A General Framework for Joint Multi-State Models

Félix Laplante[1] and Christophe Ambroise[2*]

[1*]MaIAGE, Université Paris-Saclay, INRAE, France.
[2]Laboratoire de Mathématiques et Modélisation d'Évry (LaMME), Université Paris-Saclay, CNRS, Univ Évry, France.

*Corresponding author(s). E-mail(s): christophe.ambroise@univ-evry.fr;
Contributing authors: felixlaplante0@proton.me;

Abstract

Classical joint modeling approaches often rely on competing risks or recurrent event formulations to describe complex processes involving evolving longitudinal biomarkers and discrete event occurrences, but these frameworks typically capture only limited aspects of the underlying event dynamics.

We propose a general multi-state joint modeling framework that unifies longitudinal biomarker dynamics with multi-state time-to-event processes defined on arbitrary directed graphs. The proposed framework accommodates arbitrary directed transition graphs, nonlinear longitudinal submodels, and scalable inference via stochastic gradient descent. This formulation encompasses both Markovian and semi-Markovian transition structures, allowing recurrent cycles and terminal absorptions to be naturally represented. The longitudinal and event processes are linked through shared latent structures within nonlinear mixed-effects models, extending classical joint modeling formulations.

We derive the complete likelihood, establish conditions for identifiability, and develop scalable inference procedures based on stochastic gradient descent to enable high-dimensional and large-scale applications. In addition, we formulate a dynamic prediction framework that provides individualized state-transition probabilities and personalized risk assessments along complex event trajectories. Through simulation and application to the PAQUID cohort, we demonstrate accurate parameter recovery and individualized prediction.

Keywords: joint modeling, multi-state processes, longitudinal data, survival analysis, stochastic gradient descent, dynamic prediction

1 Introduction

Joint modeling of longitudinal and time-to-event data has become an essential tool of modern biostatistics [27], particularly for dynamic prediction in clinical applications. Classical joint models typically couple a longitudinal biomarker process with a single time-to-event outcome [31], allowing for the integration of biological knowledge and individual heterogeneity via shared latent structures. However, many real-world processes involve multiple possible outcomes, intermediate stages, or recurrent events, which cannot be fully captured by a single-event framework [20]. In such settings, multi-state models provide a more flexible approach [14].

Multi-state models represent an overall joint process of clinical course as a series of discrete stages or health states that occur sequentially [25]. In biostatistics, these models are widely used for survival and reliability analysis, allowing for a richer and more accurate representation by capturing alternative paths to an event of interest, intermediate events, and progressive disease. Key components of multi-state models include transition intensity functions, which denote the instantaneous risk of moving from one state to another, and transition probability functions, which describe the probability of transition over longer intervals. Often, these models assume the Markov property, where future transitions depend only on the current state, simplifying the transition intensity functions [3].

The Markov assumption can be relaxed by adopting a semi-Markov formulation, in which the transition probabilities depend not only on the current state but also on the sojourn time, effectively resetting the time scale after each transition.

The link between multi-state models and joint models arises when a multi-state process is integrated as a component within a broader joint modeling framework . While multi-state models, such as those described in the "sequential state framework" primarily focus on movements between discrete states, joint models operate under a "parallel trajectory framework" that combines a longitudinal process with a time-to-event process.

The model proposed by Ferrer et al. [14] exemplifies this link by presenting a joint model for a longitudinal process (e.g., Prostate-Specific Antigen (PSA) measurements) and a multi-state process (e.g., clinical progressions in prostate cancer). These two submodels are interconnected by shared random effects, allowing the model to account for the correlation between the continuous longitudinal biomarker trajectory and the discrete transitions between health states.

The model we propose operates on an arbitrary directed graph, enabling representation of complex state transitions and recurrent events. We derive the complete likelihood for the nonlinear joint model, introduce an efficient stochastic approximation inference method, and develop dynamic prediction tools. To illustrate its practical relevance, we conduct both a simulation study and a synthetic biomedical case study.

This paper is organized as follows. Section 2 reviews the background on multi-state and joint modeling frameworks. Section 3.2 introduces the proposed general multi-state joint modeling framework, detailing the likelihood formulation and identifiability conditions. Section 4 describes the stochastic-gradient–based inference algorithm, while Section 5 presents the dynamic prediction framework. Section 6 reports simulation experiments that assess parameter recovery, identifiability, and convergence

properties of the proposed model, and Section 7 applies the framework to data from the PAQUID cohort. Finally, Section 8 concludes with perspectives and future research directions.

2 Background

Joint modeling of longitudinal and time-to-event data provides a unified statistical framework to analyze the interplay between continuous biomarker dynamics and event occurrence processes. In its classical form, this framework couples a mixed-effects model describing individual biomarker trajectories with a survival model for event times [31]. Such models have become central in biomedical research, particularly for dynamic prediction and personalized risk assessment [27]. However, these traditional joint models generally focus on a single terminal event, limiting their ability to represent more complex event histories that include intermediate, recurrent, or competing events. To address these limitations, multi-state extensions have been developed, providing a natural framework for describing transitions between multiple clinical or biological states over time [14, 20].

2.1 Multi-State Markov Processes

A *multi-state stochastic process* is defined as a process $S(t)$ for $t \geq 0$, where $S(t)$ can take a finite number of values (states), often labeled $1, 2, \ldots, p$. Quantities of interest typically include the probability of being in a certain state at a given time and the distribution of first passage times.

A *Markov process* (or continuous-time Markov chain) is a specific class of multi-state models in which future transitions between states *depend only upon the current state*. This *Markov property* means the process is *memoryless*. A key consequence is that the duration spent in any state follows an *exponential distribution*, implying a constant hazard rate for leaving that state [18, 29].

However, the Markov assumption can be unrealistic in many real-world applications. For example, in the study of human sleep stages, sojourn times often do not follow an exponential distribution [11, 35], and in the case of chronic diseases such as AIDS, the risk of disease progression can depend on the time elapsed since infection [2]. To address these limitations, *semi-Markov processes* (SMPs) were introduced, which allow for arbitrary distributions of sojourn times while retaining the Markov property for the embedded discrete-time chain. This flexibility makes SMPs suitable for modeling complex disease progression and patient recovery scenarios [29].

2.2 Multi-State Semi-Markov Processes

Multi-state semi-Markov processes (MSMPs) offer a natural generalization by allowing the distribution of sojourn times to be arbitrary.

In an MSMP, the process is defined by a directed graph $G = (V, E)$, where V represents the set of states and E the set of possible transitions. The transition intensities $\lambda_{k'|k}(t \mid t_0)$ from state k to state k' at time t, given entry time t_0 are therefore assumed to be non-identically zero. The sojourn time in state k is not restricted to

an exponential distribution, allowing for more realistic modeling of the time spent in each state.

MSMPs have been widely applied in a range of disciplines. In reliability, they are used to modeling degradation and repair processes [24]. In biomedical studies, they have proven useful for analyzing illness-death models or disease progression with non-exponential transitions [7, 15]. Applications in finance also exist, where semi-Markovian dynamics can model credit rating migrations or economic regimes [26]. In these contexts, MSMPs retain the Markov property in the embedded jump chain while offering increased realism through flexible dwell time modeling.

From a methodological standpoint, MSMPs extend estimation strategies developed for Markov models, such as maximum likelihood or Bayesian inference, and can accommodate interval-censored or misclassified data [16]. This makes them a powerful and general tool for multi-state event history analysis.

While both joint models and multi-state models offer powerful tools, they primarily address different facets of disease progression and have complementary strengths. Multi-state models excel at understanding the sequence and timing of discrete events, while joint models are adept at modeling continuous biomarker trajectories and their associations with event outcomes for dynamic prediction. The complexity of real-world diseases often necessitates a more integrated approach that can leverage the strengths of both frameworks. For instance, in prostate cancer, multiple types of relapse may occur successively (a multi-state process), and their risk is influenced by the dynamics of longitudinal biomarkers like PSA.

2.3 Joint Models and Multi-State Processes

Previous research has initiated such integration, with models like the one proposed by Ferrer et al. [14], which combines a linear mixed sub-model for longitudinal data with a multi-state sub-model using shared random effects. This model enables the simultaneous analysis of *repeated measurements of a biomarker* (the longitudinal process, e.g., PSA levels) and the *times of transitions between multiple health states* (the multi-state process). It explicitly accounts for the link between these two correlated processes and uses information from the biomarker dynamics to explain or predict clinical progression events. The multi-state sub-model assumes a *non-homogeneous Markov process* for the clinical progression, meaning that the future of the process depends only on the current state (Markov property) and the time elapsed since study entry (non-homogeneous property).

Although Ferrer et al. [14] note that their joint modeling framework could be adapted to semi-Markov processes in other contexts, their primary application to prostate cancer progression assumes a non-homogeneous Markov multi-state process. However, this framework is not well suited for arbitrary transition patterns other than the ones modeled by DAGs.

While both joint models and multi-state models offer powerful tools for analyzing disease progression, they emphasize different aspects of the underlying process. Multi-state models capture the sequence and timing of discrete transitions between states, whereas joint models focus on the evolution of continuous longitudinal markers and their association with event risk. Existing joint multi-state models, such as that of

Ferrer et al. [14], are restricted to linear mixed-effects submodels and to acyclic Markov transition graphs, as their likelihood relies on a sequential factorization that is not valid when transitions form cycles. However, many biological processes involve cyclic or semi-Markov transitions and nonlinear biomarker dynamics. These cannot be handled by existing frameworks, motivating the present general formulation.

In contrast, many real-world processes—such as recurrent disease relapse or recovery—require more flexible representations involving nonlinear biomarker dynamics, semi-Markovian dependencies, and recurrent cycles. The next section introduces a general framework that integrates these features within a unified likelihood-based formulation, enabling efficient inference and dynamic prediction along complex event trajectories.

3 A General Multi-State Joint Modeling Framework

Classical joint models typically link a longitudinal biomarker process to a single time-to-event outcome. While effective for terminal events, this design cannot represent more complex disease trajectories involving intermediate or recurrent events. To address these limitations, we extend the nonlinear joint modeling framework to incorporate a multi-state process, thereby enabling a unified analysis of longitudinal biomarkers and discrete transitions between multiple health states.

The proposed model defines a joint process $\{Y_i(t), S_i(t)\}_{t\geq 0}$, where $Y_i(t)$ denotes the longitudinal biomarker trajectory and $S_i(t)$ the multi-state event process. Both components are linked through shared latent structures, and the event process is represented on an arbitrary directed graph supporting both Markovian and semi-Markovian transition mechanisms.

3.1 Notation Conventions

Throughout this section, $i = 1, \ldots, n$ indexes individuals, and t_{ij} denotes the j-th observation time of individual i. The vector $Y_{ij} \in \mathbb{R}^d$ represents longitudinal biomarker measurements, and $X_i \in \mathbb{R}^p$ the associated covariates.

The latent random effects $b_i \sim \mathcal{N}(0, Q)$ capture individual heterogeneity, while $\psi_i = f(\gamma, X_i, b_i)$ denotes the subject-specific parameters in the longitudinal submodel. The filtration $\mathcal{H}_i(t)$ represents the observed history of the longitudinal process up to time t.

Transition times and states are denoted by $(T_{i\ell}, S_{i\ell})$, with right-censoring time C_i. We also define the index of the last observed transition before time t denoted by $m_i(t) \coloneqq \sup\{\ell \geq 0 : T_{i\ell} \leq t\}$.

3.2 Extending Multi-State Joint Models

Joint models traditionally focus on a single time-to-event outcome, which limits their ability to represent complex disease trajectories involving multiple intermediate and recurrent events. To address this limitation, we extend the classical nonlinear joint modeling framework to incorporate multi-state processes, enabling the joint analysis of longitudinal biomarkers and transitions between multiple health states. This is

achieved by introducing a directed graph structure encoding all permissible transitions and supports both Markov and semi-Markov assumptions.

3.2.1 Graph Structure

Let $G = (V, E)$ be a directed graph, where $V = \{1, \dots, p\}$ denotes the set of states and $E \subset V \times V$ the set of admissible transitions. The graph encodes all possible paths of the multi-state process, allowing for competing, recurrent, or absorbing transitions. Figure 1 illustrates an example of a four-state model including an absorbing state (Death). The corresponding adjacency matrix $A = (A_{kk'})_{1 \leq k,k' \leq |V|}$ satisfies $A_{kk'} = 1$ if $(k, k') \in E$ and 0 otherwise.

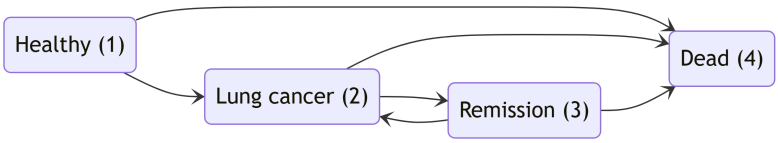


Fig. 1: Example of a 4-state transition graph $G = (V, E)$, including an absorbing state (4).

Such a representation generalizes standard joint models, which can be recovered as special cases: single-event (linear chain), competing-risks (two absorbing transitions), or recurrent-event settings (cyclic transitions) as illustrated in Figure 2.

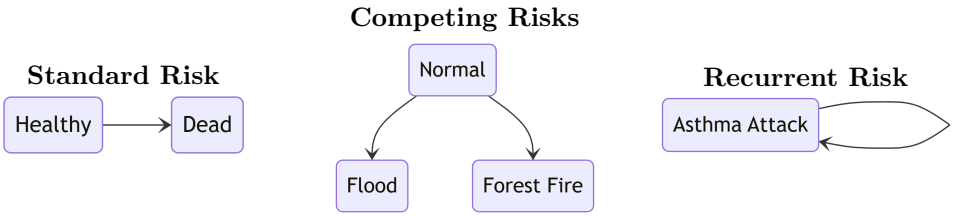


Fig. 2: Illustration of classical joint modeling approaches: (left) standard risk, (middle) competing risks, and (right) recurrent risk.

Such joint modeling offers a flexible framework that extends beyond classical approaches and is particularly useful in medical applications.

3.2.2 Individual Trajectories

Each individual i follows a latent trajectory

$$\mathcal{T}_i^* = \big((T_{i0}, S_{i0}), (T_{i1}, S_{i1}), (T_{i2}, S_{i2}), \dots\big),$$

where $T_{i\ell}$ denotes the ℓ-th transition time and $S_{i\ell} \in V$ the corresponding state. The observed trajectory is right-censored at time C_i, so that only $\{(T_{i\ell}, S_{i\ell}) : T_{i\ell} \leq C_i\}$ is observed. Between transitions, the state $S_i(t)$ is piecewise constant. We note $m_i(t) \coloneqq \sup\{\ell \geq 0 : T_{i\ell} \leq t\}$ the index of the last observed transition before time t and $m_i = m_i(C_i)$.

The process is assumed to satisfy a (possibly time-inhomogeneous) semi-Markov property

$$\mathscr{L}\big((T_{i,\ell+1}, S_{i,\ell+1}) \mid \{(T_{i\ell'}, S_{i\ell'})\}_{\ell' \leq \ell}\big) = \mathscr{L}\big((T_{i,\ell+1}, S_{i,\ell+1}) \mid (T_{i\ell}, S_{i\ell})\big), \tag{1}$$

which is relaxed compared to a stricter Markov property by allowing sojourn-time dependence.

3.2.3 Transition Intensities and Survival Functions

For any admissible transition $(k, k') \in E$, let $\lambda_i^{k'|k}(t \mid t_0)$ denote the instantaneous risk of moving from k to k' at time t, given entry time t_0

$$\lambda_i^{k'|k}(t \mid t_0) = \lim_{\delta \to 0^+} \frac{\mathbb{P}(T_{i,\ell+1} \leq t + \delta, S_{i,\ell+1} = k' \mid T_{i,\ell+1} > t, T_{i\ell} = t_0, S_{i\ell} = k)}{\delta}.$$

The cumulative intensity and the corresponding survival function are

$$\Lambda_i^{k'|k}(t \mid t_0) = \int_{t_0}^{t} \lambda_i^{k'|k}(w \mid t_0)\, dw,$$

$$\mathbb{P}(T_{i,\ell+1} > t \mid T_{i\ell}, S_{i\ell}) = \exp\left[- \sum_{s:(S_{i\ell}, s) \in E} \Lambda_i^{s|S_{i\ell}}(t \mid T_{i\ell}) \right].$$

Conditional transition probabilities are

$$\mathbb{P}(S_{i,\ell+1} = k' \mid T_{i,\ell+1}, T_{i\ell}, S_{i\ell}) = \frac{\lambda_i^{k'|S_{i\ell}}(T_{i,\ell+1} \mid T_{i\ell})}{\sum_{s:(S_{i\ell}, s) \in E} \lambda_i^{s|S_{i\ell}}(T_{i,\ell+1} \mid T_{i\ell})}.$$

3.3 Multi-State Joint Model Specification

The proposed multi-state joint model with Gaussian prior and homoscedastic Gaussian noise is specified by:

$$\begin{cases} Y_{ij} = h(t_{ij}, \psi_i) + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, R), \\ \psi_i = f(\gamma, X_i, b_i), \quad b_i \sim \mathcal{N}(0, Q), \\ \lambda_i^{k'|k}\left(t \mid X_i, T_{im_i(t)}, \mathcal{H}_i(t)\right) = \lambda_0^{k'|k}(t \mid T_{im_i(t)}) \exp\left(\alpha^{k'|k} g^{k'|k}(t, X_i, \psi_i) + \beta^{k'|k} X_i\right). \end{cases}$$

Here, h and f define the nonlinear mixed-effects submodel, $g^{k'|k}$ represents the link between biomarker dynamics and the transition intensity, and $\lambda_0^{k'|k}$ denotes the

baseline hazard associated with transition $(k \to k')$. This structure extends the classical joint model to arbitrary directed graphs and to both Markovian and semi-Markov specifications. The coefficient $\alpha^{k'|k}$ quantifies the effect of the longitudinal biomarker dynamics on the instantaneous risk of transitioning from state k to state k'. More concretely, for a one-unit increase in $g^{k'|k}(t, X_i, \psi_i)$ (holding other covariates constant), the hazard ratio for transition $(k \to k')$ is $\exp(\alpha^{k'|k})$. Likewise, $\beta^{k'|k}$ represents the effect of baseline (or time-varying) covariates X_i on that same transition: a one-unit increase in a covariate X_{ij} multiplies the hazard by $\exp(\beta_j^{k'|k})$. Thus, $\alpha^{k'|k}$ captures the degree to which the biomarker trajectory (via its latent parameter ψ_i or function $g^{k'|k}$) influences the transition risk, while $\beta^{k'|k}$ captures the direct effect of covariates on that transition's risk, beyond the biomarker pathway.

3.3.1 Model Variants

Two baseline hazard conventions are common:

$$\lambda_0^{k'|k}(t \mid T_{im_i(t)}) = \begin{cases} \lambda_0^{k'|k}(t - T_{im_i(t)}) & \text{(clock reset)}, \\ \lambda_0^{k'|k}(t) & \text{(clock forward)}. \end{cases}$$

The *clock-reset* form models risk as a function of time spent in the current state, whereas the *clock-forward* form measures risk with respect to global time since study entry.

3.3.2 Likelihood Formulation

To derive the marginal likelihood, we decompose the joint distribution of the longitudinal and multi-state processes under a set of standard conditional independence assumptions detailed in Table 3.3.2 below.

Assumptions for Likelihood Factorization

A. Latent-level independence

A1. Random effects $(b_i)_i$ are mutually independent across individuals.

B. Conditional independence within individuals

B1. Longitudinal observations $(Y_{ij})_{ij}$ are mutually independent given b_i, X_i.

B2. Trajectories $(\mathcal{T}_i^*)_i$ are mutually independent given b_i, X_i.

B3. Conditional on b_i, X_i, the longitudinal and event processes are mutually independent.

C. Censoring and process assumptions

C1. Censoring times $(C_i)_i$ are mutually independent and noninformative, *i.e.* $C_i \perp\!\!\!\perp (Y_i, \mathcal{T}_i^*) \mid b_i, X_i$.

C2. Conditionally on b_i, X_i, event trajectories satisfy the semi-Markov property 1: $p(\mathcal{T}_i^* \mid b_i, X_i) = \prod_{\ell \geq 0} p\left((T_{i,\ell+1}, S_{i,\ell+1}) \mid T_{i\ell}, S_{i\ell}, b_i, X_i\right)$.

These grouped assumptions mirror those used in classical joint modeling [31] and multi-state survival analysis [29]: independence across subjects; the conditional independence structure linking the longitudinal and event submodels through shared random effects; noninformative censoring and (semi-)Markovian dynamics.

Let $\theta = (\gamma, Q, R, \alpha, \beta)$ denote the vector of model parameters. For subject i, we observe the longitudinal measurements $Y_i = (Y_{i1}, \ldots, Y_{in_i})$ and the event trajectory $\mathcal{T}_i = \{(T_{i\ell}, S_{i\ell})_{0 \leq \ell \leq m_i}\}$. The joint likelihood then factorizes as

$$p(Y_i, \mathcal{T}_i \mid X_i, C_i, \theta) = \int p(Y_i \mid b_i, X_i, \theta)\, p(\mathcal{T}_i \mid b_i, X_i, C_i, \theta)\, p(b_i \mid \theta)\, db_i,$$

where b_i are individual random effects. This formulation generalizes the joint likelihoods of Wulfsohn and Tsiatis [38] and Ferrer et al. [14] to arbitrary multi-state event structures. Each part can then be explicitely expressed using Assumptions 3.3.2, yielding an expression very similar to that obtained by Rizopoulos [31].

Prior likelihood:

$$p(b_i \mid \theta) = \frac{(2\pi)^{-q/2}}{\det(Q)^{1/2}} \exp\left(-\tfrac{1}{2} b_i^T Q^{-1} b_i\right),$$

Longitudinal likelihood:

$$p(Y_i \mid b_i, X_i, \theta) = \prod_{j=1}^{n_i} \frac{(2\pi)^{-d/2}}{\det(R)^{1/2}} \exp\left(-\tfrac{1}{2} \left(Y_{ij} - h(t_{ij}, \psi_i)\right)^T R^{-1} \left(Y_{ij} - h(t_{ij}, \psi_i)\right)\right),$$

with $\psi_i = f(\gamma, X_i, b_i)$.

Semi-Markov likelihood:

$$\begin{aligned} p(\mathcal{T}_i \mid b_i, X_i, C_i, \theta) = \prod_{l=0}^{m_i - 1} & p\left((T_{i(l+1)}, S_{i(l+1)}) \mid b_i, X_i, (T_{il}, S_{il}), \theta\right) \\ & \exp\left(- \sum_{s:\, (S_{im_i}, s) \in E} \Lambda_i^{k' \mid S_{im_i}}(C_i \mid b_i, X_i, T_{im_i}, \theta)\right). \end{aligned} \tag{2}$$

The proof of the expression of the Semi-Markov likelihood (2) is provided in Appendix B.

The likelihood factorization above is valid and forms the basis for scalable inference procedures described in Section 4, leveraging the complete trajectory of biomarkers to refine predictions of transitions and survival [10].

Moreover, here, we assume that the initial state S_{i0} is observed. However, the framework could also incorporate a multinomial model for unobserved initial states [39].

3.4 Model Selection and Information Criteria

To compare competing model specifications, we rely on information criteria based on the marginal likelihood as defined in Section 4. The Akaike Information Criterion (AIC) [1] and the Bayesian Information Criterion (BIC) [36] are standard tools. While the AIC relies on an asymptotically unbiased estimator of the log-likelihood, the BIC aims to identify the true model with great probability as the sample size grows.

The Akaike Information Criterion is computed as follows

$$\mathrm{AIC} = -2 \log \mathcal{L}_{\mathrm{marginal}}(\hat{\theta}; X, Y, \mathcal{T}, C) + 2k,$$

where k is the number of parameters, and $\hat{\theta}$ is a maximizer of the marginal likelihood.

The Bayesian Information Criterion is similarly defined but imposes a stronger penalty for model complexity, growing in $\log n$ where n is the number of observations. Particularly in the context of mixed effects models, and even more so for multi-state joint models, the number of observations n may vary from one definition to another, either the total number of repeated measurements or the number of individuals [9]. To alleviate this problem, we recall the derivation of the BIC from the Laplace approximation [22]. For a prior distribution π on a set of models $\mathcal{M}$, the Laplace approximation approaches up to constant terms which could also be accounted to improve the approximation, the posterior probability given observed data x defined in Section 4 by

$$\forall m \in \mathcal{M},\ \log p(x \mid m) \approx \log \mathcal{L}_{\mathrm{marginal}}(\hat{\theta}; x) - \frac{1}{2} \log \det(-H_{\hat{\theta}}),$$

where $H_{\hat{\theta}}$ denotes the Hessian matrix of the marginal log-likelihood evaluated at the Maximum Likelihood Estimator (MLE). Then, under standard regularity conditions that allow differentiation and integration to be interchanged twice, we have that

$$-\frac{1}{n} H_{\hat{\theta}} \xrightarrow{\mathbb{P}} \mathcal{I}(\hat{\theta}),$$

where $\mathcal{I}(\hat{\theta})$ denotes the Fisher Information Matrix [6]. Given a reliable estimate $\widehat{\mathcal{I}}(\hat{\theta})$ of the Fisher Information Matrix (see, e.g., Delattre and Kuhn [8]), we can then approximate the BIC by substituting this estimate

$$\forall m \in \mathcal{M},\ \log p(x \mid m) \approx \log \mathcal{L}_{\mathrm{marginal}}(\hat{\theta}; x) - \frac{1}{2} \log \det \hat{\mathcal{I}}_n(\hat{\theta}),$$

where $\hat{\mathcal{I}}_n(\hat{\theta}) = n\, \hat{\mathcal{I}}(\hat{\theta})$, which in practice corresponds to the matrix readily computed by our software.

Thus, our BIC criterion can be written as

$$\mathrm{BIC}_{\mathcal{I}} = -2 \log \mathcal{L}_{\mathrm{marginal}}(\hat{\theta}; x) + \log \det \hat{\mathcal{I}}_n(\hat{\theta}).$$

In both cases, one should aim at minimizing the chosen criterion.

4 Statistical Inference

The estimation of model parameters $\theta = (\gamma, Q, R, \alpha, \beta)$ relies on two likelihood formulations: the complete-data likelihood $\mathcal{L}_{\text{complete}}$, which includes latent variables, and the marginal likelihood $\mathcal{L}_{\text{marginal}}$, obtained by integrating them out

$$\mathcal{L}_{\text{marginal}}(\theta; X, Y, \mathcal{T}, C) = \int \mathcal{L}_{\text{complete}}(\theta; X, Y, \mathcal{T}, C, b)\, db,$$

where

$$\mathcal{L}_{\text{complete}}(\theta; X, Y, \mathcal{T}, C, b) = \prod_{i=1}^{n} p(Y_i \mid b_i, X_i, \theta)\, p(\mathcal{T}_i \mid b_i, X_i, C_i, \theta)\, p(b_i \mid \theta).$$

It is to be noted that this integral is typically intractable in closed form due to the nonlinear nature of the model, necessitating numerical approximation methods such as Monte Carlo integration or Laplace approximation.

Several optimization methods adapted from the nonlinear mixed-effects literature can be applied, including Stochastic EM [21], Laplace approximation [37], Gauss-Hermite quadrature, stochastic gradient ascent with Robbins–Monro updates [34], and MCMC-based approximations such as Metropolis-Hastings and Hamiltonian Monte Carlo. These strategies are implemented in software such as **JMBayes** [33, 32].

Another approach requiring mild regularity assumptions on the log marginal likelihood, but without the need for the model to belong in the exponential family, is to consider a stochastic gradient ascent scheme [5, 4] using Fisher's identity and following the Robbins-Monro procedure [34].

Under interchangeability of integral and differentiation, setting $x = (X, Y, \mathcal{T}, C)$ for convenience, the Fisher identity writes

$$\nabla_\theta \log \mathcal{L}_{\text{marginal}}(\theta; x) = \mathbb{E}_{b \sim p(\cdot \mid x, \theta)} \left(\nabla_\theta \log \mathcal{L}_{\text{complete}}(\theta; x, b) \right).$$

This expectation is approximated using Monte Carlo samples from the posterior $p(\cdot \mid x, \theta)$, avoiding the need to evaluate the intractable marginal likelihood $\mathcal{L}_{\text{marginal}}(\theta; x)$. This formulation naturally supports stochastic gradient ascent with MCMC-based posterior sampling and is amenable to minibatch parallelization across individuals, ensuring the convergence to a critical point. The update rule is as follows:

Algorithm 1 Stochastic gradient ascent model inference

Input: data x; initial parameter $\theta^{(0)}$; step sizes $(\eta_t)_{t\geq 0}$ with $\sum_t \eta_t = +\infty$, $\sum_t \eta_t^2 < +\infty$; batch size K; sampler p_t^{MCMC}; stopping criterion.

while not converged **do**
  Draw $b_k \sim p(\cdot \mid x, \theta_n) \approx p_n^{\text{MCMC}}$.
  $\theta^{(t+1)} \leftarrow \theta^{(t)} + \frac{\eta_t}{K} \sum_{k=1}^{K} \nabla_\theta \log \mathcal{L}_{\text{complete}}(\theta^{(t)}; x, b_k)$.
end while
return $\hat{\theta}$

Preconditioning matrices may also be used, such as the Fisher Information Matrix in a natural gradient descent framework, or even specialized *optimizers* such as Adam [19], NAdam [12] or Adagrad [13]. In particular, the Adam optimizer will later be used in our numerical experiments.

5 Dynamic prediction

The simulation of individual trajectories using Algorithm 2 enables prediction in the multi-state joint modeling framework. Specifically, predictions are derived from simulated event trajectories, allowing for individualized risk assessment and forecasting of future states or event times based on a subject's observed biomarker and event history. This approach generalizes classical dynamic prediction by leveraging the rich structure of multi-state processes and the joint distribution of longitudinal and event data.

Suppose we are interested in some functional of the true (unobserved) future trajectory,

$$\chi^*(\mathcal{T}_i^*) \in \mathbb{Y},$$

where $\mathbb{Y}$ is some generic space.

For example, one can think of the state taken by the individual i at a time u. As is the case in traditional dynamic prediction however, we are only interested in predicting quantities or characteristics that depend on a yet unobserved *future*, given prior information, *i.e.* the trajectory up to some prediction time $t \leq u$ and observed longitudinal markers. Contrary to (single transition) joint models, the conditional probability distribution on $(\mathbb{R} \times \mathcal{S})^{\mathbb{N}}$ cannot be analytically derived.

Nonetheless, as seen in Algorithm 2, we are able to accurately simulate each trajectory. Therefore, a Monte Carlo estimation scheme may be devised under certain restrictions, where we estimate $\chi^*(\mathcal{T}_i^*)$ by its expected value:

$$\hat{\chi}_i = \mathbb{E}_{\mathcal{T}_i' \sim \mathscr{L}(\mathcal{T}_i^* \mid \mathcal{Y}(t), \mathcal{T}_i, t)} \left(\chi^*(\mathcal{T}_i')\right) \approx \frac{1}{B} \sum_{k=1}^{B} \chi^*(\mathcal{T}_i^{(k)}),\ \mathcal{T}_i^{(k)} \overset{\text{i.i.d.}}{\sim} \mathscr{L}(\mathcal{T}_i^* \mid \mathcal{Y}(t), \mathcal{T}_i, t).$$

In general however, χ^* may very well depend on (countably) infinitely many transitions. As a result, since the proposed estiamtion relies on simulted samples, we require the quantity to depend only on a finite subset of transitions.

Assumption 1. *Assume there exists* $\chi : \bigcup_{n\geq 1}(\mathbb{R}\times\mathbb{S})^n \to \mathbb{Y}$ *and* τ_i *a stopping time for the filtration* $\mathcal{F}_{in} = \sigma\left((T_{il}, S_{il})_{0\leq l\leq n}\right)$ *such that:*

$$\begin{cases} \tau_i < +\infty \ \textit{a.s.}, \\ \chi^*(\mathcal{T}_i^*) = \chi\left((T_{il}, S_{il})_{1\leq l\leq \tau_i}\right) \end{cases}$$

Essentially, given Assumption 1, with probability one we are guaranteed that the quantity of interested may be computed in a finite number of simulation steps. Indeed, the prediction algorithm for a single sample may be summarized as follows:

Input: $t \in \mathbb{R}$, a prediction time; $\mathcal{Y}(t)$ marker history; $\mathcal{T}_i$ trajectory up to time t; χ and stopping time τ_i.

1: $n \leftarrow 0$
2: **while** $n < \tau_i$ **do**
3: Append simulated (T_{in}, S_{in}) to $\mathcal{T}_i$ if $n > \text{len}(\mathcal{T}_i)$
4: $n \leftarrow n + 1$
5: **end while**
6: **return** $\chi(\mathcal{T}_i)$

Numerous quantities of interest may be encompassed by this framework. We give multiple examples below. First, for a directed acyclic graph $G = (V, E)$, we define the *depth* of G, denoted by $\text{depth}(G)$, as the length of the longest directed path in G, that is,

$$\text{depth}(G) = \max_{(v_0, v_1, \ldots, v_k)} k,$$

where $(v_0, v_1, \ldots, v_k)$ ranges over all directed paths in G.

Example 1 (State at time u)**.** *Let* $G = (V, E)$ *be a directed acyclic graph,* $u \in \mathbb{R}$ *be a fixed time, and* $\mathbb{Y} = V$ *be the set of possible states. Let* $\tau_i = \inf\{n \in \mathbb{N} : T_{in} \geq u \text{ or } S_{in} \text{ is absorbing}\}$*. Clearly,* $\tau_i \leq \text{depth}(G) < +\infty$*. Then, take* $\chi_u^*(\mathcal{T}_i^*) = S_{i\tau_i}$*, which corresponds to the state of the individual* i *at time* u*. In particular, if* $V = \{0, 1\}$ *and* $E = \{(0, 1)\}$*, we recover the special case used for survival probability estimation in standard joint models [30].*

Example 2 (Hitting time)**.** *Let* $G = (V, E)$ *be a directed acyclic graph,* $\mathbb{A} \subset V$ *a non-empty subset of states, and* $\mathbb{Y} = \mathbb{R} \cup \{+\infty\}$*. For any two non-empty sets* $\mathbb{A}, \mathbb{B} \subset V$*, we note* $\mathbb{A} \rightsquigarrow \mathbb{B}$ *if there exists a path from* $\mathbb{A}$ *to* $\mathbb{B}$ *in* G*. Let* $\tau_i = \inf\{n \in \mathbb{N} : S_{in} \in \mathbb{A} \vee \{S_{in}\} \not\rightsquigarrow \mathbb{A}\}$*. Then,* $\tau_i \leq \text{depth}(G) < +\infty$ *and take* $\chi_{\mathbb{A}}^*(\mathcal{T}_i^*) = T_{i\tau_i} + \mathbb{1}_{\{S_{in}\} \not\rightsquigarrow \mathbb{A}}(+\infty)$ *represents the hitting time for the set* $\mathbb{A}$*.*

In Example 1, the stopping time τ_i captures the step at which the process for individual i reaches or exceeds a given time u, or enters an absorbing state. The

resulting value $\chi^*(\mathcal{T}_i^*) = S_{i\tau_i}$ therefore represents the state of the individual at that specific time.

In Example 2, the stopping time τ_i corresponds to the first time the process reaches a target subset of states $\mathbb{A}$, or becomes unable to reach it in the future. The associated value $\chi^*_{\mathbb{A}}(\mathcal{T}_i^*)$ thus represents the *hitting time* of the set $\mathbb{A}$, which may be finite if $\mathbb{A}$ is reached, or $+\infty$ otherwise.

Other practical applications may include finding the expected number of edges in some given trajectory, the number of times an individual has returned in a specific state, the expected time between transitions...

This simulation-based dynamic prediction framework generalizes classical approaches from single-event joint models to the multi-state context. It enables individualized forecasting of future state occupancy, event risks, sojourn times, and other clinically relevant outcomes, fully exploiting the subject's observed biomarker trajectory and event history. In Section 7, we demonstrate its application to forecasting dependency trajectories in the PAQUID cohort.

6 Simulation Study

The purpose of this simulation study is to evaluate the finite-sample performance, identifiability, and convergence properties of the proposed inference algorithm under controlled conditions where the true parameters are known. We simulate data from a three-state semi-Markov process coupled with a nonlinear longitudinal biomarker trajectory, representing a simplified disease progression model. This setup allows us to assess both statistical accuracy and computational scalability of the stochastic-gradient estimation procedure.

All simulations, estimations, and figures reported in this section were produced using the open-source `jmstate` Python package, available on PyPI, which we developed to implement the multi-state joint modeling framework and inference algorithms described in Sections 3.2 and 4.

6.1 Model Specification

The simulated model involves three states—Healthy (0), Sick (1), and Terminal (2)—connected by the transitions $0 \to 1$, $0 \to 2$, and $1 \to 2$ (Figure 3). This configuration captures the typical monotone evolution of a disease process, with both direct and indirect paths to the terminal state. Transition intensities follow an exponential baseline hazard under a clock-reset convention.

The longitudinal process follows a continuous piecewise-affine regression:

$$h(t, \psi) = \psi_1 + \psi_2 t + \mathbb{1}_{t>\tau}(\psi_3 - \psi_2)(t - \tau),$$

with a known inflection time $\tau = 6$. The individual-specific parameters $\psi_i = (\psi_{i1}, \psi_{i2}, \psi_{i3})$ arise from a nonlinear mixed-effects model with Gaussian random effects $b_i \sim \mathcal{N}(0, Q)$ and independent Gaussian noise $\epsilon_{ij} \sim \mathcal{N}(0, R)$. The link function

combines the current biomarker value and its slope,

$$g(t, X_i, \psi_i) = \big(h(t, \psi_i), \frac{\partial h}{\partial t}(t, \psi_i)\big),$$

shared across transitions. Covariates X_i are normally distributed and enter linearly through $\beta^{k'|k} X_i$.

Longitudinal data were collected at up to 20 random time points per subject and censored at $C_i \sim \mathcal{U}([10, 15])$. Censoring also truncates the biomarker trajectories. The true parameter vector lies in $\mathbb{R}^{16}$. Each simulated dataset comprised $n = 1000$ individuals, generating approximately balanced transition counts (Figure 4).

This trajectory mimics a biomarker whose evolution slows after treatment or onset of symptoms, while preserving continuity at the change point.

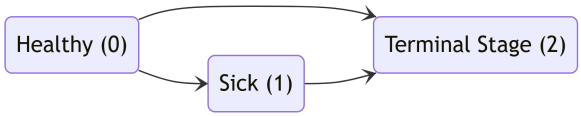


Fig. 3: State transition diagram of the simulated three-state model.

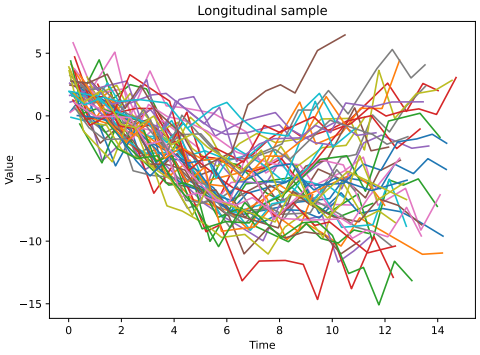
6.2 Estimator Convergence and Accuracy

Parameter estimation was performed using the stochastic-gradient inference procedure of Section 4, with the Adam optimizer (learning rate 0.5, batch size $K = 100$) and an adaptive stopping criterion based on the first and second moments of parameter differences. Convergence was typically achieved in 400–500 iterations and required around 12 seconds per run on an AMD Ryzen 5 processor.

Across $n_{\text{runs}} = 100$ independent replications, the mean relative bias of all parameters remained below 2%, and the root-mean-square error (RMSE) scaled approximately as $n^{-1/2}$. Parameters governing the biomarker trajectory (γ, Q, R) were consistently well identified, while association coefficients $(\alpha^{k'|k}, \beta^{k'|k})$ exhibited slightly higher variability for rare transitions. Representative convergence trajectories are shown in Figure 10, and quantitative results are summarized in Table 5.

Furthermore, longitudinal measurements were taken at $m_i \leq 20$ time points, and censoring times were drawn from $C_i \sim \mathcal{U}([10, 15])$. Longitudinal observations were also truncated at the censoring times.

A short summary of the longitudinal process as well as the trajectories is given by the Figure 4 below.



From → To	# Observations
$0 \to 1$	613
$0 \to 2$	375
$1 \to 2$	592

Fig. 4: Simulated data: on the left, a sample of longitudinal measurements from 50 individuals; on the right, observed transitions between states for the complete population of 1000 individuals.

6.3 Estimator Convergence

To illustrate the convergence of the estimator, we examine the optimization process for a particular run with $n = 1000$ (see Figure 10), as well as RMSE values computed from $n_{\text{runs}} = 100$ independent runs on different simulated datasets, all generated from the same underlying distribution. The results of these simulations are shown in Table 5.

To fit the model, we first have to define initial parameters, which can be particularly important for the optimization process if the model is not well-behaved. In the present case, without prior knowledge of the true parameters, we zero the values and use the identity matrix for both covariance matrices.

Formally:

$$m_i^{(t)} \leftarrow \beta_i m_i^{(t-1)} + (1-\beta_i)(\theta^{(t)} - \theta^{(t-1)})^i, \quad \hat{m}_i^{(t)} = \frac{m_i^{(t)}}{1-\beta_i^t},$$

and the optimization process is terminated when

$$|\hat{m}_1^{(t)}| \leq 10^{-6} + 10^{-1}\sqrt{\hat{m}_2^{(t)}}.$$

The Figure 10 shows the evolution of the parameters during the optimization process, and Table 5 summarizes the accuracy of the inferred parameters.

Parameter	True value	Mean	Standard error	RMSE
γ_1	2.5000	2.4960	0.0370	0.0372
γ_2	-1.3000	-1.3039	0.0257	0.0260
γ_3	0.2000	0.1936	0.0267	0.0275
$\tilde{Q}_1$	0.2554	0.2243	0.0442	0.0540
$\tilde{Q}_2$	0.8047	0.8009	0.0260	0.0263
$\tilde{Q}_3$	0.6020	0.6001	0.0276	0.0277
$\tilde{R}$	-0.2653	-0.2684	0.0083	0.0089
$\alpha_1^{1\|0}$	-0.5000	-0.49997	0.0486	0.0486
$\alpha_2^{1\|0}$	-3.0000	-2.9962	0.0800	0.0801
$\alpha_1^{2\|0}$	-1.0000	-0.9990	0.0847	0.0847
$\alpha_2^{2\|0}$	-5.0000	-4.9897	0.1191	0.1196
$\alpha_1^{2\|1}$	0.0000	0.0011	0.0327	0.0327
$\alpha_2^{2\|1}$	-1.2000	-1.2045	0.0429	0.0432
$\beta^{1\|0}$	-1.3000	-1.3015	0.0501	0.0501
$\beta^{2\|0}$	-0.9000	-0.8982	0.0670	0.0670
$\beta^{2\|1}$	-0.7000	-0.6981	0.0551	0.0551

Fig. 5: Comparison of true and estimated parameters (averaged over 100 runs).

The simulation confirms that the proposed inference method accurately recovers all parameters of the multi-state joint model under realistic sample sizes. The optimization demonstrates stable convergence (Figure 10), with minimal sensitivity to initialization. The algorithm's computational efficiency allows for extensive simulation or cross-validation studies at negligible cost.

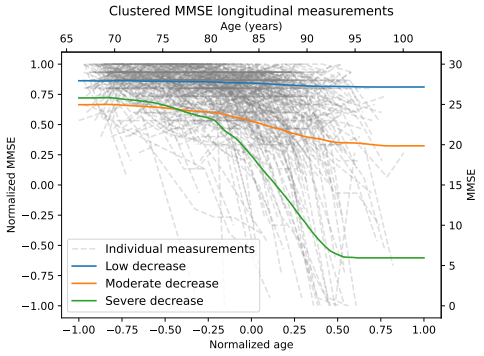
7 Application to the PAQUID Cohort

7.1 Introduction

The data used in this section originates from the PAQUID cohort [23], a large prospective population-based study initiated in southwestern France in 1988, aimed at understanding the determinants and trajectories of aging. A subsample of $n = 500$ individuals was followed over a period of up to 20 years [28], with repeated measures of cognitive and physical health, as well as socio-demographic characteristics. In particular, global cognitive functioning was assessed through the Mini-Mental State Examination (MMSE, see Figure 6), while physical dependency was evaluated using the HIER scale, which classifies subjects into four ordered states (see Figure 7): 0 (no dependency), 1 (mild dependency), 2 (moderate dependency), and 3 (severe dependency). In this work, we focus on the association between cognitive decline and the progression of physical dependency by jointly modeling the longitudinal trajectory of MMSE scores and the transitions between the four HIER states using a joint multi-state model. More specifically, the state of an individual i at a given time t is defined as the highest HIER dependency score between trial entry and t. This approach allows

us to characterize the dynamic interrelationship between cognitive and physical aging processes within a unified statistical framework.

Both time and longitudinal values were normalized in $[-1, 1]$. This technique ensured stable and fast convergence of the optimization process by keeping all parameters on roughly the same scale.



From → To	# Observations
$0 \to 1$	85
$1 \to 2$	135
$2 \to 3$	86
$0 \to 2$	44
$1 \to 3$	15
$0 \to 3$	11

Fig. 6: On the left, each light gray line represents one of the 500 observed sequences of (normalized) MMSE scores, and three functional clusters are also represented; on the right, observed transitions between states.

Therefore, the chosen transition graph is as follows, and only allows monotonic transitions from a lower level of physical dependency to a higher one.

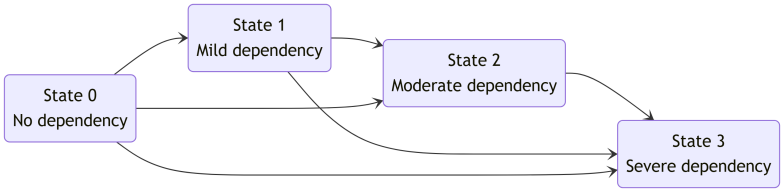


Fig. 7: Schematic representation of the four ordered HIER states, with possible transitions. HIER quantifies physical dependency (0: no dependency to 3: severe dependency) [28].

We consider an exponential baseline hazard model for transitions, with a *clock reset* specification. The parameters of these exponential baseline hazards are jointly optimized with the model parameters.

Moreover, we impose a particular substructure on the model, where $\forall (k, k') \in E$, $\beta^{k'|k} = \beta$.

In the light of Figure 6, the regression and link functions were taken to be scaled and shifted hyperbolic tangent and its derivative respectively such that the normalized

MMSE score always has a value of 1 when $t \to -\infty$ and is not increasing

$$\begin{cases} h(t,\psi) = \psi_1 \tanh\left(\frac{\psi_3 - t}{\psi_2}\right) + (1-\psi_1), \\ g(t,\psi) = \frac{\partial}{\partial t} h, \\ \psi = \begin{pmatrix} \sigma(\gamma_1 + b_1) \\ \exp(\gamma_2 + b_2) \\ \gamma_3 + b_3 \end{pmatrix}, \quad \sigma(z) = \frac{1}{1+e^{-z}}. \end{cases}$$

At last, an individual covariates correspond to a pair of binary variables indicating wheter or not the individual has a diploma and is a male $(\mathbb{1}_{\text{CEP}}(i), \mathbb{1}_{\text{male}}(i)) \in \{0,1\}^2$.

7.2 Results

7.2.1 Inference

Fitting was performed using 80% of the data, with the remaining 20% used for testing. The model was fitted using the Adam optimizer. The optimization process was run until convergence based on the same criterion as the simulation study with a tolerance of 5%, which was obtained after a little less than 500 iterations with 10 parallel chains.

Before displaying the fitted parameters, the Fisher Information Matrix was computed, implementing the method described in [4], directly taking into account shared parameters. The standard errors were then computed as the square root of the diagonal of the inverse of the Fisher Information Matrix. The results are summarized in Table 8.

Parameter	Inferred value	Std. error	Parameter	Inferred value	Std. error
γ_1	**2.1600**	**0.0470**	γ_2	**0.1590**	**0.0140**
γ_3	**1.0010**	**0.0030**	$\tilde{Q}_1$	**-0.5700**	**0.0430**
$\tilde{Q}_2$	**0.9570**	**0.0960**	$\tilde{Q}_3$	**0.3820**	**0.0430**
$\tilde{Q}_4$	-0.2050	0.1570	$\tilde{Q}_5$	**-2.3310**	**0.1380**
$\tilde{Q}_6$	**0.5220**	**0.0470**	$\tilde{R}$	**2.0830**	**0.0260**
$\alpha^{3\|1}$	0.7220	0.7150	$\alpha^{1\|0}$	-0.5400	0.4030
$\alpha^{2\|1}$	**-1.0030**	**0.2000**	$\alpha^{2\|0}$	**-0.1210**	**0.0380**
$\alpha^{3\|2}$	-0.0480	0.2520	$\alpha^{3\|0}$	**-0.3070**	**0.0740**
β_1	-0.2990	0.9130	β_2	0.0970	1.0300

Fig. 8: Inferred parameters and their estimated standard deviations. Bold values mean $p < 0.05$.

Notably, the estimated standard errors are consistent with the number of observations per transition, with larger standard errors for transitions with fewer observations. Therefore, parameters with too few observations should be interpreted with caution, if at all. The linear covariate parameters also exhibit large standard errors, which indicates the coefficients are not significative in this setting.

The association parameters α being negative, we can infer that cognitive decline (*i.e.* rate of change) is associated with an increase in physical dependency.

7.2.2 Dynamic Prediction

Using the $n_{\text{test}} = 100$ individuals not included in the model fitting, we performed dynamic prediction. For each individual, longitudinal measurements and trajectories were truncated at various times t, and predictions were made for future time points $u \geq t$ beyond the truncation. Accuracy was then assessed by comparing the most likely predicted state with the true state at each corresponding future time point u, as illustrated in Figure 9.

Formally, the prediction function χ_u^* is defined by Example 1, and the accuracy measure at time t for predicted states $(s_i)_i$ is defined as

$$\text{accuracy}(u) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}_{\chi_{u \wedge C_i}(\mathcal{T}_i) = s_i}.$$

Although $\mathcal{T}_i^*$ is unkown, the quantity $\chi_{u \wedge C_i}(\mathcal{T}_i)$ may be computed based on the censored trajectories. However, incorporating censoring times into prediction functions relies on information that is, in principle, not available at prediction time. Nevertheless, this inclusion is necessary to ensure fairness in accuracy metrics, since individuals who are not censored are generally more likely to have progressed to more advanced states of physical dependency than those who are censored.

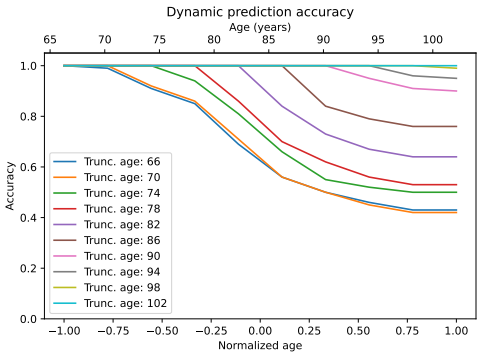


Fig. 9: Accuracy of the predicted state at each time point with respect to the true state for different truncation times.

Here, the initial drops correspond to the truncation times, before which the accuracy is always 100%.

As expected from dynamic predictions, the accuracy increases as more data is available, therefore when the truncation time grows. We also observe that the longer the prediction horizon, the lower the accuracy of the predictions.

8 Conclusion

We presented a general likelihood-based framework for joint modeling of nonlinear longitudinal biomarkers and multi-state survival processes. The model unifies both

components on arbitrary directed graphs, extending classical joint models to accommodate Markov and semi-Markov structures, recurrent transitions, and nonlinear mixed-effects submodels. This formulation provides a flexible and rigorous foundation for linking longitudinal biomarker trajectories to complex event histories within a single coherent probabilistic model.

Simulation studies confirmed robust parameter recovery, and an application to the PAQUID cohort highlighted the ability to capture the interplay between cognitive decline and physical dependency.

Since the model may potentially involve many transitions and therefore a large number of parameters, a strategy of parameter sharing across transitions could help reduce the overall number of parameters. This is especially beneficial in settings with limited data for certain transitions, or when biological or clinical knowledge suggests similar effects across multiple transitions. Parameter sharing improves statistical efficiency, reduces overfitting, and facilitates model interpretability.

Future directions include a spline-based parametrization of baseline hazards to allow for nonparametric estimation. This approach is particularly useful when the true baseline hazard is expected to be complex or non-monotonic, or when parametric forms (such as exponential or Weibull) may be too restrictive and lead to model misspecification.

Another promising direction is the development of efficient visualization tools for multi-state trajectories and dynamic predictions. Interactive representations of individual event histories, transition probabilities, and longitudinal biomarker evolution would greatly facilitate model interpretation and communication of results to applied researchers.

Funding

This work was partially funded by the Stat4Plant project ANR-20-CE45-0012.

Acknowledgments

The authors would like to thank Jean-Benoist Leger for helpful discussions.

Conflict of interest

The authors declare that they have no conflicts of interest.

Data availability

The data analyzed in this study are publicly available via the R package `timeROC` (dataset `Paquid`).

References

[1] Akaike H (2003) A new look at the statistical model identification. IEEE transactions on automatic control 19(6):716–723

[2] Andersen P, Keiding N (1999) Multi-state models for event history analysis. Statistical Methods in Medical Research 8(2):127–153. https://doi.org/10.1177/096228029900800202
[3] Asanjarani A, Liquet B, Nazarathy Y (2022) Estimation of semi-markov multi-state models: A comparison of the sojourn times and transition intensities approaches. The International Journal of Biostatistics 18(1):243–262
[4] Baey C, Delattre M, Kuhn E, et al (2023) Efficient preconditioned stochastic gradient descent for estimation in latent variable models. In: Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 202. PMLR, pp 1430–1453
[5] Caillebotte A, Kuhn E, Lemler S (2025) Estimation and variable selection in high dimension in nonlinear mixed-effects models. arXiv preprint arXiv:250320401 https://doi.org/10.48550/arXiv.2503.20401, version 2 (Aug 2025)
[6] Casella GC (2001) Theory of point estimation. Springer
[7] Commenges D, Joly P, Proust-Lima C, et al (2006) Likelihood-based approaches to illness–death models. Biostatistics 7(4):544–556
[8] Delattre M, Kuhn E (2023) Computing an empirical fisher information matrix estimate in latent variable models through stochastic approximation. Computo
[9] Delattre M, Lavielle M, Poursat MA (2014) A note on bic in mixed-effects models
[10] de Wreede L, Fiocco M, Putter H (2010) The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. Computer Methods and Programs in Biomedicine 99(3):261–274. https://doi.org/10.1016/j.cmpb.2010.01.001
[11] Dong W, Herring A, Dunson D (2008) Non-markovian modeling of sleep patterns. Biostatistics 9(4):806–820
[12] Dozat T (2016) Incorporating nesterov momentum into adam. https://openreviewnet/pdf?id=OM0jvwB8jIp57ZJjtNEZ
[13] Duchi J, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research 12(7)
[14] Ferrer L, Rondeau V, Dignam J, et al (2016) Joint modelling of longitudinal and multi-state processes: Application to clinical progressions in prostate cancer. Statistics in Medicine 35(22):3933–3948
[15] Fiocco M, Putter H, van Houwelingen H (2008) Modeling non-exponential transition times in multi-state models. Statistics in Medicine 27(30):5566–5581
[16] Frydman H (2005) Estimation in the cox model with misclassified covariates using multistate semi-markov models. Biometrics 61(3):847–854
[17] Gillespie DT (2007) Stochastic simulation of chemical kinetics. Annu Rev Phys Chem 58(1):35–55
[18] Jackson C (2011) Multi-state models for panel data: The msm package for R. Journal of Statistical Software 38(8):1–29. https://doi.org/10.18637/jss.v038.i08
[19] Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980
[20] Król A, Mauguen A, Mazroui Y, et al (2017) Tutorial in joint modeling and prediction: A statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. Journal of Statistical Software 81(3):1–52. https:

//doi.org/10.18637/jss.v081.i03
[21] Kuhn E, Lavielle M (2004) Coupling a stochastic approximation version of em with an mcmc procedure. ESAIM: Probability and Statistics 8:115–131
[22] Lebarbier É, Mary-Huard T (2006) Une introduction au critère bic: fondements théoriques et interprétation. Journal de la Société française de statistique 147(1):39–57
[23] Letenneur L, Commenges D, Dartigues JF, et al (1994) Incidence of dementia and Alzheimer's disease in elderly community residents of southwestern france. International Journal of Epidemiology 23(6):1256–1261
[24] Limnios N, Oprisan G (2001) Semi-Markov Processes and Reliability. Statistics for Industry and Technology, Birkhäuser
[25] Lovblom L, Briollais L, Perkins B, et al (2024) Modeling multiple correlated end-organ disease trajectories: A tutorial for multistate and joint models with applications in diabetes complications. Statistics in Medicine 43(5):1048–1082. https://doi.org/10.1002/sim.9984
[26] Luciano E, Vasiliev V (2006) Multistate semi-markov models in finance. Decisions in Economics and Finance 29(1):1–22
[27] Papageorgiou G, Mauff K, Tomer A, et al (2019) An overview of joint modeling of time-to-event and longitudinal outcomes. Annual Review of Statistics and Its Application 6:223–240. https://doi.org/10.1146/annurev-statistics-030718-104209
[28] Proust-Lima C, Philipps V, Liquet B (2017) Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. Journal of Statistical Software 78(2):1–56. https://doi.org/10.18637/jss.v078.i02
[29] Putter H, Fiocco M, Geskus R (2007) Tutorial in biostatistics: Competing risks and multi-state models. Statistics in Medicine 26(11):2389–2430. https://doi.org/10.1002/sim.2712
[30] Rizopoulos D (2011) Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics 67(3):819–829
[31] Rizopoulos D (2012) Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Chapman and Hall/CRC
[32] Rizopoulos D (2020) JMbayes: Joint Models for Longitudinal and Time-to-Event Data. R package JMbayes, CRAN
[33] Rizopoulos D, Miranda Afonso P, Papageorgiou G (2024) JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data. https://doi.org/10.32614/CRAN.package.JMbayes2, URL https://CRAN.R-project.org/package=JMbayes2, r package version 0.5-2
[34] Robbins H, Monro S (1951) A stochastic approximation method. The Annals of Mathematical Statistics 22(3):400–407
[35] Roever C, Meyer R, Kuchler U (2010) Statistical modelling of sleep stages: A semi-markov model approach. Computational Statistics & Data Analysis 54(8):2039–2051
[36] Schwarz G (1978) Estimating the dimension of a model. The annals of statistics pp 461–464

[37] Wolfinger R (1993) Laplace's approximation for nonlinear mixed models. Biometrika 80(4):791–795
[38] Wulfsohn M, Tsiatis A (1997) A joint model for survival and longitudinal data measured with error. Biometrics 53(2):330–339
[39] Yiu S, Farewell V, Tom B (2018) Clustered multistate models with observation level random effects, mover–stayer effects and dynamic covariates: Modelling transition intensities and sojourn times in a study of psoriatic arthritis. Journal of the Royal Statistical Society Series C: Applied Statistics 67(2):481–500

A Multi-State Simulation

In nonlinear joint models with known design and known parameters parameters $\theta = (\gamma, Q, R, \alpha, \beta)$, the occurrence times of events conditionally on the latent variables b_i and the covariates X_i can be easily sampled using the inverse cumulative distribution function transform method. This may be achieved through bisection or other root-finding algorithms.

The simulation of the transition process $\mathcal{T}_i^*$ conditionally on b_i, X_i, and (T_{i0}, S_{i0}) can be achieved drawing on the semi-Markov property by considering one transition at a time, according to Algorithm 2. The algorithme is similar to Gillespie's algorithm [17], and can also include a survival condition $T_{i1} \geq t_i^{\text{surv}}$ that uses the Chasles relation.

Algorithm 2 Simulation of trajectory i

Input: subject-specific censoring time $C_i \in \mathbb{R} \cup \{+\infty\}$; covariates X_i; latent variables b_i; initial time-state pair (T_{i0}, S_{i0}); optional survival condition $T_{i1} \geq t_i^{\text{surv}}$, defaults to $t_i^{\text{surv}} = -\infty$.

Initialize $\ell_i \leftarrow 1$, $\mathcal{T}_i \leftarrow ((T_{i0}, S_{i0}))$.

while $T_{i(\ell_i-1)} < C_i$ and $\{s : (S_{i(\ell_i-1)}, s) \in E\} \neq \emptyset$ **do**
  for each s with $(S_{i(\ell_i-1)}, s) \in E$ **do**
    Draw $T^s_{i\ell_i}$ such that
    $-\log \mathbb{P}(T^s_{i\ell_i} > t) = \int_{T_{i(\ell_i-1)} \vee t_i^{\text{surv}}}^{t} \lambda_i^{s|S_{i(\ell_i-1)}}(w \mid b_i, X_i, T_{i(\ell_i-1)}, \theta)\, dw$,
    $\forall t \geq T_{i(\ell_i-1)} \vee u$.
  end for

  Set $T_{i\ell_i} \leftarrow \min_m T^s_{i\ell_i}$, and $S_{i\ell_i} \leftarrow \arg\min_m T^s_{i\ell_i}$.
  Append $(T_{i\ell_i}, S_{i\ell_i})$ to $\mathcal{T}_i$.
  $\ell_i \leftarrow \ell_i + 1$.
end while

if $T_{i(\ell_i-1)} > C_i$ **then**
  Remove the last pair: $\mathcal{T}_i \leftarrow \mathcal{T}_i[:-1]$.
end if
return $\mathcal{T}_i$

The proof of the exactness of this algorithm is given below.

Proof First consider the case $t_i^{\text{surv}} = -\infty$. Let $t > T_{i\ell_i}$:

$$\begin{aligned}
\mathbb{P}(T_{i(\ell_i+1)} > t \mid T_{i\ell_i}, S_{i\ell_i}) &= \mathbb{P}\Big(\bigcap_{s:\,(E_{i(\ell_i-1)},s)\in E} T^s_{i(\ell_i+1)} > t \mid T_{i\ell_i}, S_{i\ell_i}\Big), \\
&= \prod_{s:\,(S_{i\ell_i},s)\in E} \mathbb{P}(T^s_{i(\ell_i+1)} > t \mid T_{i\ell_i}, S_{i\ell_i}), \\
&= \exp\Big(-\sum_{s:\,(S_{i\ell_i},s)\in E} \Lambda_i^{s|S_{i(\ell_i-1)}}\big(t \mid b_i, X_i, T_{i\ell_i}, \theta\big)\Big).
\end{aligned}$$

Furthermore:

$$\begin{aligned}
&\mathbb{P}(S_{i(\ell_i+1)} \mid T_{i(\ell_i+1)} = t, T_{i\ell_i}, S_{i\ell_i}) \\
&\propto p(\{T^{S_{i(\ell_i+1)}}_{i(\ell_i+1)} = t\} \cap \bigcap_{\substack{s:\,(S_{i\ell_i},s)\in E,\\ s\neq S_{i(\ell_i+1)}}} \{T^s_{i(\ell_i+1)} \geq t\} \mid T_{i\ell_i}, S_{i\ell_i}), \\
&= \lambda_i^{S_{i(\ell_i+1)}|S_{i\ell_i}}\big(t \mid \theta, X_i, T_{i\ell_i}, b_i\big) \exp\Big(-\sum_{s:\,(S_{i\ell_i},s)\in E} \Lambda_i^{s|S_{i\ell_i}}\big(t \mid b_i, X_i, T_{i\ell_i}, \theta\big)\Big).
\end{aligned}$$

Combining both, we retrieve the same joint density.

If $t_i^{\text{surv}} > -\infty$, one can also check that both densities are equal. Note that the condtioning only affects the first transition, as the next time $\forall \ell_i > 0$, $T_{i\ell_i} \geq t_i^{\text{surv}} \implies \forall \ell_i > 0$, $T_{i\ell_i} \vee t_i^{\text{surv}} = T_{i\ell_i}$.

□

B Expression of the Semi-Markov likelihood

Proof The proof is short and relies on the fact that, using the semi-Markov property 2

$$
\begin{aligned}
p\left((T_{i(\ell+1)}, S_{i(\ell+1)})_{\ell \leq m_i - 1} \mid b_i, X_i, \theta\right) &= \prod_{\ell=0}^{m_i-1} p\left((T_{i(\ell+1)}, S_{i(\ell+1)}) \mid b_i, X_i, (T_{i\ell'}, S_{i\ell'})_{\ell' \leq \ell}, \theta\right), \\
&= \prod_{\ell=0}^{m_i-1} p\left((T_{i(\ell+1)}, S_{i(\ell+1)}) \mid b_i, X_i, (T_{i\ell}, S_{i\ell}), \theta\right).
\end{aligned}
$$

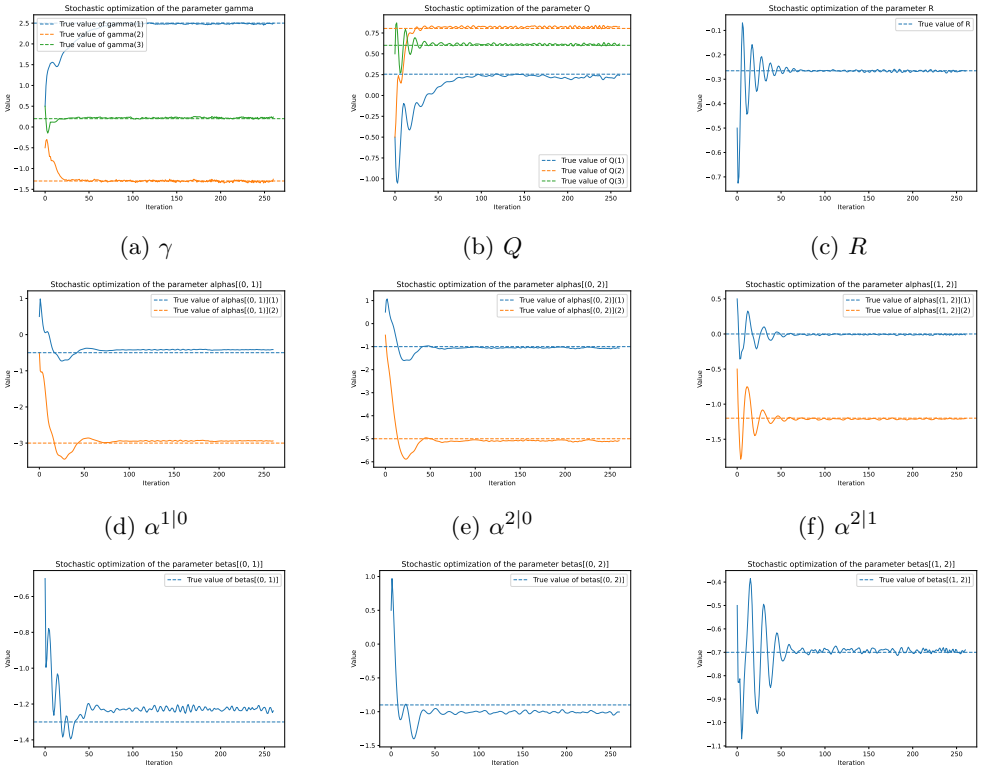
On the other hand, the probability that no additional event is observed between T_{im_i} and C_i is 1 if the last state reached is absorbing, and otherwise

$$
\mathbb{P}(T_{i(m_i+1)} > C_i \mid \theta, X_i, (T_{im_i}, S_{im_i}), b_i) = \exp\left(- \sum_{s:\, (S_{im_i}, s) \in E} \Lambda_i^{s|S_{im_i}}(C_i \mid b_i, X_i, T_{im_i}, \theta)\right),
$$

and so we recover the formula cited above with the convention that an empty sum is 0.

□

C Stochastic optimization process



(a) γ (b) Q (c) R

(d) $\alpha^{1|0}$ (e) $\alpha^{2|0}$ (f) $\alpha^{2|1}$

(g) $\beta^{1|0}$ (h) $\beta^{2|0}$ (i) $\beta^{2|1}$

Fig. 10: Evolution of the parameters during the optimization of the marginal log-likelihood using stochastic gradient descent. The dotted lines correspond to the true values.