on a larger number of time windows, $v_i = 10$. The last type of scenario (Scenario 4) evaluates the effect of an increase in the number of individuals, with $n = 2000$.

4.2 Results

Tables 1, 2, and 3 report the estimation results obtained under the four different scenarios described above, with reference to continuous, binary and count outcomes, respectively, and based on 500 datasets generated under each scenario. The performance of the proposed estimation method is evaluated in terms of mean, standard deviations (sd) and root mean square errors (RMSE) of the considered estimator. We also report the averaged estimated standard error (se) obtained on the basis of the observed information matrix, as described at the end of Section 3.2.

The results in Table 1, referred to continuous data, lead us to conclude that, under all scenarios, the means of the parameter estimates are close to the true values.

Scenario 1 confirms the *SGP II* results of Barrett et al. (2015). This allows us to validate the proposed estimation method. Under Scenario 2, we observe larger standard errors for the longitudinal parameters. Even the standard errors and the RMSE of σ_η^2 are larger than those obtained under Scenario 1. On the other hand, the standard error of the estimator of γ is smaller. Under Scenario 3, characterized by a larger number of time windows with respect to the other scenarios, we observe an improving of the behavior of the estimators of the survival model parameters, whereas the standard errors and RMSE of the estimators of σ_ϵ^2 , σ_η^2 and ρ slightly increase. Finally, as n increases (Scenario 4), the behavior of the estimators improves for all parameters. In all scenarios, the estimated standard errors obtained by the observed information matrix are in agreement with the Monte Carlo standard deviation.

Even in the context of binary data (Table 2), under Scenario 1, all parameters are sharply estimated by the proposed method, with a mean of the estimates close to the true values. Scenario 2 confirms the behavior of the estimators observed in the context

Parameters	True	Scenario 1 (benchmark)				True	Scenario 2 $\sigma^2_\eta = 4$			
		Mean	sd	se	RMSE		Mean	sd	se	RMSE
Longitudinal										
Intercept	90.000	90.003	0.122	0.127	0.122	90.000	90.015	0.227	0.225	0.228
Time	-1.700	-1.701	0.016	0.017	0.016	-1.700	-1.699	0.028	0.026	0.028
Age at t_0	-1.700	-1.700	0.005	0.006	0.005	-1.700	-1.701	0.010	0.010	0.010
Sex	2.000	2.000	0.070	0.065	0.070	2.000	1.998	0.117	0.116	0.118
Survival										
Intercept	2.000	1.985	0.220	0.224	0.220	2.000	1.996	0.227	0.224	0.227
Time	0.010	0.008	0.037	0.038	0.037	0.010	0.008	0.039	0.038	0.039
Age at t_0	0.010	0.011	0.009	0.009	0.009	0.010	0.010	0.009	0.009	0.009
Sex	0.100	0.102	0.109	0.107	0.109	0.100	0.099	0.103	0.107	0.103
γ	0.050	0.050	0.080	0.079	0.080	0.050	0.051	0.035	0.034	0.035
Others										
σ^2_ξ	1.000	1.000	0.031	0.031	0.031	1.000	1.001	0.032	0.032	0.032
σ^2_η	1.000	0.998	0.055	0.054	0.055	4.000	3.990	0.162	0.158	0.162
ρ	0.700	0.700	0.028	0.028	0.028	0.700	0.698	0.017	0.017	0.018
Scenario 3 $v_i = 10$										
Parameters	True	Mean	sd	se	RMSE	True	Mean	sd	se	RMSE
Longitudinal										
Intercept	90.000	89.994	0.130	0.127	0.130	90.000	90.000	0.090	0.090	0.090
Time	-1.700	-1.700	0.010	0.010	0.010	-1.700	-1.700	0.012	0.012	0.012
Age at t_0	-1.700	-1.700	0.006	0.006	0.006	-1.700	-1.700	0.004	0.004	0.004
Sex	2.000	1.999	0.064	0.064	0.064	2.000	1.996	0.047	0.046	0.047
Survival										
Intercept	2.000	1.978	0.164	0.171	0.165	2.000	2.001	0.159	0.158	0.159
Time	0.010	0.010	0.015	0.015	0.015	0.010	0.007	0.028	0.027	0.028
Age at t_0	0.010	0.010	0.007	0.007	0.007	0.010	0.010	0.006	0.007	0.006
Sex	0.100	0.103	0.084	0.085	0.084	0.100	0.105	0.076	0.076	0.076
γ	0.050	0.052	0.075	0.076	0.075	0.050	0.054	0.057	0.055	0.057
Others										
σ^2_ξ	1.000	1.002	0.041	0.043	0.041	1.000	0.999	0.022	0.022	0.022
σ^2_η	1.000	0.991	0.060	0.061	0.061	1.000	1.001	0.039	0.038	0.039
ρ	0.700	0.700	0.031	0.030	0.031	0.700	0.699	0.019	0.020	0.019

Table 1: *Mean of the parameter estimates, standard deviation (sd), average estimated standard error (se) and root mean square error (RMSE) for continuous data: Scenarios 1-4.*

of continuous outcomes, with a quite large standard error and RMSE of the estimator of σ_η^2 . The increase in the number of time windows, $v_i = 10$, leads to a reduction of the standard deviations of parameters in δ . However, the behavior of estimators of β and γ remains almost unchanged with respect to Scenario 1. Moreover, the worsening of the performance of the AR(1) parameter estimators is more evident than that observed for continuous data. Scenario 4 shows an improvement of all estimators when the sample size increases from $n = 1000$ to $n = 2000$. As in the context of continuous data, under all scenarios, the propose method produces reliable estimates of the standard error.

The results of the simulations involving longitudinal count data (Table 3) are in agree-

Parameters	True	Scenario 1 (benchmark)				True	Scenario 2 $\sigma_\eta^2 = 4$			
		Mean	sd	se	RMSE		Mean	sd	se	RMSE
Longitudinal										
Intercept	4.000	4.003	0.242	0.232	0.242	4.000	4.015	0.332	0.330	0.332
Time	-0.200	-0.200	0.032	0.032	0.032	-0.200	-0.200	0.041	0.043	0.041
Age at t_0	-0.200	-0.200	0.011	0.011	0.011	-0.200	-0.201	0.015	0.015	0.015
Sex	2.000	2.007	0.117	0.115	0.118	2.000	2.008	0.165	0.166	0.165
Survival										
Intercept	2.000	1.991	0.220	0.224	0.220	2.000	2.008	0.218	0.225	0.218
Time	0.010	0.010	0.039	0.038	0.039	0.010	0.010	0.040	0.038	0.040
Age at t_0	0.010	0.010	0.009	0.009	0.009	0.010	0.010	0.009	0.009	0.009
Sex	0.100	0.098	0.105	0.107	0.105	0.104	0.100	0.110	0.107	0.110
γ	0.050	0.053	0.137	0.135	0.137	0.050	0.053	0.046	0.047	0.046
Others										
σ_η^2	1.000	1.005	0.190	0.198	0.190	4.000	4.023	0.479	0.510	0.479
ρ	0.700	0.693	0.089	0.096	0.090	0.700	0.695	0.041	0.041	0.042
Parameters	True	Scenario 3 $v_i = 10$				True	Scenario 4 $n = 2000$			
		Mean	sd	se	RMSE		Mean	sd	se	RMSE
Longitudinal										
Intercept	4.000	4.010	0.265	0.264	0.265	4.000	4.015	0.166	0.164	0.167
Time	-0.200	-0.199	0.019	0.019	0.019	-0.200	-0.202	0.023	0.022	0.023
Age at t_0	-0.200	-0.201	0.012	0.012	0.012	-0.200	-0.200	0.007	0.007	0.007
Sex	2.000	2.004	0.128	0.131	0.128	2.000	1.997	0.076	0.081	0.076
Survival										
Intercept	2.000	1.989	0.174	0.172	0.174	2.000	2.005	0.162	0.158	0.162
Time	0.010	0.009	0.016	0.015	0.016	0.010	0.008	0.026	0.027	0.026
Age at t_0	0.010	0.011	0.007	0.007	0.007	0.010	0.010	0.007	0.007	0.007
Sex	0.100	0.101	0.084	0.085	0.084	0.100	0.099	0.073	0.076	0.073
γ	0.050	0.058	0.135	0.141	0.135	0.050	0.055	0.100	0.094	0.100
Others										
σ_η^2	1.000	1.028	0.275	0.282	0.276	1.000	0.999	0.135	0.138	0.135
ρ	0.700	0.685	0.103	0.110	0.104	0.700	0.700	0.063	0.065	0.063

Table 2: Mean of the parameter estimates, standard deviation (*sd*), average estimated standard error (*se*) and root mean square error (*RMSE*) for binary data: Scenarios 1-4.

ment with the results already discussed for the previous cases. In summary, the standard deviation and RMSE of the parameter estimates in β are larger when we assume a higher variance of the random effects. On the other hand, the behavior of the estimator of the association parameter γ improves with $\sigma_\eta^2 = 1$. The accuracy of the estimated parameters of the survival process increases when the number of time windows v_i increases. Even in this case, and as expected, a larger value of n leads to better results in terms of accuracy and efficiency of the parameter estimates.

The additional scenarios reported in Appendix allows us to conclude that, in both continuous, binary, and count data, an increase of the maximum number of repeated measurements per individual, m_i , leads to a lower standard error and RMSE of the estimated parameters of the longitudinal process. Moreover, also the standard errors and

Parameters	True	Scenario 1 (benchmark)				True	Scenario 2 $\sigma_\eta^2 = 1$			
		Mean	sd	se	RMSE		Mean	sd	se	RMSE
Longitudinal										
Intercept	2.000	1.999	0.064	0.061	0.064	2.000	1.996	0.104	0.089	0.104
Time	-0.020	-0.020	0.007	0.008	0.007	-0.020	-0.020	0.013	0.011	0.013
Age at t_0	-0.020	-0.020	0.003	0.003	0.003	-0.020	-0.020	0.005	0.004	0.005
Sex	0.500	0.502	0.031	0.031	0.031	0.500	0.506	0.060	0.046	0.061
Survival										
Intercept	2.000	1.998	0.223	0.224	0.223	2.000	1.988	0.242	0.224	0.242
Time	0.010	0.007	0.036	0.038	0.036	0.010	0.011	0.038	0.038	0.038
Age at t_0	0.010	0.010	0.009	0.009	0.009	0.010	0.010	0.010	0.009	0.010
Sex	0.100	0.103	0.106	0.107	0.106	0.100	0.106	0.100	0.107	0.101
γ	0.050	0.048	0.144	0.149	0.144	0.050	0.046	0.068	0.067	0.068
Others										
σ_η^2	0.250	0.250	0.012	0.012	0.012	1.000	1.009	0.039	0.026	0.040
ρ	0.700	0.699	0.021	0.022	0.021	0.700	0.701	0.016	0.015	0.016
Parameters	True	Scenario 3 $v_i = 10$				True	Scenario 4 $n = 2000$			
		Mean	sd	se	RMSE		Mean	sd	se	RMSE
Longitudinal										
Intercept	2.000	2.003	0.061	0.060	0.061	2.000	2.003	0.042	0.043	0.042
Time	-0.020	-0.020	0.005	0.005	0.005	-0.020	-0.021	0.006	0.005	0.006
Age at t_0	-0.020	-0.020	0.003	0.003	0.003	-0.020	-0.020	0.002	0.002	0.002
Sex	0.500	0.500	0.030	0.031	0.030	0.500	0.499	0.022	0.022	0.022
Survival										
Intercept	2.000	1.994	0.176	0.171	0.176	2.000	1.993	0.161	0.158	0.161
Time	0.010	0.010	0.015	0.015	0.015	0.010	0.009	0.028	0.027	0.028
Age at t_0	0.010	0.010	0.008	0.007	0.008	0.010	0.010	0.007	0.007	0.007
Sex	0.100	0.100	0.086	0.085	0.086	0.100	0.103	0.075	0.076	0.075
γ	0.050	0.049	0.145	0.142	0.145	0.050	0.048	0.104	0.105	0.104
Others										
σ_η^2	0.250	0.250	0.012	0.012	0.012	0.250	0.250	0.008	0.008	0.008
ρ	0.700	0.698	0.022	0.023	0.023	0.700	0.700	0.015	0.016	0.015

Table 3: Mean of the parameter estimates, standard deviation (sd), average estimated standard error (se) and root mean square error (RMSE) for count data: Scenarios 1-4.

RMSE of the estimators of σ_η^2 and ρ slightly decreases, in contrast to the results registered in Scenario 3, when the number of time windows increases. Moreover, for continuous data, a larger variance of the repeated measures, $\sigma_\epsilon^2 = 4$, leads to a worsening of the behavior of the estimators of the longitudinal parameters, together with a loss of efficiency of the estimates of σ_η^2 , ρ , and σ_ϵ^2 with respect to the benchmark design. We also note that the association parameter γ and the autocorrelation parameter ρ do not seem to have a substantial influence on the estimation results.

5 Applications

We propose three applications of the class of JMs specified in Sections 2.1 and 2.2 on some medical datasets, characterized by different types of longitudinal outcome: continuous, count with excess of zeros, and bivariate with a normal and a binary response. The first example, concerning a continuous longitudinal response variable, is mainly used to show the validity of our proposal in comparison with that of Barrett et al. (2015). Then, the more general capability of application of the proposed model is illustrated through the second and the third examples. For each of them we provide the estimates of model parameters under the general assumption of MNAR and AR(1) random effects and under the following two special cases: MNAR assumption with time-constant random effects and MAR assumption (with AR(1) random effects).

5.1 Example 1: lung functioning in patients affected by cystic fibrosis

In order to validate our model and the related estimation method, we compare it with the *SGP II* model of Barrett et al. (2015), which is equivalent to our proposal in the case of a normally distributed repeated outcome. We consider a dataset coming from the UK Cystic Fibrosis registry and covering years 2007-2013. The data concern 3,627 patients affected by cystic fibrosis, which is a genetic chronic disease influencing the physiological functioning of lungs, pancreas, liver, kidneys and intestine; usually, patients die as a consequence of respiratory complications. An important biomarker of the lung functioning of a patient is represented by the percentage of predicted forced expiratory volume in 1 second (FEV1), which is usually measured once a year. The total number of FEV1 measurements is 21,578 and ranges from 1 to 7 for each patient with average 3.68 (standard deviation 1.92) and median 4.0. Our aim consists in detecting a possible association between values of FEV1 and survival, other than estimating the time trend, after adjusting for sex and age at baseline. As the longitudinal outcome is normally distributed, the longitudinal sub-model is specified according to (2).

In Table 4 we show the estimation results referred to the *SGP II* model of Barrett et al.

(2015)¹ (columns 2-4) and to the JM under MNAR assumption with AR(1) random effects (columns 5-7). We observe that the proposed estimation method provides parameter estimates and standard errors that strongly resemble those obtained through the exact likelihood based approach. These results may be seen as a further evidence that our proposal leads to the same results of the method of Barrett et al. (2015) while being more general of the latter.

	Model <i>SGP II</i> of Barrett et al. (2015)			MNAR with AR(1) random effects		
	Est.	se	p-value	Est.	se	p-value
<i>Longitudinal sub-model</i>						
Intercept	89.126	0.929	0.000	89.086	0.929	0.000
Time	-1.245	0.053	0.000	-1.232	0.054	0.000
Age at baseline	-0.695	0.033	0.000	-0.696	0.033	0.000
Sex (male)	0.109	0.734	0.882	0.117	0.733	0.873
<i>Survival sub-model</i>						
Intercept	6.508	0.235	0.000	8.179	0.372	0.000
Time	-0.011	0.024	0.656	-0.018	0.033	0.583
Age at baseline	-0.045	0.006	0.000	-0.066	0.008	0.000
Sex (male)	0.164	0.117	0.161	0.234	0.161	0.146
$\hat{\gamma}$	0.056	0.004	0.000	0.085	0.005	0.000
<i>Others</i>						
$\hat{\sigma}_\varepsilon^2$	37.661	1.227	0.000	36.760	1.125	0.000
$\hat{\sigma}_\eta^2$	534.883	22.434	0.000	535.018	11.556	0.000
$\hat{\rho}$	0.951	0.015	0.000	0.948	0.002	0.000

Table 4: *Parameter estimates for the value of FEV1, under the SGP II model of Barrett et al. (2015) and under the assumption of nonignorable dropout (MNAR) with AR(1) random effects.*

On the basis of both approaches we conclude for a significant negative effect of age at baseline on FEV1 and survival and a positive, although not significant, effect on FEV1 and survival of males with respect to females. Moreover, the estimated value of parameter γ denotes a positive association between the lung functioning and the probability of survival.

Finally, it is worth noting that the proposed estimation algorithm is highly stable. In particular, in order to assess the convergence to the global maximum of the model log-likelihood, we performed a series of 100 random initializations of the proposed estimation algorithm, following the strategy described in Section 3.2. We observed that the log-likelihood at convergence equals the best solution 81 times out of 100.

¹ The R functions for the estimation of the *SGP II* model are available at <http://wileyonlinelibrary.com/journal/rss-datasets>.

5.2 Example 2: effect of β -carotene in prevention of nonmelanoma skin cancer

We analyze the effect of β -carotene for the prevention of nonmelanoma skin cancer in high-risk individuals. The dataset used in this application comes from a randomized study (for details see Greenberg et al., 1990) involving 1,683 individuals randomized to placebo or to a treatment based on β -carotene for a period of 5 years. The response variable of interest is the count of yearly new skin cancers. The total number of observations is 7,081 with a strong prevalence of zeros (83.55%). As the longitudinal outcome is well described by a ZIP variable, the longitudinal sub-model is specified as in equation (6).

We observe that the number of counts ranges from 1 to 5 for each individual with just 811 patients (48.19%) having a complete pattern of five measurements. Therefore, it is reasonable to account for the possible association between the longitudinal response process and the dropout process, where the status of an individual is equal to 1 if he/she is in the follow up at a given time and to 0 if the individual drops out from the follow up (and, then, the count of yearly new skin cancers is missing).

Following Hasan et al. (2009), who analyze the same data through a pattern-mixture ZIP model to account for informative dropout, we consider in the proposed model the following baseline covariates, which affect both the longitudinal and the survival processes: age (in years), type of skin (1 for burned individuals, 0 otherwise), sex (0 for females and 1 for males), number of previous skin cancers (exposure), treatment (0 for placebo and 1 for β -carotene), time (in years, from 1 to 5).

In Table 5 we report the parameter estimates of the JM under the general assumption of MNAR with AR(1) random effects (columns 2-4) and under the specific assumption of MNAR with time-constant random effects (columns 5-7), other than the constrained model that assumes ignorable dropout (columns 8-10).

We observe that results concerning the longitudinal process are perfectly in agreement with those obtained by Hasan et al. (2009), both in terms of estimated coefficients and in terms of statistical significance. More in detail, the risk of skin cancer is significantly higher for older males with burned skin and with a previous history of skin cancer rather than for younger females having a nonburned skin and not previous skin cancers. On the

	MNAR with AR(1) r.e.			MNAR with time-constant r.e.			MAR		
	Est.	se	<i>p</i> -value	Est.	se	<i>p</i> -value	Est.	se	<i>p</i> -value
<i>Longitudinal sub-model</i>									
Intercept	-4.535	0.333	0.000	-4.210	0.324	0.000	-4.528	0.334	0.000
Age	0.019	0.005	0.000	0.018	0.005	0.000	0.019	0.005	0.000
Skin (burns)	0.327	0.085	0.000	0.328	0.085	0.000	0.331	0.086	0.000
Sex (males)	0.650	0.099	0.000	0.644	0.099	0.000	0.650	0.099	0.000
Exposure	0.177	0.010	0.000	0.176	0.010	0.000	0.178	0.010	0.000
Treatment (β -carotene)	0.119	0.085	0.158	0.129	0.085	0.126	0.118	0.085	0.165
Time	0.014	0.024	0.568	0.022	0.020	0.278	0.007	0.024	0.781
<i>Survival sub-model</i>									
Intercept	3.894	0.275	0.000	3.893	0.275	0.000	—	—	—
Age	-0.001	0.004	0.821	-0.001	0.004	0.821	—	—	—
Skin (burns)	-0.152	0.075	0.044	-0.151	0.075	0.044	—	—	—
Sex (males)	-0.230	0.083	0.005	-0.231	0.083	0.005	—	—	—
Exposure	-0.006	0.011	0.615	-0.005	0.011	0.624	—	—	—
Treatment (β -carotene)	-0.162	0.075	0.030	-0.162	0.075	0.030	—	—	—
Time	-0.507	0.029	0.000	-0.507	0.029	0.000	—	—	—
$\hat{\gamma}$	-0.090	0.059	0.126	-0.086	0.062	0.164	—	—	—
<i>Others</i>									
$\hat{\tau}$	0.862	0.040	0.000	0.739	0.026	0.000	0.861	0.040	0.000
$\hat{\sigma}_{\eta}^2$	1.426	0.126	0.000	1.141	0.105	0.000	1.429	0.126	0.000
$\hat{\rho}$	0.852	0.032	0.000	—	—	—	0.853	0.031	0.000

Table 5: *Parameter estimates for the number of yearly new skin cancers, under the assumptions of nonignorable dropout (MNAR) with AR(1) random effects, MNAR with time-constant random effects, and ignorable dropout (MAR).*

other hand, the data at issue do not provide any evidence of a significant effect of the treatment based on β -carotene (p -values greater than 0.10), coherently with the results obtained by Greenberg et al. (1990) and Hasan et al. (2009). Moreover, although the effect of time is not significant, the estimated value of the autocorrelation coefficient ($\hat{\rho} = 0.85$) denotes that the count of new skin cancers at a given year highly depends on the count at the previous year. This result argues in favor of AR(1) random effects rather than time-constant random effects, especially as the model specified under the former assumption has a better goodness-of-fit than the model under the latter assumption (BIC values equal to 12,957.22 and 12,966.81, respectively).

Concerning the dropout process, the association parameter (γ) estimated both under the assumption of AR(1) random effects and under the assumption of time-constant random effects is negative, but not statistically significant (p -values greater than 0.10). This result agrees with the conclusions of Hasan et al. (2009) and provides evidence in favor of the MAR assumption. As further evidence, we observe that the parameter estimates obtained under the MAR assumption (Table 5, columns 8-10) are very similar to those obtained under the MNAR assumption (Table 5, columns 2-4 and 5-7).

Different to the pattern-mixture model of Hasan et al. (2009), our proposed JM (both

with AR(1) and with time-constant random effects) provides a deeper insight of the dropout process. We observe that males with burned skin and treated with β -carotene tend to exit from the follow-up significantly before than other individuals.

Finally, as outlined in the previous section, also in this context the proposed estimation algorithm is highly stable: the log-likelihood at convergence obtained through a random rule equals the best solution 99 out of 100 times and corresponds to the solution reached by the deterministic initialization strategy.

5.3 Example 3: Bivariate analysis of two biomarkers of primary biliary cirrhosis

We analyze a dataset concerning 312 patients affected by primary biliary cirrhosis, which represents a chronic and fatal liver disease characterized by destruction of bile ducts and, eventually, by cirrhosis of the liver. The data were collected by the Mayo Clinic from 1974 to 1984 (Murtaugh et al., 1994) during a randomized study for the treatment of primary biliary cirrhosis through a D-penicillamine based drug. Patients affected by this disease present several biomarkers associated with the disease progression. Here we focus on the (logarithm of) serum bilirubin (in mg/dl) and on the possible presence of edema that are a consequence of accumulation of toxic compounds and fluids. The total number of measurements is 1,945 and ranges from 1 to 16 for each patient with average 6.2 (standard deviation 3.8) and median 5, then the dropout represents a relevant aspect of the study that enforces the assumption of nonignorable missingness.

We analyze the effect of D-penicillamine on the logarithm of serum bilirubin and on the presence of edema jointly with the status of patients, being the status equal to 1 if the patient is free-transplantation alive and 0 in case of death or transplantation at a certain time. Since the first longitudinal outcome is described by a continuous normal variable and the second longitudinal outcome is described by a binary variable, the link functions of the sub-model (10), with $r = 2$, specify as an identity function and a logit function, similarly to equations (2) and (3), respectively. Following Bartolucci and Farcomeni (2015), who perform a similar analysis on the same data adopting a mixed latent Markov model, we also account for the following baseline covariates: age, gender, albumin in

mg/dl, logarithm of alkaline phosphatase in U/L, logarithm of transaminase (SGOT) in U/ml. Besides, we adjust for the effect of time and for the interaction between time and treatment. We remind that, coherently with our proposed JM, in the sub-models for the longitudinal process time is a continuous variable representing the time of measurement, whereas in the sub-model for the survival process time is a discrete variable representing the time interval (in semesters), in which the dropout may occur and a varying number of measurements of the serum bilirubin and of the presence of edema may be observed.

In Table 6 we illustrate the parameter estimates of the JM under the general assumption of MNAR with AR(1) random effects (columns 2-4) and under the specific assumption of MNAR with time-constant random effects (columns 5-7), other than the constrained model that assumes ignorable dropout (columns 8-10). We also report the survival curves (Figure 1) under the assumptions of AR(1) and time-constant random effects for a generic patient having average characteristics at baseline and randomized to a placebo (left panel) or a treatment based on D-penicillamine (right panel). In the case of AR(1) random effects, we adopt the average of 1,000 values simulated according to the AR(1) process of equation (7), whereas in the case of time-constant random effects we simulate 1,000 independent values from a Gaussian distribution with mean 0.

From the results obtained under the estimated models we observe that the treatment based on D-penicillamine has a beneficial effect on the log serum bilirubin and on the presence of edema; however, this effect is not statistically significant. The same result is obtained for the probability of free-transplantation survival (see also Figure 1).

Concerning the other covariates, the expected values of log serum bilirubin increase with the levels of alkaline phosphatase and transaminase (SGOT) at baseline, whereas they reduce with the level of albumin and are lower for females with respect to males; the effect of age is not significant. On the other hand, the probability of edema increases with the age and with the levels of transaminase at the baseline and, on the contrary, it reduces in presence of increasing levels of albumine, whereas the level of alkaline phosphatase is not significant. Besides, the effect of gender on the presence of edema is opposite compared to the effect on the log serum bilirubin: indeed, edema are more common in females rather than in males.

	MNAR with AR(1) r.e.			MNAR with time-constant r.e.			MAR		
	Est.	se	p-value	Est.	se	p-value	Est.	se	p-value
<i>Longitudinal sub-model - Log(serum Bilirubin)</i>									
Intercept	-3.677	0.985	0.000	-3.607	0.925	0.000	-3.699	0.969	0.000
Treatment	-0.108	0.113	0.342	-0.119	0.101	0.239	-0.107	0.111	0.336
Age/10	0.032	0.054	0.553	0.029	0.050	0.568	0.028	0.053	0.601
Female	-0.421	0.168	0.012	-0.418	0.158	0.008	-0.408	0.166	0.014
Albumin	-0.711	0.135	0.000	-0.693	0.126	0.000	-0.688	0.133	0.000
Log-alkaline ph.	0.249	0.075	0.001	0.234	0.071	0.001	0.242	0.074	0.001
Log-SGOT	1.135	0.127	0.000	1.136	0.119	0.000	1.130	0.125	0.000
Time	0.118	0.011	0.000	0.094	0.006	0.000	0.109	0.011	0.000
Treatment.time	-0.003	0.016	0.852	0.010	0.009	0.239	0.000	0.016	0.990
$\hat{\phi}_1$	0.969	0.036	0.000	0.831	0.036	0.000	0.950	0.036	0.000
<i>Longitudinal sub-model - Edema</i>									
Intercept	-1.084	1.459	0.458	-1.290	1.496	0.389	-1.112	1.443	0.441
Treatment	-0.216	0.204	0.291	-0.227	0.204	0.265	-0.215	0.203	0.288
Age/10	0.688	0.083	0.000	0.688	0.085	0.000	0.682	0.082	0.000
Female	0.599	0.248	0.016	0.580	0.252	0.021	0.609	0.245	0.013
Albumin	-1.423	0.208	0.000	-1.429	0.212	0.000	-1.397	0.206	0.000
Log-alkaline ph.	0.054	0.106	0.611	0.058	0.109	0.593	0.048	0.105	0.645
Log-SGOT	0.779	0.187	0.000	0.822	0.192	0.000	0.773	0.185	0.000
Time	0.275	0.030	0.000	0.267	0.028	0.000	0.266	0.030	0.000
Treatment.time	-0.031	0.039	0.419	-0.018	0.036	0.626	-0.028	0.039	0.467
$\hat{\phi}_2$	0.924	0.076	0.000	0.915	0.084	0.000	0.898	0.074	0.000
<i>Survival sub-model</i>									
Intercept	2.760	2.398	0.250	3.537	2.238	0.114	—	—	—
Treatment	0.035	0.371	0.925	0.108	0.339	0.749	—	—	—
Age/10	-0.362	0.128	0.005	-0.359	0.122	0.003	—	—	—
Female	0.858	0.376	0.023	0.813	0.354	0.022	—	—	—
Albumin	2.120	0.339	0.000	1.980	0.315	0.000	—	—	—
Log-alkaline ph.	0.003	0.171	0.988	-0.057	0.162	0.727	—	—	—
Log-SGOT	-1.306	0.304	0.000	-1.329	0.289	0.000	—	—	—
Time	-0.296	0.048	0.000	-0.286	0.043	0.000	—	—	—
Treatment.time	0.053	0.061	0.383	0.032	0.051	0.533	—	—	—
$\hat{\psi}$	-1.590	0.140	0.000	-1.356	0.135	0.000	—	—	—
<i>Others</i>									
$\hat{\sigma}_\varepsilon^2$	0.074	0.004	0.000	0.247	0.009	0.000	0.077	0.004	0.000
$\hat{\rho}$	0.972	0.003	0.000	—	—	—	0.972	0.003	0.000

Table 6: *Parameter estimates for the log serum bilirubin and the presence of edema, under the assumptions of nonignorable dropout (MNAR) with AR(1) random effects, MNAR with time-constant random effects, and ignorable dropout (MAR).*

Regarding the survival process, the probability of free-transplantation survival significantly decreases with age and with the levels of transaminase, whereas it is higher for females and patients having higher levels of albumine. The status of any patient is also negatively correlated ($\hat{\psi}$ equals to -1.59 in the case of AR(1) random effects and to -1.35 in the case of time-constant random effects) with the two longitudinal processes, confirming that the log serum bilirubin and the presence of edema are two relevant biomarkers of the health status and cannot be ignored.

As expected, both the log serum bilirubin and the probability of edema increase over time and, at the same time, the probability of free-transplantation survival reduces (Figure 1).

Finally, we observe that the results in terms of free-transplantation survival under the two different assumptions about the random effects are quite different, although the esti-

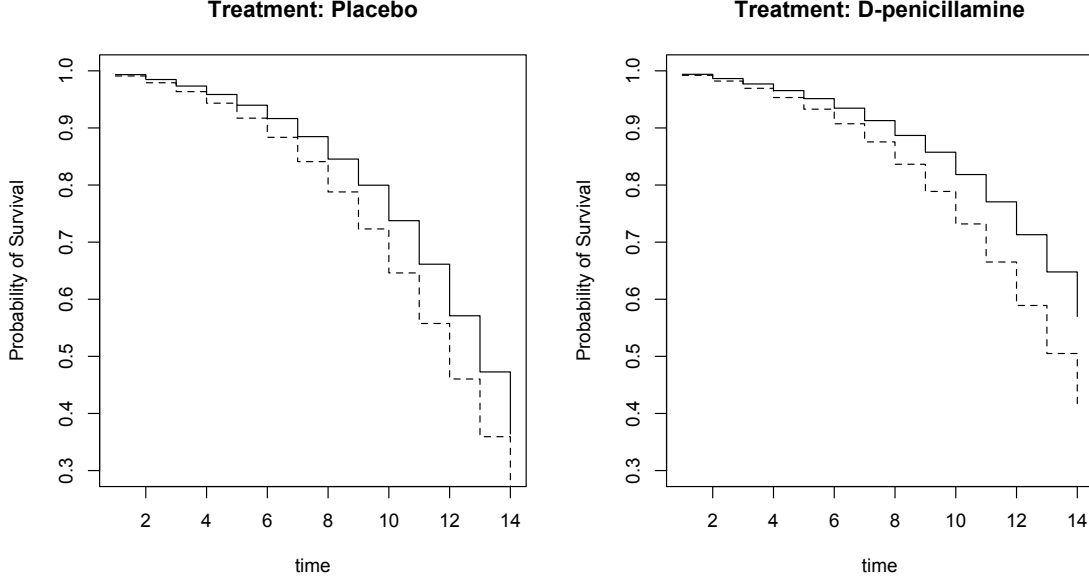


Figure 1: *Survival curves for placebo patients (left) and for patients treated with D-penicillamine (right), under assumptions of AR(1) (solid lines) and time-constant (dashed lines) random effects.*

mated coefficients appear similar. As shown in Figure 1, the model with autocorrelated random-effects (solid lines) provides a more positive trend of the probability of survival rather than the restricted one (dashed lines), both under the placebo and under the treatment based on D-penicillamine. We also observe that the formulation of simpler models (i.e., with time-constant random-effects or ignoring the dropout) leads to underestimating the expected values of log serum bilirubin with respect to the model formulation that accounts for dropout through AR(1) random effects (Figure 2).

With respect to the analysis on the same variables proposed by Bartolucci and Farcomeni (2015), we observe that our results are only partially aligned with those obtained by the authors under their “best” model (see their Table 3, columns 8-10). However, it is important to note that the mixed latent Markov model proposed by the authors is less parsimonious with respect to our model and, most of all, it provides rather unstable results, which depend on the number of latent classes that have to be a priori chosen (for instance, compare their Table 3, columns 5-7 versus columns 8-10).

Moreover, the comparison in terms of goodness-of-fit allows us to conclude that the

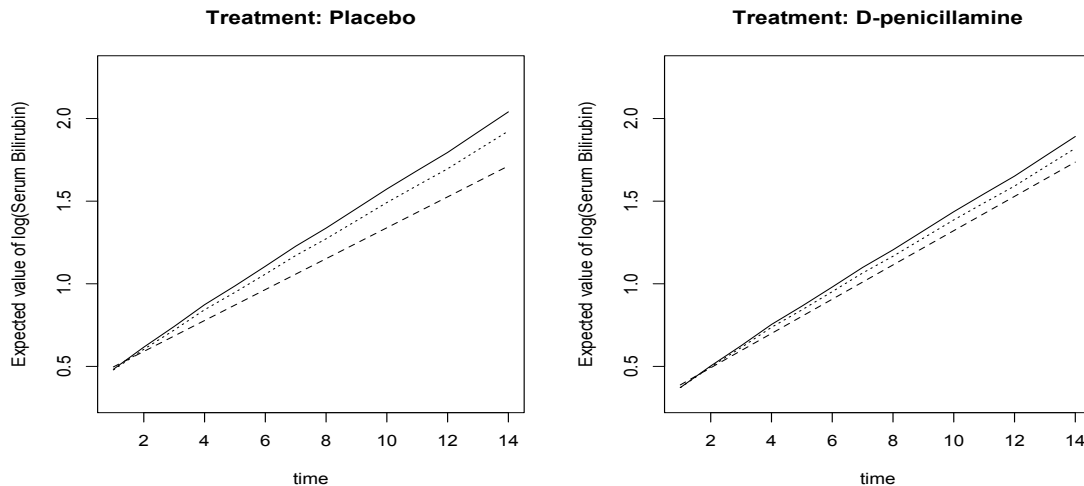


Figure 2: *Trend of log serum bilirubin for placebo patients (left) and for patients treated with D-penicillamine (right), under MNAR assumption with AR(1) random effects (solid lines), MNAR assumption with time-constant random effects (dashed lines), and MAR assumption (dotted lines).*

proposed model under the MNAR assumption with AR(1) random effects (maximum log-likelihood at convergence equal to -2,710.57, BIC value equal to 5,604.92 and a number of parameters equal to 32) always outperforms the models of Bartolucci and Farcomeni (2015) which do not include time-fixed latent variables (see the first row of their Table 2, with a number of support points of the time-constant latent variable, k_1 , equal to 1). Even with reference to the same number of parameters, the proposed model has a better goodness-of-fit than the corresponding model of Bartolucci and Farcomeni (2015). Finally, the approach of Bartolucci and Farcomeni (2015), being based on a discrete Markov chain, suffers from the problem of local maxima of the model log-likelihood, as shown in the lower panel of their Table 2. As observed for the two previously described applications, the estimation algorithm here proposed does not suffer from this drawback: in the case of bivariate data, the 100% of log-likelihood values at convergence obtained through the random rule are equal to the best solution, which corresponds to the solution reached by the deterministic strategy.

6 Conclusions

In this article we propose a Joint Modeling (JM) approach to deal with longitudinal data affected by nonignorable dropout, which is a very relevant field of application especially, but not only, in medicine. The approach is based on accounting for the unobserved heterogeneity between subjects in a dynamic fashion, by adopting an AR(1) process. This latent process has also the role of connecting the longitudinal response process with the survival process, and the strength of this association is measured on the basis of suitable parameters. In this way we generalize the typical JM approach (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Tsiatis and Davidian, 2004; Rizopoulos, 2012) in which the random effects are usually assumed to be time constant.

Our approach is very flexible, as it may deal with response variables of any nature through a generalized linear model parametrization, and even with counts having an excess of zeros and categorical/ordinal responses. Moreover, the approach may be also used in the presence of multivariate longitudinal data in which we observe a vector of outcomes at each occasion. Parameter estimation of the proposed model relies on a sequential quadrature algorithm that is rather simple to be implemented and that has been used, in the simple context in which longitudinal responses are not affected by dropout, in Bartolucci et al. (2014).

Overall, the proposed approach is comparable with two very recent approaches in the literature about joint modeling a longitudinal response and a survival process. We refer in particular to the proposals of Barrett et al. (2015) and Bartolucci and Farcomeni (2015). Both these contributions are based on a discretization of the time-to-event measurements, however they differ in the way the time constancy of the random effects assumption is relaxed. Barrett et al. (2015) adopt a linear mixed Gaussian model for the longitudinal process and a sequential probit model for the survival process, so as to accommodate different specifications of random effects. On the contrary, Bartolucci and Farcomeni (2015) assume that the random effects follow a first order Markov chain with a finite number of latent states. Different from the proposal of Barrett et al. (2015), the approach we here propose is more general, as it allows us to deal with response variables of any

type and, with respect to the approach of Bartolucci and Farcomeni (2015), the present one has advantages in terms of stability.

These advantages are illustrated by three applications on medical real data. The first application involves a continuous longitudinal outcome and allows us to validate the proposed estimation approach through a comparison with the same model estimated by means of the algorithm proposed by Barrett et al. (2015). The second application concerns count data with excess of zeros and shows the potentialities of our proposal in applications involving longitudinal outcomes whose distribution does not necessarily belong to the exponential family. Finally, the third application involves a bivariate longitudinal outcome characterized by two variables of different nature (continuous and binary). In all examples, the proposed approach leads to highly stable estimation results.

Finally, we stress that a key point of the proposed approach is that it has a simple interpretation and, at the same time, it is flexible and estimable by an algorithm rather simple to implement (the R software is made available upon request). Given this, we avoid to explicit consider versions of the proposed model for multivariate data which are based on a specific AR(1) process for each longitudinal response process. In fact, as we remark at the end of Section 2.2.2, the resulting model would require much more sophisticated computational tools, based on adaptive quadrature methods (Rizopoulos, 2012), or even Bayesian inference tools. However, as we experimented by some attempts for bivariate longitudinal data, the resulting improvement in terms of goodness-of-fit might be not enough to justify the additional computational effort.

Appendix

Here we report the results of the additional scenarios simulated according to what described in Section 4.1. In particular, for continuous (Table 7), binary (Table 8), and count data (Table 9), Scenario 5 is characterized by an increase in the number of repeated measurements, $m_i = 20$, Scenario 6 shows the results of the simulation with $\rho = 0.9$ and Scenario 7 assumes a different value of the association parameter $\gamma = 0.5$. Finally, for the case of continuous data, Scenario 8 is characterized by an higher variance of the repeated measurements $\sigma_\epsilon^2 = 4$.

Parameters	True	Scenario 5 $m_i = 20$				Scenario 6 $\rho = 0.9$				
		Mean	sd	se	RMSE	True	Mean	sd	se	RMSE
Longitudinal										
Intercept	90.000	89.998	0.115	0.115	0.115	90.000	90.003	0.130	0.137	0.130
Time	-1.700	-1.701	0.014	0.014	0.014	-1.700	-1.699	0.014	0.015	0.015
Age at t_0	-1.700	-1.700	0.005	0.005	0.005	-1.700	-1.700	0.006	0.006	0.006
Sex	2.000	1.999	0.058	0.059	0.058	2.000	2.004	0.069	0.071	0.069
Survival										
Intercept	2.000	1.991	0.221	0.224	0.221	2.000	2.011	0.228	0.224	0.229
Time	0.010	0.011	0.039	0.038	0.039	0.010	0.009	0.037	0.038	0.037
Age at t_0	0.010	0.010	0.009	0.009	0.009	0.010	0.010	0.009	0.009	0.009
Sex1	0.100	0.095	0.108	0.107	0.108	0.100	0.090	0.104	0.107	0.105
γ	0.050	0.046	0.072	0.069	0.072	0.050	0.051	0.074	0.073	0.074
Others										
σ_ϵ^2	1.000	1.000	0.019	0.019	0.019	1.000	0.998	0.029	0.029	0.029
σ_η^2	1.000	0.994	0.041	0.043	0.042	1.000	0.997	0.058	0.058	0.058
ρ	0.700	0.697	0.019	0.020	0.020	0.900	0.900	0.015	0.017	0.015
Parameters	True	Scenario 7 $\gamma = 0.5$				Scenario 8 $\sigma_\epsilon^2 = 4$				
		Mean	sd	se	RMSE	True	Mean	sd	se	RMSE
Longitudinal										
Intercept	90.000	89.996	0.127	0.127	0.127	90.000	90.003	0.157	0.168	0.157
Time	-1.700	-1.700	0.017	0.017	0.017	-1.700	-1.701	0.025	0.025	0.025
Age at t_0	-1.700	-1.700	0.006	0.006	0.006	-1.700	-1.700	0.007	0.007	0.007
Sex	2.000	2.000	0.066	0.065	0.066	2.000	1.998	0.083	0.084	0.083
Survival										
Intercept	2.000	1.997	0.227	0.229	0.227	2.000	2.009	0.227	0.224	0.228
Time	0.010	0.009	0.038	0.038	0.038	0.010	0.009	0.037	0.038	0.037
Age at t_0	0.010	0.010	0.009	0.009	0.009	0.010	0.010	0.009	0.009	0.009
Sex	0.100	0.093	0.106	0.109	0.106	0.100	0.101	0.108	0.107	0.108
γ	0.500	0.506	0.084	0.083	0.084	0.050	0.053	0.112	0.110	0.112
Others										
σ_ϵ^2	1.000	1.000	0.032	0.031	0.032	4.000	4.004	0.122	0.116	0.122
σ_η^2	1.000	0.994	0.054	0.054	0.055	1.000	1.000	0.110	0.110	0.110
ρ	0.700	0.699	0.029	0.028	0.029	0.700	0.698	0.065	0.065	0.065

Table 7: Mean of parameter estimates, standard deviation (sd), average estimated standard error (se) and root mean square error (RMSE) for continuous data: Scenarios 5-8.

Acknowledgments

The authors acknowledge the financial support from award RBFR12SHVV of the Italian Government (FIRB “Mixture and latent variable models for causal inference and analysis of socio-economic data”, 2012). The authors are also grateful to the UK Cystic Fibrosis registry for making available the data.

References

Azzalini, A. (2005). The skew-normal distribution and related multivariate families (with discussion). *Scandinavian Journal of Statistics*, 32:159–200.

Parameters	True	Scenario 5 $m_i = 20$				Scenario 6 $\rho = 0.9$				
		Mean	sd	se	RMSE	True	Mean	sd	se	RMSE
Longitudinal										
Intercept	4.000	4.003	0.178	0.180	0.178	4.000	3.999	0.234	0.235	0.234
Time	-0.200	-0.200	0.023	0.024	0.023	-0.200	-0.199	0.031	0.030	0.031
Age at t_0	-0.200	-0.200	0.008	0.008	0.008	-0.200	-0.200	0.010	0.011	0.010
Sex	2.000	2.006	0.087	0.089	0.087	2.000	1.999	0.119	0.117	0.119
Survival										
Intercept	2.000	1.999	0.236	0.224	0.236	2.000	1.998	0.220	0.225	0.220
Time	0.010	0.008	0.041	0.038	0.041	0.010	0.007	0.038	0.038	0.039
Age at t_0	0.010	0.010	0.009	0.009	0.009	0.010	0.010	0.010	0.009	0.010
Sex	0.100	0.101	0.103	0.107	0.103	0.100	0.105	0.108	0.107	0.109
γ	0.050	0.046	0.112	0.106	0.112	0.050	0.051	0.114	0.118	0.114
Others										
σ_η^2	1.000	0.999	0.121	0.116	0.121	1.000	1.004	0.179	0.181	0.179
ρ	0.700	0.697	0.052	0.053	0.052	0.900	0.900	0.061	0.292	0.061
Scenario 7 $\gamma = 0.5$										
Parameters	True	Mean	sd	se	RMSE					
Longitudinal										
Intercept	4.000	3.998	0.226	0.233	0.226					
Time	-0.200	-0.199	0.032	0.032	0.032					
Age at t_0	-0.200	-0.200	0.010	0.011	0.010					
Sex	2.000	2.007	0.120	0.115	0.120					
Survival										
Intercept	2.000	2.001	0.236	0.235	0.236					
Time	0.010	0.009	0.040	0.039	0.040					
Age at t_0	0.010	0.010	0.010	0.010	0.010					
Sex	0.100	0.097	0.110	0.110	0.110					
γ	0.500	0.528	0.171	0.161	0.173					
Others										
σ_η^2	1.000	1.003	0.189	0.195	0.189					
ρ	0.700	0.698	0.094	0.096	0.094					

Table 8: Mean of parameter estimates, standard deviation (sd), average estimated standard error (se) and root mean square error (RMSE) for binary data: Scenarios 5-7.

Barrett, J., Diggle, P., Henderson, R., and Taylor-Robinson, D. (2015). Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society, Series B*, 77:131–148.

Bartolucci, F., Bacci, S., and Pennoni, F. (2014). Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *Journal of the Royal Statistical Society-Series C*, 63:267–288.

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

Parameters	True	Scenario 5 $m_i = 20$				Scenario 6 $\rho = 0.9$				
		Mean	sd	se	RMSE	True	Mean	sd	se	RMSE
Longitudinal										
Intercept	2.000	2.000	0.056	0.056	0.056	2.000	2.000	0.070	0.065	0.070
Time	-0.020	-0.020	0.006	0.006	0.006	-0.020	-0.020	0.007	0.006	0.007
Age at t_0	-0.020	-0.020	0.002	0.003	0.002	-0.020	-0.020	0.003	0.003	0.003
Sex	0.500	0.499	0.028	0.029	0.028	0.500	0.500	0.033	0.034	0.033
Survival										
Intercept	2.000	2.006	0.219	0.224	0.220	2.000	1.990	0.219	0.224	0.219
Time	0.010	0.009	0.038	0.038	0.038	0.010	0.012	0.036	0.038	0.036
Age at t_0	0.010	0.010	0.009	0.009	0.009	0.010	0.010	0.009	0.009	0.009
Sex	0.100	0.092	0.103	0.107	0.103	0.100	0.090	0.108	0.107	0.108
γ	0.050	0.048	0.132	0.132	0.132	0.050	0.058	0.138	0.138	0.138
Others										
σ_η^2	0.250	0.250	0.010	0.010	0.010	0.250	0.249	0.013	0.013	0.013
ρ	0.700	0.699	0.017	0.017	0.017	0.900	0.899	0.011	0.011	0.011
Scenario 7 $\gamma = 0.5$										
Parameters	True	Mean	sd	se	RMSE					
Longitudinal										
Intercept	2.000	1.993	0.061	0.060	0.061					
Time	-0.020	-0.020	0.008	0.008	0.008					
Age at t_0	-0.020	-0.020	0.003	0.003	0.003					
Sex	0.500	0.499	0.031	0.031	0.031					
Survival										
Intercept	2.000	1.997	0.227	0.225	0.227					
Time	0.010	0.008	0.038	0.038	0.038					
Age at t_0	0.010	0.010	0.009	0.009	0.009					
Sex	0.100	0.101	0.105	0.107	0.105					
γ	0.500	0.505	0.153	0.152	0.153					
Others										
σ_η^2	0.250	0.249	0.011	0.012	0.011					
ρ	0.700	0.697	0.022	0.022	0.022					

Table 9: Mean of parameter estimates, standard deviation (sd), average estimated standard error (se) and root mean square error (RMSE) for count data: Scenarios 5-7.

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

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

Cox, D. (2007). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B*, 34:187–220.

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

- Follmann, D. and Wu, M. (1995). An approximate generalized liner model with random effects for informative missing data. *Biometrics*, 51:151–168.
- Greenberg, E. R., Baron, J. A., Stukel, T. A., Stevens, M. M., Mandel, J. S., Spencer, S. K., Elias, P. M., Lowe, N., Nierenberg, D. W., Bayrd, G., Vance, J. C., Freeman, D. H., Clendenning, W. E., Kwan, T., and the Skin Cancer Prevention Study Group (1990). A clinical trial of beta carotene to prevent basal-cell and squamous-cell cancers of the skin. *New England Journal of Medicine*, 323:789–795.
- Hasan, M. T., Sneddon, G., and Ma, R. (2009). Pattern-mixture zero-inflated mixed models for longitudinal unbalanced count data with excessive zeros. *Biometrical Journal*, 51:946–960.
- Heiss, F. (2008). Sequential numerical integration in nonlinear state space models for microeconomic panel data. *Journal of Applied Econometrics*, 23:373–389.
- Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1:465–480.
- Hogan, J. and Laird, N. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, 16:239–258.
- Hogan, J. and Laird, N. (1998). Increasing efficiency from censored survival data by using random effects to model longitudinal covariates. *Statistical Methods in Medical Research*, 7:28–48.
- Huang, X., Stefanski, L., and Davidian, M. (2009). Latent-model robustness in joint models for a primary endpoint and a longitudinal process. *Biometrics*, 64:719–727.
- Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14.
- Lavalley, M. P. and DeGruttula, V. (1996). Model for empirical Bayes estimators of longitudinal CD4 counts. *Statistics in Medicine*, 15:2289–2305.

- Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of American Statistical Association*, 88:125–134.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley Sons Hoboken NJ.
- Lystig, T. C. and Hughes, J. (2002). Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics*, 11:678–689.
- McCulloch, C. and Searle, S. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.
- Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5:1–19.
- Murtaugh, P., Dickson, E., Van Dam, G., Malincho, M., Grambsch, P., Langworthy, A., and Gips, C. (1994). Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, 20:126–134.
- Rizopoulos, D. (2012). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics & Data Analysis*, 56(3):491–501.
- Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika*, 95:63–74.
- Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data with applications in R*. Chapman&Hall/CRC Press, Boca Raton, FL.
- Song, X., Davidian, M., and Tsiatis, A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58:742–753.
- Taylor, J. M. G. (1994). A stochastic model for analysis of longitudinal data. *Journal of the American Statistical Association*, 89:727–776.

- Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14:809–834.
- Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, 96:895–905.
- Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, 53:1–13.
- Wu, M. and Bailey, K. (1988). Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine*, 7:337–346.
- Wu, M. and Bailey, K. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, 45:939–955.
- Wu, M. and Carroll, R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44:175–188.
- Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339.
- Xu, J. and Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*, 50:375–387.
- Xu, S., Jones, R., and Grunwald, G. (2007). Analysis of longitudinal count data with serial correlation. *Biometrical Journal*, 49:416–428.
- Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC Press, Boca Raton, FL.