A General Estimation Framework for Multi-State Markov Processes with Flexible Specification of the Transition Intensities

Alessia Eletti * Giampiero Marra† Rosalba Radice ‡

Abstract

When interest lies in the progression of a disease rather than on a single outcome, non-homogeneous multi-state Markov models constitute a natural and powerful modelling approach. Constant monitoring of a phenomenon of interest is often unfeasible, hence leading to an intermittent observation scheme. This setting is challenging and existing models and their implementations do not yet allow for flexible enough specifications that can fully exploit the information contained in the data. To widen significantly the scope of multi-state Markov models, we propose a closed-form expression for the local curvature information of a key quantity, the transition probability matrix. Such development allows one to model any type of multi-state Markov process, where the transition intensities are flexibly specified as functions of additive predictors. Parameter estimation is carried out through a carefully structured, stable penalised likelihood approach. The methodology is exemplified via two case studies that aim at modelling the onset of cardiac allograft vasculopathy, and cognitive decline. To support applicability and reproducibility, all developed tools are implemented in the `R` package `flexmsm`.

Keywords: additive predictor; information matrix; longitudinal survival data; Markov model; penalised log-likelihood; regression spline.

1 Introduction

With the increase in the availability of longitudinal survival data, continuous-time multi-state Markov models have established themselves as powerful tools to model the progression of a phenomenon, while accounting for background information recorded for each individual throughout

*Department of Statistical Science, University College London, Gower Street, WC1E 6BT London, UK, alessia.eletti.19@ucl.ac.uk.

†Department of Statistical Science, University College London, Gower Street, WC1E 6BT London, UK, giampiero.marra@ucl.ac.uk.

‡Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London, 106 Bunhill Row, EC1Y 8TZ London, UK, rosalba.radice@city.ac.uk.

the follow-up period; see Yiu et al. (2017), Williams et al. (2020) and Gorfine et al. (2021) for some examples. In many applications, a non-homogeneous Markov process is assumed, i.e. the risks of moving across states depend on the current state and on time. This is typically addressed by employing parametric functional forms for the transition hazards, but some examples of more flexible (e.g., spline-based) specifications can be found as well (e.g., Cook and Lawless, 2018; Joly et al., 2002; Machado et al., 2021; Titman, 2011; Van Den Hout, 2017).

Constant monitoring of the progression of a phenomenon of interest is often not possible since it may be too expensive or altogether not feasible due to the nature of the event of interest. When this is the case, the process is only observed at a fixed set of times and is thus said to be intermittently observed or interval-censored. The lack of knowledge of the times in which the transitions occurred represents a methodological challenge. The literature on the subject is vast, however existing computational methods for fitting non-homogeneous multi-state Markov models in such setting are mainly based on the estimation approach developed by Kalbfleisch and Lawless (1985), which relies on approximating the information matrix using the analytical score of the log-likelihood. The advantage of this method is that, in principle, it permits a great degree of generality by allowing for any number of states, forward and backward transitions, and any type of functional form for the transition intensities. However, only simpler models are supported in practice, with the most commonly used implementation provided via the `R` package `msm` (Jackson, 2022). Yet, convergence failures occur when the numbers of states and covariates increase; this can be attributed to the absence of the analytical information matrix which would provide valuable exact curvature information exploitable in model fitting. Using the approach of Kalbfleisch and Lawless (1985), Machado et al. (2021) presented a model that includes a smooth function of the time variable, and provided a bespoke code for the simple and well-known three-state Illness-Death Model (IDM). Alternatives, such as the pseudo-values based approach discussed in Sabathé et al. (2020) or the semiparametric regression model proposed in Gu et al. (2022), also fall short of the desired generality as they only support the IDM case, are limited in the functional forms allowed for the transition intensities and/or fail to provide software.

To widen significantly the scope of non-homogeneous multi-state Markov models, we propose an analytical expression for the local curvature information of the transition probability matrix. This allows us to introduce a modelling framework which is general and flexible, and that is applicable to far more complex empirical problems than those previously explorable in the literature. Specifically, the proposal allows for any type of multi-state process, with several states and various combinations of observation schemes (e.g., intermittent, exactly observed, censored), and for the transition intensities to be flexibly modelled through additive predictors. Parameter estimation is carried out by adapting to this context the stable and efficient estimation algorithm of Marra and Radice (2020) which can fully exploit the newly derived analytical observed information matrix. To allow for reproducible and transparent research, the framework is implemented in the `R` package `flexmsm` (Eletti et al., 2023) which is very easy and intuitive to use; for instance, time and

covariate effects of multi-state Markov models can be flexibly specified using the same syntax as that for generalised additive models in `R` (Wood, 2017).

In Section 2, we introduce the mathematical setting of multi-state Markov models and describe the regression spline-based approach employed for modelling the transition intensities. The penalised log-likelihood is presented in Section 3, while parameter estimation and how this is intertwined with the problem of computing the transition probabilities from the transition intensities are discussed in Section 4. This section also presents the newly derived analytical expression for the local curvature information of the transition probability matrix; the proof is provided in the Appendix. Section 5 describes how inference is carried out. Section 6 illustrates the potential of the proposal via a classical study, based on the IDM, that aims at modelling the onset of cardiac allograft vasculopathy, and a more complex one, about cognitive decline, which requires the use of a five-state process with both forward and backward transitions as well as an absorbing death state. Section 7 concludes the paper with some directions of future research. On-line Supplementary Materials A, B and C provide details on the log-likelihood contributions, the `R` package `flexmsm` and the algorithm employed for parameter estimation. Supplementary Material D illustrates the empirical effectiveness of the proposal via two simulation studies. Supplementary Material E contains a list of the mathematical symbols used and their meaning.

2 Multi-state processes with flexible transition intensities

Let $\{Z(t), t > 0\}$ be a continuous-time Markov process, $\mathcal{S} = \{1, 2, \ldots, C\}$ its discrete state space, where C is the total number of states, and $\mathcal{A} = \{(r, r') \mid r \neq r' \in \mathcal{S}, \exists\, r \rightarrow r'\}$ the set of transitions. The transition intensity function, i.e. the instantaneous rate of transition to a state r' for an individual who is currently in another state r, is defined as follows

$$q^{(rr')}(t) = \lim_{h \downarrow 0} \frac{P(Z(t+h) = r' \mid Z(t) = r)}{h}, \quad r \neq r',$$

with $q^{(rr')}(t) = 0$ if *r* is an absorbing state and $q^{(rr)}(t) = -\sum_{r \neq r'} q^{(rr')}(t)$. The matrix with (r, r') element given by $q^{(rr')}(t)$ for every $r, r' \in \mathcal{S}$ is called transition intensity matrix or generator matrix and can be denoted with $\mathbf{Q}(t)$. Similarly, the transition probability matrix associated with the time interval (t, t') is defined as the matrix with (r, r') element given by $p^{(rr')}(t, t') = P(Z(t') = r' \mid Z(t) = r)$ and can be denoted with $\mathbf{P}(t, t')$. Here, we assume a time-dependent process as opposed to the rather restrictive time-homogeneous process (i.e., $\mathbf{Q}(t) = \mathbf{Q}\ \forall t > 0$) often adopted in the literature for mathematical convenience.

The intensity for transition $r \rightarrow r'$, with $r \neq r'$, is generally represented using the proportional hazards specification, where the baseline intensity is typically specified using the exponential or Gompertz distribution (Van Den Hout, 2017). A more flexible representation for the transition

intensity is

$$q^{(rr')}(t_\iota) = \exp\left[\eta_\iota^{(rr')}(t_\iota, \mathbf{x}_\iota; \boldsymbol{\beta}^{(rr')})\right], \tag{1}$$

where t_ι and $\mathbf{x}_\iota$ are the time and the vector of characteristics for observation ι respectively, $\boldsymbol{\beta}^{(rr')}$ is the associated regression coefficient vector and $\eta_\iota^{(rr')}(t_\iota, \mathbf{x}_\iota; \boldsymbol{\beta}^{(rr')}) \in \mathbb{R}$ is an additive predictor, discussed in detail in the following section, which includes a baseline smooth function of time and several types of covariate effects.

2.1 Additive predictor

For simplicity, the dependence on covariates and parameters has been dropped when discussing the construction of $\eta_\iota^{(rr')}$.

An additive predictor allows for various types of covariate effects and is defined as

$$\eta_\iota^{(rr')} = \beta_0^{(rr')} + \sum_{k=1}^{K^{(rr')}} s_k^{(rr')}(\tilde{\mathbf{x}}_{k\iota}), \quad \iota = 1, \ldots, \check{n}, \tag{2}$$

where $\check{n}$ is the sample size, $\beta_0^{(rr')} \in \mathbb{R}$ is an overall intercept, $\tilde{\mathbf{x}}_{k\iota}$ denotes the k^{th} sub-vector of the complete vector $\tilde{\mathbf{x}}_\iota = (t_\iota, \mathbf{x}_\iota^\mathsf{T})^\mathsf{T}$ and the $K^{(rr')}$ functions $s_k^{(rr')}(\tilde{\mathbf{x}}_{k\iota})$ represent effects which are chosen according to the type of covariate(s) considered. For example, if we were interested in modelling a time-dependent effect of the covariate age_ι, then $\tilde{\mathbf{x}}_{\iota k}$ would be the vector $(age_\iota, t_\iota)^\mathsf{T}$ and $s_k^{(rr')}(age_\iota, t_\iota)$ the corresponding joint effect. Each $s_k^{(rr')}(\tilde{\mathbf{x}}_{k\iota})$ can be represented as a linear combination of $J_k^{(rr')}$ known basis functions $\mathbf{b}_k^{(rr')}(\tilde{\mathbf{x}}_{k\iota}) = \left(b_{k1}^{(rr')}(\tilde{\mathbf{x}}_{k\iota}), \ldots, b_{kJ_k^{(rr')}}^{(rr')}(\tilde{\mathbf{x}}_{k\iota})\right)^\mathsf{T}$ and regression coefficients $\boldsymbol{\beta}_k^{(rr')} = \left(\beta_{k1}^{(rr')}, \ldots, \beta_{kJ_k^{(rr')}}^{(rr')}\right)^\mathsf{T} \in \mathbb{R}^{J_k^{(rr')}}$, that is $s_k^{(rr')}(\tilde{\mathbf{x}}_{k\iota}) = \mathbf{b}_k^{(rr')}(\tilde{\mathbf{x}}_{k\iota})^\mathsf{T}\boldsymbol{\beta}_k^{(rr')}$ (e.g., Wood, 2017). The above formulation implies that the vector of evaluations $\left\{s_k^{(rr')}(\tilde{\mathbf{x}}_{k1}), \ldots, s_k^{(rr')}(\tilde{\mathbf{x}}_{k\check{n}})\right\}^\mathsf{T}$ can be written as $\tilde{\mathbf{X}}_k^{(rr')}\boldsymbol{\beta}_k^{(rr')}$, where $\tilde{\mathbf{X}}_k^{(rr')}$ is the design matrix whose ι^{th} row is given by $\mathbf{b}_k^{(rr')}(\tilde{\mathbf{x}}_{k\iota})^\mathsf{T}$ for $\iota = 1, \ldots, \check{n}$. This allows the predictor in equation (2) to be written as $\boldsymbol{\eta}^{(rr')} = \beta_0^{(rr')}\mathbf{1}_{\check{n}} + \tilde{\mathbf{X}}_1^{(rr')}\boldsymbol{\beta}_1^{(rr')} + \ldots + \tilde{\mathbf{X}}_{K^{(rr')}}^{(rr')}\boldsymbol{\beta}_{K^{(rr')}}^{(rr')}$, where $\mathbf{1}_{\check{n}}$ is an $\check{n}$-dimensional vector made up of ones. This can also be represented in a more compact way as $\boldsymbol{\eta}^{(rr')} = \tilde{\mathbf{X}}^{(rr')}\boldsymbol{\beta}^{(rr')}$, where $\tilde{\mathbf{X}}^{(rr')} = (\mathbf{1}_n, \tilde{\mathbf{X}}_1^{(rr')}, \ldots, \tilde{\mathbf{X}}_{K^{(rr')}}^{(rr')})$ and $\boldsymbol{\beta}^{(rr')} = \left(\beta_0^{(rr')\mathsf{T}}, \boldsymbol{\beta}_1^{(rr')\mathsf{T}}, \ldots, \boldsymbol{\beta}_{K^{(rr')}}^{(rr')\mathsf{T}}\right)^\mathsf{T}$. Each $\boldsymbol{\beta}_k^{(rr')}$ has an associated quadratic penalty $\lambda_k^{(rr')}\boldsymbol{\beta}_k^{(rr')\mathsf{T}}\mathbf{D}_k^{(rr')}\boldsymbol{\beta}_k^{(rr')}$, used in fitting, whose role is to enforce specific properties on the k^{th} function, such as smoothness. Matrix $\mathbf{D}_k^{(rr')}$ only depends on the choice of the basis functions. Smoothing parameter $\lambda_k^{(rr')} \in [0, \infty)$ has the crucial role of controlling the trade-off between fit and smoothness and hence it determines the shape of the corresponding estimated smooth function. The overall penalty is defined as $\boldsymbol{\beta}^{(rr')\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}^{(rr')}}^{(rr')}\boldsymbol{\beta}^{(rr')}$, where $\mathbf{S}_{\boldsymbol{\lambda}^{(rr')}}^{(rr')} = \mathrm{diag}(0, \lambda_1^{(rr')}\mathbf{D}_1^{(rr')}, \ldots, \lambda_{K^{(rr')}}^{(rr')}\mathbf{D}_{K^{(rr')}}^{(rr')})$

and $\boldsymbol{\lambda}^{(rr')} = (\lambda_1^{(rr')}, \ldots, \lambda_{K^{(rr')}}^{(rr')})^\mathsf{T}$ is the transition-specific overall smoothing parameter vector. Note that smooth functions are subject to centering (identifiability) constraints which are imposed as described in Wood (2017). Several definitions of basis functions and penalty terms are supported by `flexmsm`; these include thin plate, cubic and P-regression splines, tensor products, Markov random fields, random effects, and Gaussian process smooths (see Wood (2017) for details).

An example of predictor specification is $\eta_\iota^{(rr')} = \beta_0^{(rr')} + s_1^{(rr')}(t_\iota) + \beta_2^{(rr')} sex_\iota$. Parametric effects usually, but not exclusively, relate to binary and categorical variables such as sex_ι. The spline representation introduced above thus simplifies to $s_2^{(rr')}(sex_\iota) = \beta_2^{(rr')} sex_\iota$. No penalty is typically assigned to parametric effects, hence the associated penalty is 0. However, there might be instances where some form of regularisation is required in which case a suitable penalisation scheme can be employed (e.g., Wood, 2017, Section 5.8). To explore a potentially nonlinear effect of t_ι, $s_1^{(rr')}(t_\iota)$ is specified as $\mathbf{b}_1^{(rr')}(t_\iota)^\mathsf{T} \boldsymbol{\beta}_1^{(rr')}$, where $\mathbf{b}_1^{(rr')}(t_\iota)$ are cubic regression spline bases, for example. In this case, the penalty is defined as

$$\boldsymbol{\beta}_1^{(rr')^\mathsf{T}} \mathbf{D}_1^{(rr')} \boldsymbol{\beta}_1^{(rr')} = \int_{u_1}^{u_{J_1^{(rr')}}} \left(\frac{\partial^2}{\partial t^2} s_1^{(rr')}(u) \right)^2 du,$$

where u_1 and $u_{J_1^{(rr')}}$ are the locations of the first and last knots. For a smooth term in one dimension, such as $s_1^{(rr')}(t_\iota)$, the specific choice of spline definition (e.g., thin plate, cubic) will not have an impact on the estimated curve. As for $J_1^{(rr')}$, or more generally $J_k^{(rr')}$, this is typically set to 10 since such value offers enough flexibility in most applications. However, analyses using larger values can be attempted to assess the sensitivity of the results to $J_k^{(rr')}$. Regarding the selection of knots, these can be placed evenly throughout (or using the percentiles of) the values of the variable the smooth term refers to. For a thin-plate regression spline only $J_k^{(rr')}$ has to be chosen. See Wood (2017) for a thorough discussion.

As mentioned previously, our framework poses no limits on the types of splines that can be employed for specifying the transition intensities. For instance, as illustrated in Section 6.1, two-dimensional splines can be used to incorporate time-dependent effects. This would take the form of an interaction term involving, e.g., age_ι and the time variable through the smooth term $s_k^{(rr')}(age_\iota, t_\iota)$. Here we have two penalties, one for each of the arguments of the smooth function. These are summed after being weighted by smoothing parameters, which serve the purpose of controlling the trade-off between fit and smoothness in each of the two directions, thus allowing for a great degree of flexibility (Wood, 2017, Section 5.6).

3 Penalised log-likelihood

Let N be the number of statistical units, n_i the number of times the i^{th} unit is observed, $0 = t_{i0} < t_{i1} < \cdots < t_{in_i}$ the follow-up times, $z_{i0}, z_{i1}, \ldots, z_{in_i}$ the (possibly unobserved, i.e. censored) states occupied, and $\breve{n} = \sum_{i=1}^{N}(n_i - 1)$ the sample size. If $L_{ij}(\boldsymbol{\theta})$ is the likelihood contribution for the j^{th} observation of the i^{th} unit and $\boldsymbol{\theta} = \{\boldsymbol{\beta}^{(rr')} \mid (r, r') \in \mathcal{A}\}$ the model parameter vector, then the log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \log\left(L_{ij}(\boldsymbol{\theta})\right), \tag{3}$$

where we have

$$L_{ij}(\boldsymbol{\theta}) = \begin{cases} p^{(z_{ij-1}z_{ij})}(t_{ij-1}, t_{ij}), & \text{if } z_{ij} \text{ is a living state} \\ \exp\left[\int_{t_{ij-1}}^{t_{ij}} q^{(z_{ij-1}z_{ij-1})}(u)du\right] q^{(z_{ij-1}z_{ij})}(t_{ij}), & \text{if } z_{ij} \text{ is an exactly observed living state} \\ \sum_{c=1}^{C} p^{(z_{ij-1}c)}(t_{ij-1}, t_{ij}), & \text{if } z_{ij} \text{ is censored} \\ \sum_{\substack{c=1 \\ c\neq z_{ij}}}^{C} p^{(z_{ij-1}c)}(t_{ij-1}, t_{ij}) q^{(cz_{ij})}(t_{ij}), & \text{if } z_{ij} \text{ is an exactly observed absorbing state} \end{cases}.$$

That is, the likelihood contribution for a given observation will depend on the nature of the states between which the transition occurred and the way in which it was observed. Supplementary Material A provides details on each contribution type, whereas Supplementary Material B describes the use of the `R` package `flexmsm` in such a general context.

To calibrate the trade-off between parsimony and complexity, we augment the objective function (3) with a quadratic penalty term. This results in the penalised log-likelihood

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\theta}, \tag{4}$$

where $\mathbf{S}_{\boldsymbol{\lambda}} = \operatorname{diag}\left(\{\mathbf{S}^{(rr')}_{\boldsymbol{\lambda}^{(rr')}} \mid (r, r') \in \mathcal{A}\}\right)$ which is a block diagonal matrix where each block is given by the transition-specific penalty matrix $\mathbf{S}^{(rr')}_{\boldsymbol{\lambda}^{(rr')}}$, and $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^{(rr')} \mid (r, r') \in \mathcal{A}\}$ is the overall multiple smoothing parameter vector. Both $\mathbf{S}^{(rr')}_{\boldsymbol{\lambda}^{(rr')}}$ and $\boldsymbol{\lambda}^{(rr')}$ are defined for a generic transition (r, r') in Section 2.1.

4 Stable estimation through exact local curvature information

Building a general and flexible multi-state Markov modelling framework hinges on the availability of the analytical information matrix of the transition probability matrix, for which we propose a

version here. Parameter estimation is achieved by adapting to our setting the stable and efficient approach proposed in Marra and Radice (2020), which combines a trust region algorithm with automatic multiple smoothing parameter selection. The trust region method is known to perform better than its line search counterparts and has certain optimal convergence properties as long as the analytical observed information matrix is provided (Chapter 4, Nocedal and Wright, 2006). As for the smoothing parameters, we employ a general and fast estimation framework which removes the need for computationally expensive grid search-based methods and ad-hoc optimisers (see Supplementary Material C for details). From (3), the w^{th} element of the gradient vector $\mathbf{g}(\boldsymbol{\theta})$ and the (w, w') element of the Hessian matrix $\mathbf{H}(\boldsymbol{\theta})$, for $w, w' = 1, \ldots, W$ with $W = \sum_{(r,r')\in\mathcal{A}} \left(1 + \sum_{k=1}^{K^{(rr')}} J_k^{(rr')}\right)$, are defined as

$$\frac{\partial}{\partial\theta_w}\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i} L_{ij}(\boldsymbol{\theta})^{-1}\frac{\partial}{\partial\theta_w}L_{ij}(\boldsymbol{\theta}),$$

$$\frac{\partial^2}{\partial\theta_w\partial\theta_{w'}}\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(L_{ij}(\boldsymbol{\theta})^{-1}\frac{\partial^2}{\partial\theta_w\partial\theta_{w'}}L_{ij}(\boldsymbol{\theta}) - L_{ij}(\boldsymbol{\theta})^{-2}\frac{\partial}{\partial\theta_w}L_{ij}(\boldsymbol{\theta})\frac{\partial}{\partial\theta_{w'}}L_{ij}(\boldsymbol{\theta})\right),$$

where $\dfrac{\partial L_{ij}(\boldsymbol{\theta})}{\partial\theta_w}$ is given by

$$\begin{cases} \dfrac{\partial}{\partial\theta_w}p^{(z_{ij-1}z_{ij})}(t_{ij-1}, t_{ij}), & \text{if } z_{ij} \text{ is a living state} \\ \exp\left(\int_{t_{ij-1}}^{t_{ij}} q^{(z_{ij-1}z_{ij-1})}(u)du\right)\left[\dfrac{\partial}{\partial\theta_w}q^{(z_{ij-1}z_{ij})}(t_{ij}) \right. & \text{if } z_{ij} \text{ is an exactly observed living state} \\ \left. +q^{(z_{ij-1}z_{ij})}(t_{ij})\int_{t_{ij-1}}^{t_{ij}}\dfrac{\partial}{\partial\theta_w}q^{(z_{ij-1}z_{ij-1})}(u)du\right], & \\ \sum_{c=1}^{C}\dfrac{\partial}{\partial\theta_w}p^{(z_{ij-1}c)}(t_{ij-1}, t_{ij}), & \text{if } z_{ij} \text{ is censored} \\ \sum_{c=1}^{C}\dfrac{\partial}{\partial\theta_w}p^{(z_{ij-1}c)}(t_{ij-1}, t_{ij})q^{(cz_{ij})}(t_{ij}) & \text{if } z_{ij} \text{ is an exactly observed absorbing state} \\ \quad +p^{(z_{ij-1}c)}(t_{ij-1}, t_{ij})\dfrac{\partial}{\partial\theta_w}q^{(cz_{ij})}(t_{ij}), & \end{cases},$$

and $\dfrac{\partial^2 L_{ij}(\boldsymbol{\theta})}{\partial\theta_w\partial\theta_{w'}}$ is given by

$$
\begin{cases}
\dfrac{\partial^2}{\partial\theta_w\partial\theta_{w'}}p^{(z_{ij-1}z_{ij})}(t_{ij-1},t_{ij}), & \text{if } z_{ij} \text{ is a living state} \\
\dfrac{\partial}{\partial\theta_w}L_{ij}(\boldsymbol{\theta})\int\limits_{t_{ij-1}}^{t_{ij}}\dfrac{\partial}{\partial\theta_{w'}}q^{(z_{ij-1}z_{ij-1})}(u)du & \text{if } z_{ij} \text{ is an exactly observed living state} \\
+\exp\left(\int\limits_{t_{ij-1}}^{t_{ij}}q^{(z_{ij-1}z_{ij-1})}(u)du\right)\left[\dfrac{\partial^2 q^{(z_{ij-1}z_{ij})}(t_{ij})}{\partial\theta_w\partial\theta_{w'}}\right. & \\
+\dfrac{\partial}{\partial\theta_{w'}}q^{(z_{ij-1}z_{ij})}(t_{ij})\int\limits_{t_{ij-1}}^{t_{ij}}\dfrac{\partial}{\partial\theta_w}q^{(z_{ij-1}z_{ij-1})}(u)du & \\
\left.+q^{(z_{ij-1}z_{ij})}(t_{ij})\int\limits_{t_{ij-1}}^{t_{ij}}\dfrac{\partial^2 q^{(z_{ij-1}z_{ij-1})}(u)}{\partial\theta_w\partial\theta_{w'}}du\right], & \\
\sum\limits_{c=1}^{C}\dfrac{\partial^2}{\partial\theta_w\partial\theta_{w'}}p^{(z_{ij-1}c)}(t_{ij-1},t_{ij}), & \text{if } z_{ij} \text{ is censored} \\
\sum\limits_{c=1}^{C}\dfrac{\partial^2}{\partial\theta_w\partial\theta_{w'}}p^{(z_{ij-1}c)}(t_{ij-1},t_{ij})q^{(cz_{ij})}(t_{ij}) & \text{if } z_{ij} \text{ is an exactly observed absorbing state} \\
\quad+\dfrac{\partial}{\partial\theta_w}p^{(z_{ij-1}c)}(t_{ij-1},t_{ij})\dfrac{\partial}{\partial\theta_{w'}}q^{(cz_{ij})}(t_{ij}) & \\
\quad+\dfrac{\partial}{\partial\theta_{w'}}p^{(z_{ij-1}c)}(t_{ij-1},t_{ij})\dfrac{\partial}{\partial\theta_w}q^{(cz_{ij})}(t_{ij}) & \\
\quad+p^{(z_{ij-1}c)}(t_{ij-1},t_{ij})\dfrac{\partial^2}{\partial\theta_w\partial\theta_{w'}}q^{(cz_{ij})}(t_{ij}), &
\end{cases}
$$

The quantities needed for parameter estimation are the $C \times C$ dimensional matrices $\mathbf{P}(t_{ij-1}, t_{ij})$, $\partial\mathbf{P}(t_{ij-1}, t_{ij})/\partial\theta_w$ and $\partial^2\mathbf{P}(t_{ij-1}, t_{ij})/\partial\theta_w\partial\theta_{w'}$ for $w, w' = 1, \ldots, W$. Given the transition intensity matrix $\mathbf{Q}(t)$, the transition probability matrix is the solution of the Kolmogorov forward differential equations $\partial\mathbf{P}(t,t')/\partial t' = \mathbf{P}(t,t')\mathbf{Q}(t')$, which are not in general tractable. Kalbfleisch and Lawless (1985) proposed analytical expressions for $\mathbf{P}(t_{ij-1}, t_{ij})$ and $\partial\mathbf{P}(t_{ij-1}, t_{ij})/\partial\theta_w$ but not for $\partial^2\mathbf{P}(t_{ij-1}, t_{ij})/\partial\theta_w\partial\theta_{w'}$, which is needed to derive the observed information matrix. The next section presents a closed-form expression for $\partial^2\mathbf{P}(t_{ij-1}, t_{ij})/\partial\theta_w\partial\theta_{w'}$.

4.1 Observed information matrix of the transition probability matrix

In the following, time-dependency of (1) is taken into account by employing the commonly adopted piecewise-constant approximation approach. As for the time grid over which such approximation is defined, we let it coincide with the observation times of the dataset at hand; this allows for satisfactory estimation of the model parameters at a contained computational cost (Van Den Hout, 2017). Grids can be defined differently if required (e.g., Van den Hout and Matthews, 2008).

For each individual $i = 1, \ldots, N$, let the observed follow-up times $t_{i0} < t_{i1} < \cdots < t_{in_i}$ define

the extremities of the intervals over which the transition intensities are assumed to be constant. The convention is to assume that the transition intensities remain constant on the value taken in the left extremity of each time interval. Then, for $t \in [t_{ij}, t_{ij+1})$, with $j = 0, 1, \ldots, n_i - 1$, making explicit the dependence on the model parameters, we have $\mathbf{Q}(t; \boldsymbol{\theta}) = \mathbf{Q}_j(\boldsymbol{\theta})$ and

$$\mathbf{P}(t_{ij}, t_{ij+1}) = \mathbf{P}(t_{ij+1} - t_{ij}) = \exp[(t_{ij+1} - t_{ij})\mathbf{Q}_j(\boldsymbol{\theta})] = \sum_{\zeta=0}^{\infty} \frac{[(t_{ij+1} - t_{ij})\mathbf{Q}_j(\boldsymbol{\theta})]^{\zeta}}{\zeta!}. \quad (5)$$

Computing the transition probability matrix and its derivatives entails calculating a number of matrix exponentials and their derivatives. The eigendecomposition approach popularised by Kalbfleisch and Lawless (1985) is appealing because it provides a closed-form solution for the power series in (5) and for $\partial \mathbf{P}(t_{ij-1}, t_{ij})/\partial\theta_w$. Theorem 4.1 gives the result for $\partial^2 \mathbf{P}(t_{ij-1}, t_{ij})/\partial\theta_w \partial\theta_{w'}$. Note that solving the power series for this quantity is a rather involved process because of matrix-multiplication being non-commutative. However, the final expression is compact. For simplicity, let us drop the dependence on i, j and $\boldsymbol{\theta}$ and define $\delta t = t_{ij+1} - t_{ij}$.

Theorem. *Let $\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^{-1}$ be the eigendecomposition of the transition intensity matrix, which is constant over the generic time interval of length δt, with $\boldsymbol{A}$ the matrix of eigenvectors and $\boldsymbol{\Gamma} = \mathrm{diag}[\gamma_1, \ldots, \gamma_Y]$ the diagonal matrix containing Y distinct eigenvalues. Then*

$$\frac{\partial^2}{\partial\theta_w \partial\theta_{w'}} \boldsymbol{P}(\delta t) = \boldsymbol{A}(\check{\boldsymbol{U}}_{ww'} + \dot{\boldsymbol{U}}_{ww'} + \dot{\boldsymbol{U}}_{w'w})\boldsymbol{A}^{-1}.$$

The (l, m) element of $\check{\boldsymbol{U}}_{ww'}$ is

$$\check{\boldsymbol{U}}_{ww'}[l, m] = \begin{cases} G_{lm}^{(ww')} \dfrac{e^{\gamma_l \delta t} - e^{\gamma_m \delta t}}{\gamma_l - \gamma_m}, & l \neq m \\ G_{ll}^{(ww')} \delta t e^{\gamma_l \delta t}, & l = m \end{cases},$$

where $G_{lm}^{(ww')}$ is the (l, m) element of matrix $\boldsymbol{G}^{(ww')} = \boldsymbol{A}^{-1} \dfrac{\partial^2 \boldsymbol{Q}}{\partial\theta_w \partial\theta_{w'}} \boldsymbol{A}$. Element (l, m) of $\dot{\boldsymbol{U}}_{ww'}$ is

$$\dot{\boldsymbol{U}}_{ww'}[l, m] = \begin{cases} \displaystyle\sum_{\substack{y=1 \\ y \neq l, m}}^{Y} G_{ly}^{(w)} G_{ym}^{(w')} \left(\frac{e^{\gamma_l \delta t} - e^{\gamma_y \delta t}}{(\gamma_l - \gamma_y)(\gamma_y - \gamma_m)} - \frac{e^{\gamma_l \delta t} - e^{\gamma_m \delta t}}{(\gamma_l - \gamma_m)(\gamma_y - \gamma_m)} \right) \\ \quad + G_{ll}^{(w)} G_{lm}^{(w')} \left(\dfrac{t e^{\gamma_l \delta t}}{\gamma_l - \gamma_m} - \dfrac{e^{\gamma_l \delta t} - e^{\gamma_m \delta t}}{(\gamma_l - \gamma_m)^2} \right) \\ \quad + G_{lm}^{(w)} G_{mm}^{(w')} \left(\dfrac{e^{\gamma_l \delta t} - e^{\gamma_m \delta t}}{(\gamma_l - \gamma_m)^2} - \dfrac{t e^{\gamma_m \delta t}}{\gamma_l - \gamma_m} \right), & l \neq m \\ \displaystyle\sum_{\substack{y=1 \\ y \neq l}}^{Y} G_{ly}^{(w)} G_{yl}^{(w')} \left(\dfrac{t e^{\gamma_l \delta t}}{\gamma_l - \gamma_y} - \dfrac{e^{\gamma_l \delta t} - e^{\gamma_y \delta t}}{(\gamma_l - \gamma_y)^2} \right) + \frac{1}{2} G_{ll}^{(w)} G_{ll}^{(w')} \delta t^2 e^{\gamma_l \delta t}, & l = m \end{cases},$$

where $G_{ly}^{(w)}$ *is the* (l, y) *element of matrix* $\boldsymbol{G}^{(w)} = \boldsymbol{A}^{-1}\dfrac{\partial \boldsymbol{Q}}{\partial \theta_w}\boldsymbol{A}$ *and* $G_{ym}^{(w')}$ *is the* (y, m) *element of matrix* $\boldsymbol{G}^{(w')} = \boldsymbol{A}^{-1}\dfrac{\partial \boldsymbol{Q}}{\partial \theta_{w'}}\boldsymbol{A}$. $\dot{\boldsymbol{U}}_{w'w}$ *is obtained in the same way as* $\dot{\boldsymbol{U}}_{ww'}$ *but with* w *and* w' *swapped wherever they appear.*

The proof is provided in the Appendix.

Note that $\partial\mathbf{Q}/\partial\theta_w$ and $\partial^2\mathbf{Q}/\partial\theta_w\partial\theta_{w'}$ are matrices whose (r, r') elements are given, respectively, by $\partial q^{(rr')}(t_{ij})/\partial\theta_w$ and $\partial^2 q^{(rr')}(t_{ij})/\partial\theta_w\partial\theta_{w'}$ for $w, w' = 1, \ldots, W$. Further, the first derivatives of the transition intensity matrix are already available from the computation of the first derivatives of the transition probabilities, hence only second derivatives have to be computed anew. Matrices $\mathbf{A}$, $\mathbf{A}^{-1}$ and $\mathbf{\Gamma}$ also need to be computed only once, when obtaining matrix $\mathbf{P}$. In fact, from Kalbfleisch and Lawless (1985), $\mathbf{P}(\delta t) = \mathbf{A}\,\mathrm{diag}\left[\exp(\gamma_1\delta t), \ldots, \exp(\gamma_Y\delta t)\right]\mathbf{A}^{-1}$ and $\partial\mathbf{P}(\delta t)/\partial\theta_w = \mathbf{A}\dot{\mathbf{U}}_w\mathbf{A}^{-1}$ with the (l, m) element of $\dot{\mathbf{U}}_w$ given by

$$\dot{\mathbf{U}}_w[l, m] = \begin{cases} G_{lm}^{(w)}\dfrac{e^{\gamma_l\delta t} - e^{\gamma_m\delta t}}{\gamma_l - \gamma_m}, & l \neq m \\ G_{ll}^{(w)}e^{\gamma_l\delta t}\delta t, & l = m \end{cases}.$$

Theorem (4.1) requires distinct eigenvalues. In practice, in the literature and our own extensive experimentation, this has always been found to be the case for non-ill-defined problems. Regarding the implementation of the algorithm, the number of operations grows quickly as n_i, N, C and W increase. Specifically, $\mathbf{Q}$ (and its eigendecomposition), $\partial\mathbf{Q}/\partial\theta_w$, $\partial^2\mathbf{Q}/\partial\theta_w\partial\theta_{w'}$, $\mathbf{P}$, $\partial\mathbf{P}/\partial\theta_w$ and $\partial^2\mathbf{P}/\partial\theta_w\partial\theta_{w'}$, for $w, w' = 1, \ldots, W$, have to be computed $\sum_{i=1}^{N} n_i - N$ times and then combined. To reduce computational cost, the proposed implementation exploited the upper-triangle form of the above mentioned matrices and the presence of structural zero-values in them. We also exploited parallel computing to obtain the log-likelihood, analytical score and information matrix more quickly; the overall run-time of the algorithm can be cut by a factor proportional to the number of cores in the user's computer. Section 6.1 provides an example of the resulting difference with the implementation provided by Machado et al. (2021) which was found to be an order of magnitude slower than the `R` package `flexmsm`.

5 Inference

To obtain confidence intervals, instead of using the classically derived frequentist covariance matrix $-\mathbf{H}_p^{-1}(\boldsymbol{\theta})\mathbf{H}(\boldsymbol{\theta})\mathbf{H}_p^{-1}(\boldsymbol{\theta})$, we follow Wood et al. (2016) and employ the Bayesian large sample approximation $\boldsymbol{\theta} \dot{\sim} \mathcal{N}(\widehat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})$, where $\mathbf{V}_{\boldsymbol{\theta}} = -\mathbf{H}_p(\widehat{\boldsymbol{\theta}})^{-1}$ with $\hat{\boldsymbol{\theta}}$ the estimated model parameter and $\mathbf{H}_p(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta}) - \mathbf{S}_{\boldsymbol{\lambda}}$ the penalised Hessian. Using $\mathbf{V}_{\boldsymbol{\theta}}$ gives close to across-the-function frequentist coverage probabilities because it accounts for both sampling variability and smoothing bias, a feature that is particularly relevant at finite sample sizes and that is not shared by the

frequentist covariance matrix. Note that applying the Bayesian approach to the modelling framework discussed in this paper follows the notion that penalisation in estimation implicitly assumes that wiggly models are less likely than smoother ones, which translates into the following prior specification for $\boldsymbol{\theta}$, $f_{\boldsymbol{\theta}} \propto \exp\left\{-\boldsymbol{\theta}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\theta}/2\right\}$.

Intervals for linear functions of the model coefficients, e.g. smooth components, are obtained using the result just shown for $\boldsymbol{\theta}$. For nonlinear functions of the model coefficients, intervals can be conveniently obtained by posterior simulation. For example, to derive the $(1-\alpha)100\%$ intervals for the (r, r') transition intensity, the following procedure is employed:

1. Draw n_{sim} random vectors $\boldsymbol{\beta}^{(1,rr')}, \ldots, \boldsymbol{\beta}^{(n_{sim},rr')}$ from $\mathcal{N}(\widehat{\boldsymbol{\beta}^{(rr')}}, \mathbf{V}_{\boldsymbol{\beta}^{(rr')}})$, where $\widehat{\boldsymbol{\beta}^{(rr')}}$ is the estimated model parameter.

2. Calculate n_{sim} simulated realisations of the quantity of interest, such as $q^{(rr')}(t)$. For fixed $\mathbf{x}$ and t, one would obtain $\mathbf{q}_{sim}^{(rr')} = (q^{(1,rr')}, \ldots, q^{(n_{sim},rr')})^{\mathsf{T}}$ using $\boldsymbol{\beta}^{(1,rr')}, \ldots, \boldsymbol{\beta}^{(n_{sim},rr')}$ respectively.

3. Using $\mathbf{q}_{sim}^{(rr')}$, calculate the lower, $\alpha/2$, and upper, $1-\alpha/2$, quantiles.

A small value of $n_{sim} = 100$ typically gives accurate results, whereas α is usually set to 0.05. Note that the distribution of nonlinear functions of the model parameters need not be symmetric. Intervals for the transition probabilities can be obtained by applying the above procedure to the $\mathbf{Q}$ matrices and then deriving the corresponding $\mathbf{P}$ matrices, as explained in Section 4.1.

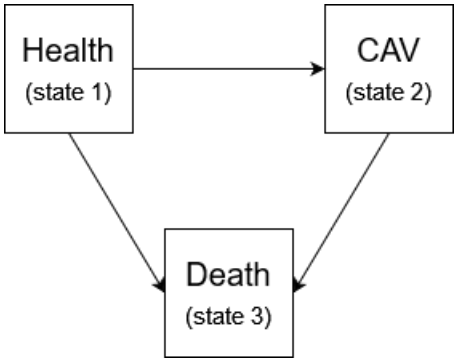
P-values for the terms in the model are obtained by using the results summarised in Wood (2017, Section 6.12), which are based on $-\mathbf{H}_p(\boldsymbol{\theta})^{-1}$. Model building can be aided using tools such as the Akaike information criterion (AIC, Akaike, 1998) and the Bayesian information criterion (BIC, Schwarz, 1978). The AIC and BIC are defined as $-2\boldsymbol{\ell}(\boldsymbol{\theta}) + 2edf$ and $-2\boldsymbol{\ell}(\boldsymbol{\theta}) + \log(\check{n})edf$, respectively, where the log-likelihood is evaluated at the penalised parameter estimates, $\check{n}$ is the sample size and the effective degrees of freedom are given by $edf = \text{tr}(\mathbf{O})$, with $\text{tr}(\cdot)$ the trace function and $\mathbf{O} = \sqrt{-\mathbf{H}(\boldsymbol{\theta})}\left(-\mathbf{H}_p(\boldsymbol{\theta})\right)^{-1}\sqrt{-\mathbf{H}(\boldsymbol{\theta})}$ (Marra and Radice, 2020).

6 Case studies

The proposal is illustrated through two case studies. The first one uses flexible IDMs to model the onset of cardiac allograft vasculopathy (CAV), a deterioration of the arterial walls in heart transplant patients. The second one aims at modelling cognitive decline in the English Longitudinal Study of Ageing (ELSA) population through a flexible five-state model with both forward and backward transitions as well as an absorbing death state.

6.1 CAV case study

The heart transplant monitoring data used here are openly accessible from the `R` package `msm`. The dataset contains 2846 observations, relating to 622 patients, and is about angiographic (approximately yearly) examinations of heart transplant recipients where the grade of CAV (not present, mild/moderate or severe) is recorded. The additional time event of death is also registered and known exactly (within one day). It follows that the likelihood contributions involved here are those relating to interval censored living states and to exactly observed absorbing states. Available baseline covariates include age of the donor (`dage`) and primary diagnosis of ischaemic heart disease (IHD, `pdiag`) which are known to be major risk factors for CAV onset. In line with Machado et al. (2021), we remove eight individuals for which the principal diagnosis is not known and exclude observations which occurred beyond 15 years from the transplant. The resulting dataset contains $\sum_{i=1}^{N} n_i = 2803$ observations of $N = 614$ patients. We consider flexible IDMs where the states are (1) health (2) CAV onset (mild/moderate or severe) and (3) death. A diagram representing the process is displayed in Figure 1 while Table 1 reports the number of observations available for each pair of states in the dataset. Note that the sum of these counts provides the sample size, $\check{n} = 2189$.



	state 1	state 2	state 3
state 1	1314	223	136
state 2	0	411	105
state 3	0	0	0

Table 1: Number of observations for each pair of states in the CAV dataset.

Figure 1: Diagram of the possible IDM disease trajectories.

The most flexible IDM considered in the literature for the CAV case study is based on Machado et al. (2021)

$$q^{(rr')}(t_{ij}) = \exp\left[\beta_0^{(rr')} + s_1^{(rr')}(t_{ij}) + \beta_2 \texttt{dage}_{ij} + \beta_3 \texttt{pdiag}_{ij}\right], \tag{6}$$

for $(r, r') \in \{(1,2),(1,3),(2,3)\}$, where *t* is the time since transplant, the smooth term is represented by a cubic regression spline with 10 basis functions and second order penalty, and β_2 and β_3 are covariate effects which are constrained to be equal across the three transitions, hence the lack of superscript. Model fitting was conducted using the bespoke `R` code provided by Machado et al. (2021) which took 3.5 days to reach convergence, on a laptop with Windows 10, Intel 2.20 GHz core, 16 GB of RAM and eight cores. The resulting AIC was 2931.7. No justification was provided for setting $\beta_2^{(rr')} = \beta_2$ and $\beta_3^{(rr')} = \beta_3$ which may be too restrictive to estimate adequately

the effects of dage and pdiag.

Using the proposed methodology, we considered the more general specification

$$q^{(rr')}(t_{ij}) = \exp\left[\beta_0^{(rr')} + s_1^{(rr')}(t_{ij}) + \beta_2^{(rr')}\texttt{dage}_i + \beta_3^{(rr')}\texttt{pdiag}_i\right], \tag{7}$$

which produced an AIC of 2915.2. The run-time of flexmsm was 59 minutes. Using different spline definitions and increasing $J_1^{(rr')}$ did not lead to tangible empirical differences.

Table 2 reports the effects for dage and pdiag, and their standard errors, resulting from models (6) and (7). As the table shows, the constrained coefficients are, roughly speaking, the averages of the respective unconstrained ones. In this case, setting restrictions does not allow one to uncover the differing effects of the risk factors in the different trajectories. Specifically, the model (7) results indicate that dage and pdiag increase the risks of moving from state 1 to state 2 and from state 1 to state 3, and that these variables do not play a role in the transition $2 \to 3$. The curve estimates for the $s_1^{(rr')}(t_{ij})$ (not reported here) were similar across the two models.

	dage	pdiag
$1 \to 2$	0.023 (0.006)	0.414 (0.132)
$1 \to 3$	0.040 (0.011)	0.341 (0.255)
$2 \to 3$	−0.016 (0.009)	0.002 (0.178)
$1 \to 2, 1 \to 3, 2 \to 3$	0.018 (0.004)	0.274 (0.096)

Table 2: Estimated covariate effects and related standard errors (between brackets) for donor age (dage) and principal diagnosis of IHD (pdiag) obtained using the proposed model fitted by flexmsm (first three lines) and the constrained model of Machado et al. (2021) fitted using the related bespoke R code.

Figure 2 shows the estimated transition intensities, and 95% intervals, when dage is equal to 26 years and pdiag is equal to 1 (i.e., the principal diagnosis is IHD). The risk of moving from state 1 to state 2 increases until about 7 years since transplant; after that the situation is uncertain. The risk for the transition $1 \to 3$ is fairly low and constant until about 10 years, after which it starts increasing steeply. For transition $2 \to 3$, the risk increases overall. As expected, the intervals are wide when the data are scarce. The same exercise can be repeated for different combinations of dage and pdiag.

Estimated transition intensities provide valuable information about the risks of moving across states. However, interpretation is more intuitive and easier when transition probabilities are considered. Setting dage = 26 and pdiag = 1 and assuming yearly piecewise constant transition intensities, the estimated five-year transition probabilities can be obtained by exploiting the Chapman-Kolmogorov equations (Cox and Miller, 1977). These allow us to write $\hat{\mathbf{P}}(0,5) = \hat{\mathbf{P}}(0,1) \times \hat{\mathbf{P}}(1,2) \times \cdots \times \hat{\mathbf{P}}(4,5)$, where the probabilities over each sub-interval are obtained using the corresponding transition intensity matrix, i.e. $\hat{\mathbf{Q}}(t)$, for $t = 0, 1, \ldots, 4$ respectively. The resulting estimated transition probability matrix and 95% intervals (obtained through the method

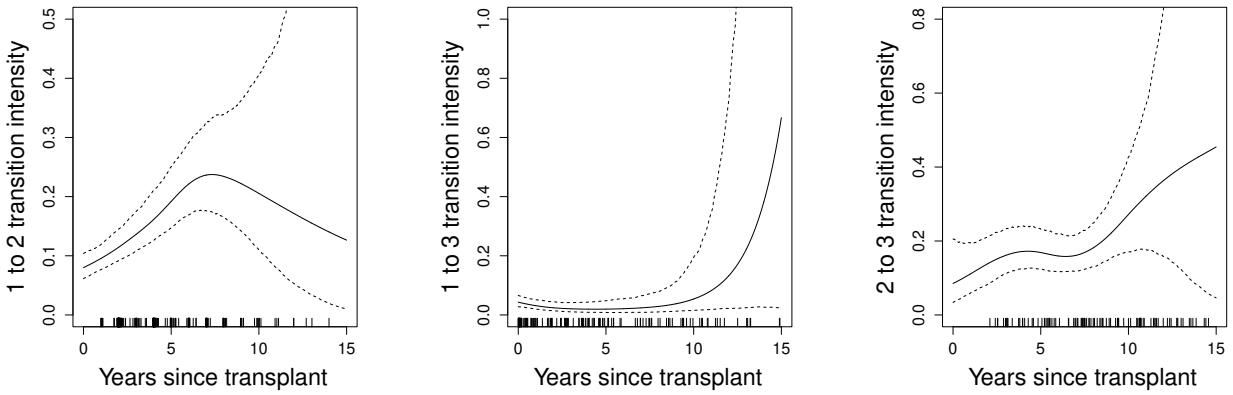


Figure 2: Estimated transition intensities obtained with `flexmsm` for $q^{(12)}(\cdot)$, $q^{(13)}(\cdot)$ and $q^{(23)}(\cdot)$ (from left to right) when `dage = 26` and `pdiag = 1`, with 95% intervals derived as detailed in Section 5. The 'rug plot', at the bottom of each graph, shows the empirical distribution of the transition times. Because we are dealing with an intermittent observation scheme, the time intervals have been represented by plotting the right extremity of each observed interval (the left extremity or mid-point could have been equivalently chosen). Recall that the aim of the rug plot is to highlight regions where the occurrence of a specific transition is rare, hence explaining the width of the intervals across sections.

detailed in Section 5) are

$$\hat{\mathbf{P}}(0,5) = \begin{bmatrix} 0.48 & (0.42, 0.53) & 0.29 & (0.24, 0.34) & 0.23 & (0.19, 0.29) \\ 0 & & 0.51 & (0.35, 0.63) & 0.49 & (0.37, 0.64) \\ 0 & & 0 & & 1 & \end{bmatrix}.$$

For instance, given a healthy starting point, there is a 29% chance of developing CAV five years after the transplant procedure occurred. Similarly, there is a 23% chance of dying within the same time frame, given the same starting point.

We also assessed the possible presence of nonlinear effects of `dage`. This was achieved by replacing $\beta_2^{(rr')}\texttt{dage}_{ij}$ with $s_2^{(rr')}(\texttt{dage}_{ij})$ in model (7), where the smooth terms were represented as for $s_1^{(rr')}(t_{ij})$; the effects were found to be linear. Finally, to illustrate the generality of the proposal, we considered the specification

$$q^{(rr')}(t_{ij}) = \exp\left[\beta_0^{(rr')} + s_1^{(rr')}(t_{ij}) + s_2^{(rr')}(\texttt{dage}_{ij}) + s_3^{(rr')}(t_{ij}, \texttt{dage}_{ij}) + \beta_4^{(rr')}\texttt{pdiag}_{ij}\right],$$

where $s_3^{(rr')}(t_{ij}, \texttt{dage}_{ij})$ is a tensor product interaction between `dage` and time whose marginals are cubic regression splines. Here, the main effects and their interaction are modelled separately, thus leading to more flexibility in determining the complexity of the effects (Wood, 2017, Section 5.6.3). Figure 3 shows the results for transition $1 \to 2$. In the left panel, we report the estimated transition intensity surface, which is a bivariate function of time and `dage`. This plot can be read by sectioning the surface, with respect to either of the two arguments, and assessing how the resulting curve varies with respect to the other covariate. In the right panel, we report two sections

of the surface obtained by fixing `dage` at 26 and 56 years, along with their 95% confidence intervals. The scarcity of data for the two sections helps explaining the wide confidence intervals, particularly past a certain point. For this reason, we will focus the interpretation on the first few years since the transplant took place. One can see that the risk of developing CAV is almost three times higher with a 56 year old donor than it is with a 26 year old donor right after the transplant, and remains higher overall in the following few years. This is in line with expectations that older donors are associated with higher chances of disease occurrence.

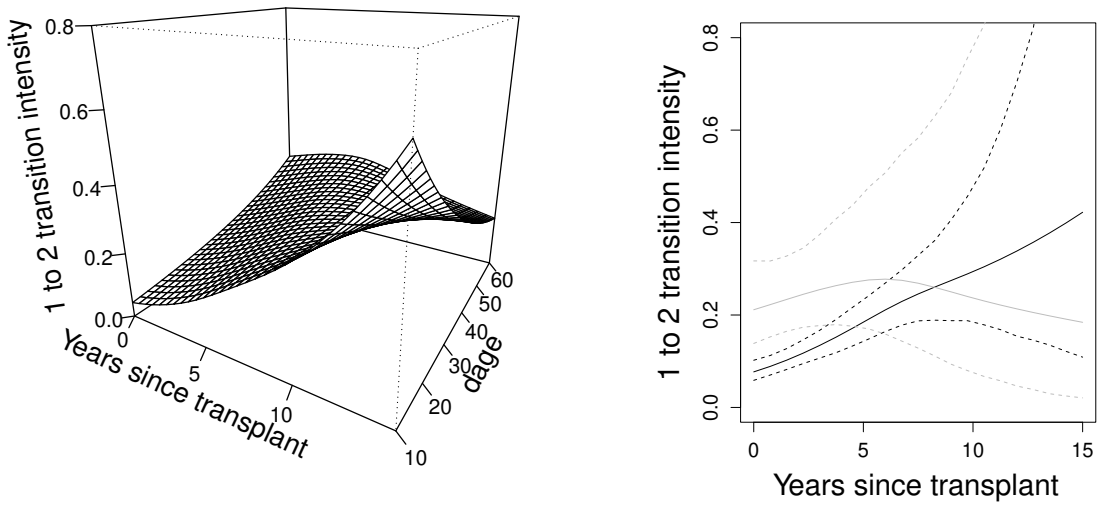


Figure 3: Left panel: estimated transition intensity surface, obtained with `flexmsm` when including a time-dependent effect of the donor age. Right panel: sections of the estimated transition intensity surface at `dage` $= 26$ (black) and `dage` $= 56$ (grey), along with their respective 95% confidence intervals (black and grey dashed lines, respectively).

Supplementary Material D.1 discusses a simulation study based on the IDM. The results support the empirical effectiveness of the proposed modelling framework and the related implementation in `flexmsm`.

6.2 ELSA case study

The ELSA collects data from people aged over 50 to understand all aspects of ageing in England. More than 18000 people have taken part in the study since it started in 2002, with the same people re-interviewed every two years, hence giving rise to an intermittently observed scheme. ELSA collects information on physical and mental health, wellbeing, finances and attitudes around ageing, and tracks how these change over time. The data can be downloaded from the UK Data Service by registering and accepting an End User Licence.

For this study, interest lies in assessing cognitive function in the older population. This is

measured through the score obtained on a test in which participants are asked to remember words in a delayed recall from a list of ten, with the score given by the number of words remembered. In line with Machado et al. (2021), we use a random sample of $N = 1000$ individuals from the full population, leading to 4597 observations, and create four score groups to obtain a five-state process with the fifth state given by the occurrence of death (which is an exactly observed absorbing state). The intermediate states are given by $\{10, 9, 8, 7\}$, $\{6, 5\}$, $\{4, 3, 2\}$ and $\{1, 0\}$ words remembered, respectively. Both forward and backward transitions are allowed between the intermediate states to account for possible improvements or fluctuations through the years in the cognitive function of the participants. In fact, although interest lies mostly in cognitive decline, the opposite trend is also observed as shown in Table 3. A diagram representing the assumed process is reported in Figure 4. Further, 221 participants die during the observation period. The time scale is defined by subtracting 49 years to the age of the individuals.

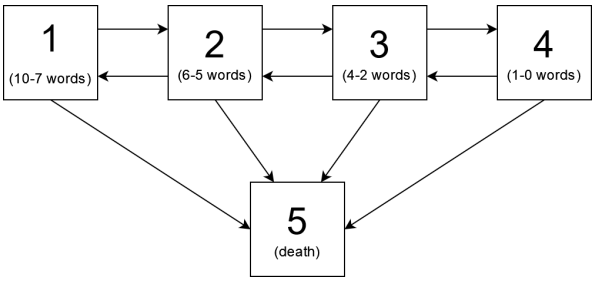


Figure 4: Diagram of the possible five-state process disease trajectories.

	state 1	state 2	state 3	state 4	state 5
state 1	225	194	58	5	11
state 2	209	600	384	54	46
state 3	59	383	732	152	94
state 4	8	42	117	154	70

Table 3: Number of observations for each pair of states in the ELSA dataset.

The most flexible five-state model considered in the literature for the ELSA data is based on Machado et al. (2021)

$$q^{(rr')}(t_{ij}) = \begin{cases} \exp\left[\beta_0^{(rr')} + s_1^{(rr')}(t_{ij})\right] & \text{for } (r,r') \in \{(1,2),(2,3),(2,5),(3,4),(3,5),(4,5)\} \\ \exp\left[\beta_0^{(rr')}\right] & \text{for } (r,r') \in \{(1,5),(2,1),(3,2),(4,3)\} \end{cases},$$

where each smooth term is represented by a cubic regression spline with $J_1^{(rr')} = 5$ and second order penalty, and upper bounds for the smoothing parameters were set at $\exp(20)$. The authors justify the specifications for the $q^{(rr')}(t_{ij})$ and the other settings by arguing that the limited information across the age range is probably what causes algorithmic convergence failures in more general models.

Using the proposed methodology, we considered the general specification

$$q^{(rr')}(t_{ij}) = \exp\left[\beta_0^{(rr')} + s_1^{(rr')}(t_{ij})\right] \text{ for } (r,r') \in \mathcal{A}, \tag{8}$$

with $J_1^{(rr')} = 10$ cubic regression spline bases instead. Figure 5 shows the estimated transition intensities, and related 95% intervals, obtained with `flexmsm`. As expected, the instantaneous

risks of dying are overall smaller than the risks of experiencing further cognitive impairment. As the starting stage reflects more advanced decline, the risk of transitioning to a worse stage becomes a progressively flatter function of time. This shows that once the individuals in the population reach a stage of cognitive impairment, they will typically stay there for the rest of the observation period. Note that there is added value from having modelled the backward transitions through smooth functions of time. For example, we find that the instantaneous chance of improving back to state 3 from a state of cognitive impairment of level 4 decreases considerably faster through time than that of returning to state 1 from state 2. This is in line with expectations as the intermediate stages of cognitive health, i.e. stages 2 and 3, are by far the most frequently observed, with 72% of the population that is still alive at the end of the observation period found in these categories. The wide 95% intervals for transitions $1 \rightarrow 5$ and $2 \rightarrow 5$ can be explained by observing, from Table 3, that these transitions are characterised by the lowest number of observations. Model (8) is general in that no prior assumptions are made with regard to the way each transition depends on time. Instead, they are all defined through flexible functional forms by means of splines. The proposed estimation approach then suppresses any complexity not supported by the data, resulting in final estimated shapes which may be either flat, linear or non-linear. This avoids the need for setting manual constraints or enforcing ad-hoc fixes.

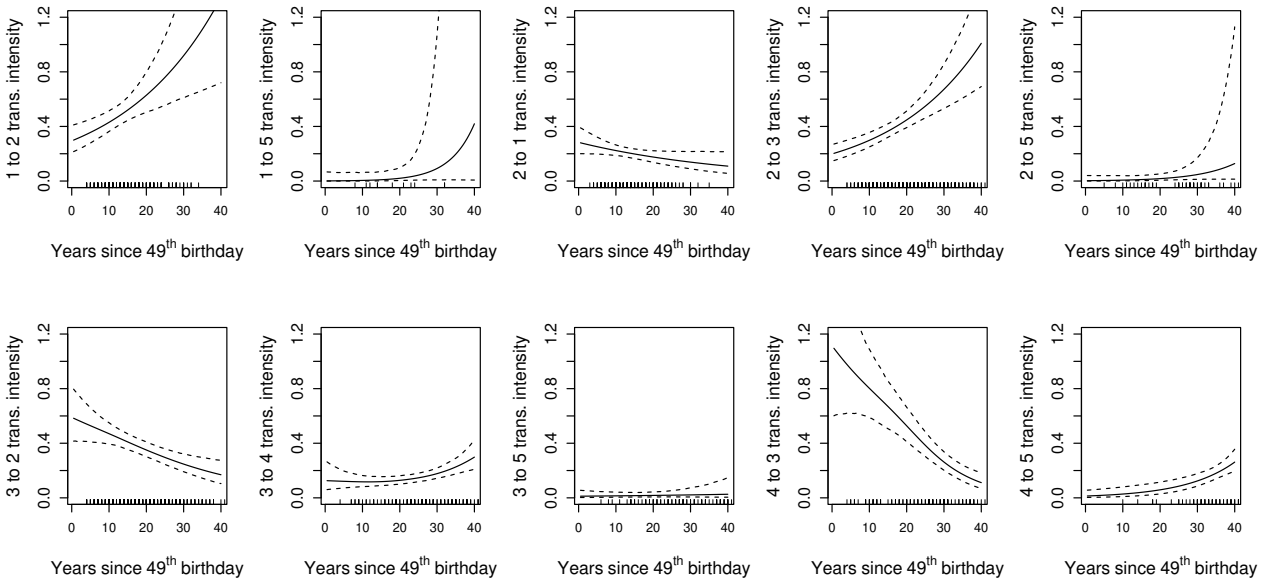


Figure 5: Estimated transition intensities obtained with `flexmsm` with the 95% confidence intervals derived as detailed in Section 5.

We also quantified the effects of two commonly investigated risk factors: `sex` (0 for male and 1 for female) and `higherEdu` (0 if the individual has had less than 10 years of education and 1 otherwise) as extracted from the ELSA datasets. This was achieved by simply including $\beta_2^{(rr')}\texttt{sex}_{ij}$ and $\beta_3^{(rr')}\texttt{higherEdu}_{ij}$ in (8). We found, for example, that older people with a higher level of education have better memory function, although this does not protect them from

cognitive decline as they age (e.g., Cadar et al., 2017). Overall, the effect of `sex` was found not to be significant.

Finally, in Figure 6, we present transition probability plots over 10 years for a 60 year old male with less than 10 years of education. We observe, e.g., that for such individual with stage 2 cognitive health and $\texttt{higherEdu} = 0$, the probability of reaching stage 3 by the age of 65 is approximately 40.3%, with 95% interval $(33.3\%, 44.7\%)$.

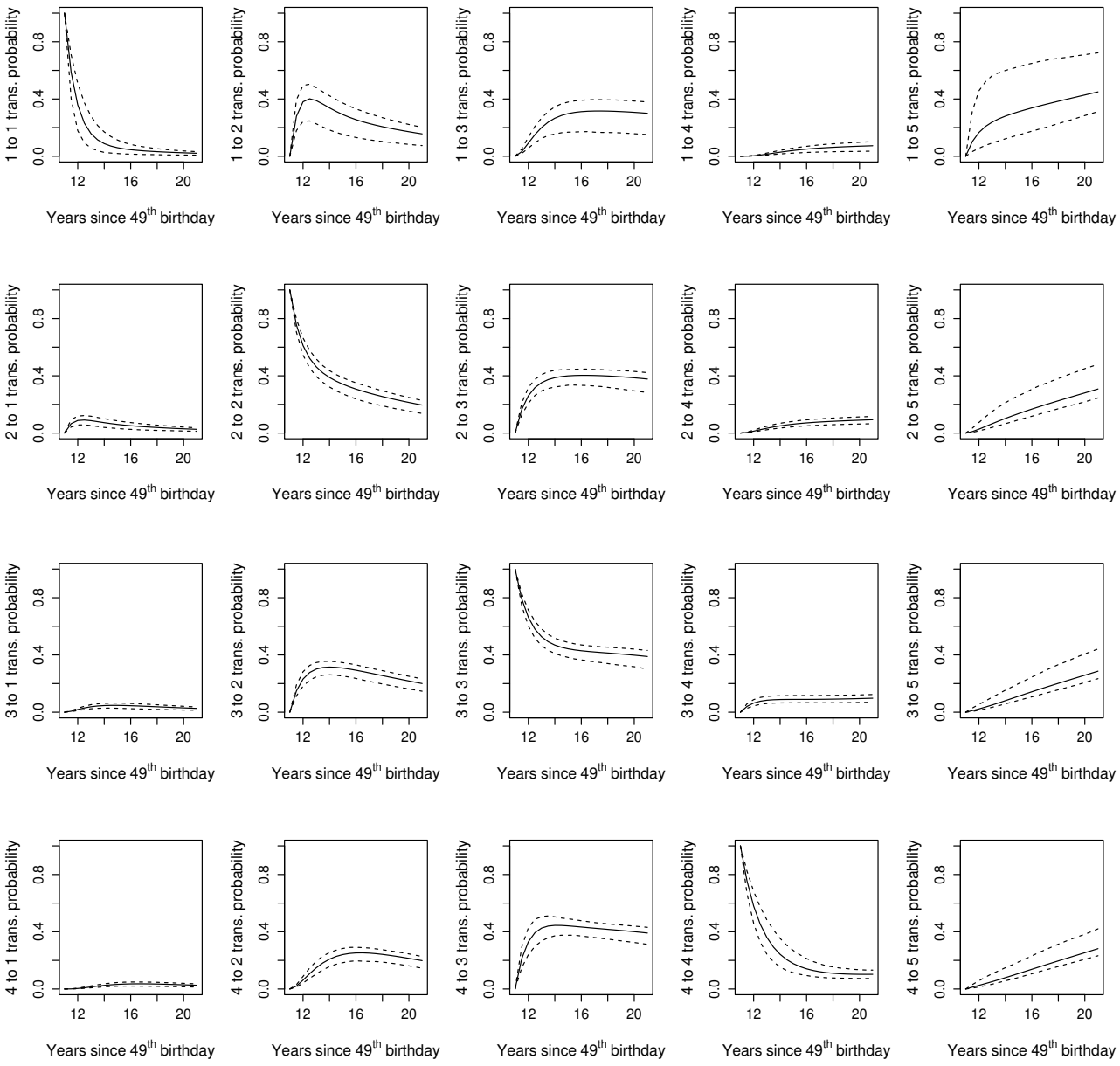


Figure 6: Transition probabilities for a male individual with less than 10 years of education estimated between 11 and 21 years from their 49^{th} birthday, i.e. $\hat{\mathbf{P}}(11,\ t)$ and $t \in (11, 21)$. The dashed lines represent the corresponding 95% intervals.

Supplementary Material D.2 discusses a simulation study based on a five-state process. The results show our framework's ability to recover the true underlying transition intensities in a context

which strays from the traditionally explored IDM.

7 Discussion

We propose a general framework for multi-state Markov modelling that allows for different types of process, with several states and various observation schemes, and that supports time-dependent flexible transition intensities with any type of covariate effects. This is motivated by the interest in modelling the evolution through time of diseases, with the aim of making statements on their course given specific scenarios or risk factors. The degree of flexibility allowed for the specification of the transition intensities determines the extent to which we can explore and describe the different factors influencing the evolution of a disease. Previous methodological developments have mainly focused on simple parametric forms and time-constant transition intensities, which can be attributed to the lack of an estimation framework capable of supporting more realistic specifications. Attempts addressing this have not been backed by adequate estimation procedures and software implementations.

The key development of the paper is the derivation of the local curvature information of a crucial quantity, the transition probability matrix, which has not been attempted to date. Access to this source of information has allowed us to introduce a modelling framework that has unlocked a host of processes and specifications which were not previously attainable, as demonstrated via the two case studies on cardiac allograft vasculopathy and cognitive decline. To support applicability and reproducibility, we also introduced the `R` package `flexmsm`, which is easy and intuitive to use.

Future work will look into further improving the run-time required for model fitting. We are also interested in exploring transformations alternative to the exponential, to enhance the flexibility allowed by the framework. Note that we have assumed a Markov process throughout. Checking whether this property was appropriate for the data considered in this paper was outside of the scope of this work. Future efforts will look into goodness-of-fit tests (e.g., Titman, 2009) as well as the possibility of extending the current model to relax the Markov assumption. There is, however, theoretical and empirical evidence that assuming the Markov property when the true underlying process is non-Markov will still lead to a model that performs well and that has desirable properties (Datta and Satten, 2001; Nießl et al., 2023). Finally, there are circumstances which give rise to multiple dependent multi-state processes, such as the analysis of the evolution of a disease in paired organ systems. In these cases, interest lies in jointly modelling the evolution through time of these events, as the course of one is expected to affect the course of the others. Existing approaches rely on very simple specifications for the marginal processes and restrictive dependence structures among them. The framework proposed in this article will serve as the foundation for the flexible modelling of joint multi-state processes.

Acknowledgments

The ELSA dataset was made available through the UK Economic and Social Data Service. The authors of ELSA do not bear any responsibility for the analyses and interpretations presented in this article.

AE was supported by the UCL Departmental Teaching Assistantship Scholarship. GM and RR were supported by the EPSRC grant EP/T033061/1.

Appendix: Proof of the Theorem

From the definition of matrix exponential we have that

$$\begin{aligned}\frac{\partial^2}{\partial\theta_w\partial\theta_{w'}}\mathbf{P}(\delta t) &= \frac{\partial^2}{\partial\theta_w\partial\theta_{w'}}\sum_{\zeta=0}^{\infty}\frac{(\mathbf{Q}\delta t)^\zeta}{\zeta!}\\ &= \sum_{\zeta=1}^{\infty}\frac{\delta t^\zeta}{\zeta!}\sum_{\rho=0}^{\zeta-1}\mathbf{Q}^\rho\left(\frac{\partial^2\mathbf{Q}}{\partial\theta_w\partial\theta_{w'}}\right)\mathbf{Q}^{\zeta-1-\rho}\\ &+\sum_{\zeta=2}^{\infty}\frac{\delta t^\zeta}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho}\left[\mathbf{Q}^\rho\left(\frac{\partial\mathbf{Q}}{\partial\theta_w}\right)\mathbf{Q}^\kappa\left(\frac{\partial\mathbf{Q}}{\partial\theta_{w'}}\right)\mathbf{Q}^{\zeta-2-\rho-\kappa}\right.\\ &\qquad\left.+\mathbf{Q}^\rho\left(\frac{\partial\mathbf{Q}}{\partial\theta_{w'}}\right)\mathbf{Q}^\kappa\left(\frac{\partial\mathbf{Q}}{\partial\theta_r}\right)\mathbf{Q}^{\zeta-2-\rho-\kappa}\right],\end{aligned}$$

where we used the product rule to re-write the derivative of $\mathbf{Q}^\zeta$. The first summation can be re-written using the same considerations made in Kalbfleisch and Lawless (1985) to find the expression for the gradient. This gives the term $\mathbf{A}\check{\mathbf{U}}_{ww'}\mathbf{A}^{-1}$ reported in the theorem statement. The second summation is made of two addends which are equivalent to each other except that w and w' are swapped, thus we will just show the derivation for the former. In particular, we observe that from the eigendecomposition of $\mathbf{Q}$

$$\begin{aligned}&\mathbf{Q}^\rho\left(\frac{\partial\mathbf{Q}}{\partial\theta_w}\right)\mathbf{Q}^\kappa\left(\frac{\partial\mathbf{Q}}{\partial\theta_{w'}}\right)\mathbf{Q}^{\zeta-2-\rho-\kappa} = (\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^{-1})^\rho\left(\frac{\partial\mathbf{Q}}{\partial\theta_w}\right)(\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^{-1})^\kappa\left(\frac{\partial\mathbf{Q}}{\partial\theta_{w'}}\right)(\mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^{-1})^{\zeta-2-\rho-\kappa}\\ &= \mathbf{A}\boldsymbol{\Gamma}^\rho\underbrace{\mathbf{A}^{-1}\left(\frac{\partial\mathbf{Q}}{\partial\theta_w}\right)\mathbf{A}}_{\mathbf{G}^{(w)}}\boldsymbol{\Gamma}^\kappa\underbrace{\mathbf{A}^{-1}\left(\frac{\partial\mathbf{Q}}{\partial\theta_{w'}}\right)\mathbf{A}}_{\mathbf{G}^{(w')}}\boldsymbol{\Gamma}^{\zeta-2-\rho-\kappa}\mathbf{A}^{-1} = \mathbf{A}\boldsymbol{\Gamma}^\rho\mathbf{G}^{(w)}\boldsymbol{\Gamma}^\kappa\mathbf{G}^{(w')}\boldsymbol{\Gamma}^{\zeta-2-\rho-\kappa}\mathbf{A}^{-1}.\end{aligned}$$

Then the summation $\sum_{\zeta=2}^{\infty}\frac{\delta t^\zeta}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho}\mathbf{A}\boldsymbol{\Gamma}^\rho\mathbf{G}^{(w)}\boldsymbol{\Gamma}^\kappa\mathbf{G}^{(w')}\boldsymbol{\Gamma}^{\zeta-2-\rho-\kappa}\mathbf{A}^{-1}$ becomes

$$\mathbf{A}\left[\sum_{\zeta=2}^{\infty}\frac{\delta t^\zeta}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho}\boldsymbol{\Gamma}^\rho\mathbf{G}^{(w)}\boldsymbol{\Gamma}^\kappa\mathbf{G}^{(w')}\boldsymbol{\Gamma}^{\zeta-2-\rho-\kappa}\right]\mathbf{A}^{-1},$$

where the quantity between brackets leads to the quantity called $\dot{\mathbf{U}}_{ww'}$ in the theorem statement. In fact, note that the $(l,m)^{th}$ element of the inmost term can be re-written as

$$\{\mathbf{\Gamma}^{\rho}\mathbf{G}^{(w)}\mathbf{\Gamma}^{\kappa}\mathbf{G}^{(w')}\mathbf{\Gamma}^{\zeta-2-\rho-\kappa}\}_{lm} = \sum_{y=1}^{Y} G_{ly}^{(w)} G_{ym}^{(w')} \gamma_l^{\rho}\gamma_y^{\kappa}\gamma_m^{\zeta-2-\rho-\kappa}.$$

When we plug this back into the summation defining $\dot{\mathbf{U}}_{ww'}$ we need to distinguish a number of cases (breaking up these cases is allowed by the property that, when the series converges, series of sums or differences are equals to the sums or differences of series):

- When $l = m = y$

$$\begin{aligned}
&\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho} G_{ly}^{(w)} G_{ym}^{(w')} \gamma_l^{\rho}\gamma_y^{\kappa}\gamma_m^{\zeta-2-\rho-\kappa} = \sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho} G_{ll}^{(w)} G_{ll}^{(w')} \gamma_l^{\zeta-2} \\
&= G_{ll}^{(w)} G_{ll}^{(w')} \sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\frac{\zeta(\zeta-1)}{2}\gamma_l^{\zeta-2} = \frac{\delta t^2}{2} G_{ll}^{(w)} G_{ll}^{(w')} \sum_{\zeta=0}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\gamma_l^{\zeta} \\
&= \frac{1}{2} G_{ll}^{(w)} G_{ll}^{(w')} \delta t^2 e^{\gamma_l \delta t}.
\end{aligned}$$

- When $l = m$ and $y \neq l$

$$\begin{aligned}
&\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho} G_{ly}^{(w)} G_{ym}^{(w')} \gamma_l^{\rho}\gamma_y^{\kappa}\gamma_m^{\zeta-2-\rho-\kappa} = G_{ly}^{(w)} G_{yl}^{(w')} \sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho} \gamma_l^{\zeta-2-\kappa}\gamma_y^{\kappa} \\
&= G_{ly}^{(w)} G_{yl}^{(w')} \sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\left[\frac{(\zeta-1)\gamma_l^{\zeta-1}}{\gamma_l-\gamma_y} - \frac{\gamma_l^{\zeta-1}-\gamma_y^{\zeta-1}}{(\gamma_l-\gamma_y)^2}\gamma_y\right],
\end{aligned}$$

  where we used the fact that

$$\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho} \gamma_l^{\zeta-2-\kappa}\gamma_y^{\kappa} = \sum_{\kappa=0}^{\zeta-2}\sum_{\rho=0}^{\kappa} \gamma_l^{\zeta-2-\kappa+\rho}\gamma_y^{\kappa-\rho} = \sum_{\kappa=0}^{\zeta-2}\gamma_l^{\zeta-2-\kappa}\sum_{\rho=0}^{\kappa}\gamma_l^{\rho}\gamma_y^{\kappa-\rho},$$

  through summation swap and change of indices. We then use the result known for the difference of the powers of two terms on the inmost summation, i.e. $(\gamma_l-\gamma_y)\sum_{\rho=0}^{\kappa}\gamma_l^{\rho}\gamma_y^{\kappa-\rho} = \gamma_l^{\kappa+1}-\gamma_y^{\kappa+1}$, and then again on the resulting summations

$$\begin{aligned}
&\sum_{\kappa=0}^{\zeta-2}\gamma_l^{\zeta-2-\kappa}\frac{\gamma_l^{\kappa+1}-\gamma_y^{\kappa+1}}{\gamma_l-\gamma_y} = \frac{(\zeta-1)\gamma_l^{\zeta-1}}{\gamma_l-\gamma_y} - \frac{\gamma_y}{\gamma_l-\gamma_y}\sum_{\kappa=0}^{\zeta-2}\gamma_l^{\zeta-2-\kappa}\gamma_y^{\kappa} \\
&= \frac{(\zeta-1)\gamma_l^{\zeta-1}}{\gamma_l-\gamma_y} - \frac{\gamma_y}{\gamma_l-\gamma_y}\frac{\gamma_l^{\zeta-1}-\gamma_y^{\zeta-1}}{\gamma_l-\gamma_y},
\end{aligned}$$

  which leads to the result above. The series can then be solved by recognising multiple times

the power series expansion of the exponential

$$G_{ly}^{(w)}G_{yl}^{(w')}\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\left[\frac{(\zeta-1)\gamma_l^{\zeta-1}}{\gamma_l-\gamma_y}-\frac{\gamma_l^{\zeta-1}-\gamma_y^{\zeta-1}}{(\gamma_l-\gamma_y)^2}\gamma_y\right]=G_{ly}^{(w)}G_{yl}^{(w')}\left(\frac{\delta t e^{\gamma_l\delta t}}{\gamma_l-\gamma_y}-\frac{e^{\gamma_l\delta t}-e^{\gamma_y\delta t}}{(\gamma_l-\gamma_y)^2}\right).$$

- When $l\neq m$ and $y=l$

$$\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho}G_{ly}^{(w)}G_{ym}^{(w')}\gamma_l^{\rho}\gamma_y^{\kappa}\gamma_m^{\zeta-2-\rho-\kappa}=G_{ll}^{(w)}G_{lm}^{(w')}\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho}\gamma_l^{\rho+\kappa}\gamma_m^{\zeta-2-\rho-\kappa}$$
$$=G_{ll}^{(w)}G_{lm}^{(w')}\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\left[\frac{(\zeta-1)\gamma_l^{\zeta-1}}{\gamma_l-\gamma_m}-\frac{\gamma_l^{\zeta-1}-\gamma_m^{\zeta-1}}{(\gamma_l-\gamma_m)^2}\gamma_m\right],$$

with similar considerations to those of the previous point. Then, similarly to above

$$G_{ll}^{(w)}G_{lm}^{(w')}\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\left[\frac{(\zeta-1)\gamma_l^{\zeta-1}}{\gamma_l-\gamma_m}-\frac{\gamma_l^{\zeta-1}-\gamma_m^{\zeta-1}}{(\gamma_l-\gamma_m)^2}\gamma_m\right]=G_{ll}^{(w)}G_{lm}^{(w')}\left(\frac{\delta t e^{\gamma_l\delta t}}{\gamma_l-\gamma_m}-\frac{e^{\gamma_l\delta t}-e^{\gamma_m\delta t}}{(\gamma_l-\gamma_m)^2}\right).$$

- When $l\neq m$ and $y=m$

$$\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho}G_{ly}^{(w)}G_{ym}^{(w')}\gamma_l^{\rho}\gamma_y^{\kappa}\gamma_m^{\zeta-2-\rho-\kappa}=G_{lm}^{(w)}G_{mm}^{(w')}\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho}\gamma_l^{\rho}\gamma_m^{\zeta-2-\rho}$$
$$=G_{lm}^{(w)}G_{mm}^{(w')}\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\left[\frac{\gamma_l^{\zeta-1}-\gamma_m^{\zeta-1}}{(\gamma_l-\gamma_m)^2}\gamma_l-\frac{(\zeta-1)\gamma_m^{\zeta-1}}{\gamma_l-\gamma_m}\right],$$

with similar considerations to those of the previous point. We then obtain

$$G_{lm}^{(w)}G_{mm}^{(w')}\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\left[\frac{\gamma_l^{\zeta-1}-\gamma_m^{\zeta-1}}{(\gamma_l-\gamma_m)^2}\gamma_l-\frac{(\zeta-1)\gamma_m^{\zeta-1}}{\gamma_l-\gamma_m}\right]=G_{lm}^{(w)}G_{mm}^{(w')}\left(\frac{e^{\gamma_l\delta t}-e^{\gamma_m\delta t}}{(\gamma_l-\gamma_m)^2}-\frac{\delta t e^{\gamma_m\delta t}}{\gamma_l-\gamma_m}\right).$$

- Finally, when $l\neq m\neq y$

$$\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\sum_{\rho=0}^{\zeta-2}\sum_{\kappa=0}^{\zeta-2-\rho}G_{ly}^{(w)}G_{ym}^{(w')}\gamma_l^{\rho}\gamma_y^{\kappa}\gamma_m^{\zeta-2-\rho-\kappa}=\frac{G_{ly}^{(w)}G_{ym}^{(w')}}{\gamma_y-\gamma_m}\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\left[\frac{\gamma_l^{\zeta}-\gamma_y^{\zeta}}{\gamma_l-\gamma_y}-\frac{\gamma_l^{\zeta}-\gamma_m^{\zeta}}{\gamma_l-\gamma_m}\right],$$

which is obtained by repeatedly applying the result known for the difference of the power of two numbers. The series is then solved again by repeatedly recognising the power series expansion of the exponential

$$\frac{G_{ly}^{(w)}G_{ym}^{(w')}}{\gamma_y-\gamma_m}\sum_{\zeta=2}^{\infty}\frac{\delta t^{\zeta}}{\zeta!}\left[\frac{\gamma_l^{\zeta}-\gamma_y^{\zeta}}{\gamma_l-\gamma_y}-\frac{\gamma_l^{\zeta}-\gamma_m^{\zeta}}{\gamma_l-\gamma_m}\right]=\frac{G_{ly}^{(w)}G_{ym}^{(w')}}{\gamma_y-\gamma_m}\left(\frac{e^{\gamma_l\delta t}-e^{\gamma_y\delta t}}{\gamma_l-\gamma_y}-\frac{e^{\gamma_l\delta t}-e^{\gamma_m\delta t}}{\gamma_l-\gamma_m}\right).$$

This concludes the proof of the theorem.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pp. 199–213. Springer.

Cadar, D., A. Robitaille, S. Clouston, S. M. Hofer, A. M. Piccinin, and G. Muniz-Terrera (2017). An international evaluation of cognitive reserve and memory changes in early old age in 10 european countries. *Neuroepidemiology 48*(1-2), 9–20.

Cook, R. J. and J. F. Lawless (2018). *Multistate models for the analysis of life history data*. CRC Press.

Cox, D. R. and H. D. Miller (1977). *The theory of stochastic processes*, Volume 134. CRC press.

Datta, S. and G. A. Satten (2001). Validity of the aalen–johansen estimators of stage occupation probabilities and nelson–aalen estimators of integrated transition hazards for non-markov models. *Statistics & probability letters 55*(4), 403–411.

Eletti, A., G. Marra, and R. Radice (2023). *flexmsm: A General Framework for Flexible Multi-State Survival Modelling*. R package version 0.1.0.

Gorfine, M., N. Keret, A. Ben Arie, D. Zucker, and L. Hsu (2021). Marginalized frailty-based illness-death model: application to the uk-biobank survival data. *Journal of the American Statistical Association 116*(535), 1155–1167.

Gu, Y., D. Zeng, G. Heiss, and D. Lin (2022). Maximum likelihood estimation for semiparametric regression models with interval-censored multi-state data. *arXiv preprint arXiv:2209.07708*.

Jackson, C. (2022). *msm: Multi-State Markov and Hidden Markov Models in Continuous Time*. R package version 1.7.

Jackson, C. H. et al. (2011). Multi-state models for panel data: the msm package for R. *Journal of statistical software 38*(8), 1–29.

Joly, P., D. Commenges, C. Helmer, and L. Letenneur (2002). A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics 3*(3), 433–443.

Kalbfleisch, J. D. and J. F. Lawless (1985). The analysis of panel data under a markov assumption. *Journal of the American Statistical Association 80*(392), 863–871.

Machado, R. J. M., A. Van den Hout, and G. Marra (2021). Penalised maximum likelihood estimation in multi-state models for interval-censored data. *Computational Statistics & Data Analysis 153*, 107057.

Marra, G. and R. Radice (2020). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association 115*(530), 886–895.

Moré, J. J. and D. C. Sorensen (1983). Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing 4*(3), 553–572.

Nießl, A., A. Allignol, J. Beyersmann, and C. Mueller (2023). Statistical inference for state occupation and transition probabilities in non-markov multi-state models subject to both random left-truncation and right-censoring. *Econometrics and Statistics 25*, 110–124.

Nocedal, J. and S. J. Wright (2006). *Numerical Optimization*. Springer-Verlag, New York.

Sabathé, C., P. K. Andersen, C. Helmer, T. A. Gerds, H. Jacqmin-Gadda, and P. Joly (2020). Regression analysis in an illness-death model with interval-censored data: A pseudo-value approach. *Statistical methods in medical research 29*(3), 752–764.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Titman, A. C. (2009). Computation of the asymptotic null distribution of goodness-of-fit tests for multi-state models. *Lifetime Data Analysis 15*(4), 519–533.

Titman, A. C. (2011). Flexible nonhomogeneous markov models for panel observed data. *Biometrics 67*(3), 780–787.

Van Den Hout, A. (2017). *Multi-state survival models for interval-censored data*. CRC Press.

Van den Hout, A. and F. E. Matthews (2008). Multi-state analysis of cognitive ability data: a piecewise-constant model and a weibull model. *Statistics in Medicine 27*(26), 5440–5455.

Williams, J. P., C. B. Storlie, T. M. Therneau, C. R. J. Jr, and J. Hannig (2020). A bayesian approach to multistate hidden markov models: application to dementia progression. *Journal of the American Statistical Association 115*(529), 16–31.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association 99*(467), 673–686.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction With R*. Second Edition, Chapman & Hall/CRC, London.

Wood, S. N. and M. Fasiolo (2017). A generalized fellner-schall method for smoothing parameter optimization with application to tweedie location, scale and shape models. *Biometrics 73*(4), 1071–1081.

Wood, S. N., N. Pya, and B. Säfken (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association 111*(516), 1548–1563.

Yiu, S., V. T. Farewell, and B. D. Tom (2017). Exploring the existence of a stayer population with mover–stayer counting process models: application to joint damage in psoriatic arthritis. *Journal of the Royal Statistical Society. Series C, Applied Statistics 66*(4), 669.

Supplementary Material: "A General Estimation Framework for Multi-State Markov Processes with Flexible Specification of the Transition Intensities"

A Log-likelihood contributions

This section follows Jackson et al. (2011). For a time-inhomogeneous Markov process, the likelihood contribution for the j^{th} observation of the i^{th} unit can take any of the following forms

$$L_{ij}(\boldsymbol{\theta}) = \begin{cases} p^{(z_{ij-1}z_{ij})}(t_{ij-1}, t_{ij}), & \text{if } z_{ij} \text{ is a living state} \\ \exp\left[\int_{t_{ij-1}}^{t_{ij}} q^{(z_{ij-1}z_{ij-1})}(u)du\right] q^{(z_{ij-1}z_{ij})}(t_{ij}), & \text{if } z_{ij} \text{ is an exactly observed living state} \\ \sum_{c=1}^{C} p^{(z_{ij-1}c)}(t_{ij-1}, t_{ij}), & \text{if } z_{ij} \text{ is censored} \\ \sum_{\substack{c=1 \\ c\neq z_{ij}}}^{C} p^{(z_{ij-1}c)}(t_{ij-1}, t_{ij}) q^{(cz_{ij})}(t_{ij}), & \text{if } z_{ij} \text{ is an exactly observed absorbing state} \end{cases}.$$

for $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, with N the total number of statistical units and n_i the number of observations for unit i and where $p^{(z_{ij-1}z_{ij})}(t_{ij-1}, t_{ij}) = P(Z(t_{ij}) = z_{ij} \mid Z(t_{ij-1}) = z_{ij-1})$. In other words, each pair of consecutively observed states contributes one term to the likelihood. Specifically, if a transition between two transient states is observed and the transition time is interval-censored then the contribution is

$$L_{ij}(\boldsymbol{\theta}) = P(Z(t_{ij}) = z_{ij} \mid Z(t_{ij-1}) = z_{ij-1}),$$

If, due to the nature of the process, the transitions to some living states are exactly observed, the contribution is

$$L_{ij}(\boldsymbol{\theta}) = \exp\left[\int_{t_{ij-1}}^{t_{ij}} q^{(z_{ij-1}z_{ij-1})}(u)du\right] q^{(z_{ij-1}z_{ij})}(t_{ij}),$$

since the process is known to have stayed in state z_{ij-1} between t_{ij-1} and t_{ij} and then jumped from state z_{ij-1} to state z_{ij} at exactly t_{ij}. The first term can be explained by observing that

$$\exp\left[\int_{t_{ij-1}}^{t_{ij}} q^{(z_{ij-1}z_{ij-1})}(u)du\right] = \exp\left[-\int_{t_{ij-1}}^{t_{ij}} \sum_{c\neq z_{ij-1}} q^{(z_{ij-1}c)}(u)du\right]$$
$$= \prod_{c\neq z_{ij-1}} \frac{\exp\left[-\int_0^{t_{ij}} q^{(z_{ij-1}c)}(u)du\right]}{\exp\left[-\int_0^{t_{ij-1}} q^{(z_{ij-1}c)}(u)du\right]},$$

which implies that no transition exiting state z_{ij-1} has occurred at time t_{ij} given that it had not occurred by time t_{ij-1} either.

If the state occupied at a given time is unknown then it is said to be censored. In this case, the contribution to the likelihood has to account for all the possible trajectories that may have occurred from the last known occupied state to the current observation time. Therefore, the sum over the various probabilities is taken, which will be null if the transition is not allowed. In particular,

$$L_{ij}(\boldsymbol{\theta}) = \sum_{c=1}^{C} P(Z(t_{ij}) = c \mid Z(t_{ij-1}) = z_{ij-1}).$$

Finally, if the last observed state is an absorbing one then the time at which the transition occurred is generally assumed to be known. In this case, one needs to account for the possibility that the state occupied before the absorbing state is unknown and thus the contribution to the likelihood is summed over the possible states occupied by the process. The information of the exact observation time t_{in_i} is included through the transition intensity computed in that time. Here, we have

$$L_{ij}(\boldsymbol{\theta}) = \sum_{\substack{c=1 \\ c\neq z_{ij}}}^{C} p^{(z_{ij-1}c)}(t_{ij-1}, t_{ij}) q^{(cz_{ij})}(t_{ij}).$$

B R package `flexmsm`

To support applicability and reproducibility, the proposed modelling framework has been implemented in the `R` package `flexmsm`. The package is straightforward to use, especially if the user is already familiar with the syntax of generalised linear models and generalised additive models (GAMs) in `R`. The key function is `fmsm()`, which carries out model fitting and inference, and is exemplified with some of its main arguments in the following code snippet

```
out <- fmsm(formula = formula, data = df,
            id = ID, state = state,
            death = TRUE, living.exact = NULL, cens.state = -99,
            sp.method = 'perf',
            constraint = NULL, parallel = TRUE, ...)
```

where the user specifies the model through the argument `formula` as a `list()` containing the model specifications for the transition intensities, and the dataset has to be provided through the argument `data`. This will always have at least three columns: the state column (whose name is provided through the argument `state`), the column containing the unique IDs (whose name is provided through the argument `id`) identifying each individual, and a column containing the (intermittent) observation times. The arguments `death`, `living.exact` and `cens.state` allow the user to specify the observation type. If the last state in the process is an absorbing state then the user must specify `death = TRUE`; if there are exactly observed living states then the dataset must contain an additional column with `TRUE` (or `1`) if the data point is exactly observed and `FALSE` (or `0`) otherwise; the name of this column must be passed through the argument `living.exact`, which defaults to `NULL`. If there are any censored states then the user must specify the code used to indicate this through the argument `cens.state`, which defaults to -99. The `sp.method` argument specifies the method employed for multiple smoothing parameter estimation (this can be set to `'perf'` or `'efs'`). The argument `constraint` allows the user to specify equality constraints on the covariates. The `parallel` argument allows the user to exploit parallel computing, in Windows, for the likelihood, gradient and Hessian, thus cutting the run-time of the algorithm by factor proportional to the number of cores on the computer.

The `formula` is a `list()` object whose elements are the off-diagonal elements of the transition intensity matrix. The order of the elements is that given by reading the $\mathbf{Q}$ matrix from the first row to the last and from left to right. The equation corresponding to each non-zero transition intensity has to be specified with syntax similar to that used for GAMs, with the response given by the time-to-event variable. Trivially, zero elements have to be specified with a `0`. For instance, we may consider the following model, with a smooth effect of time t and two covariates x_1 and

x_2, one included linearly and the other as a time-dependant flexible effect, for a transition $r \to r'$

$$q^{(rr')}(t_{ij}) = \exp\left[\beta_0^{(rr')} + s_1^{(rr')}(t_{ij}) + \beta_2^{(rr')} x_{1ij} + s_3^{(rr')}(x_{2ij}) + s_4^{(rr')}(t_{ij}, x_{2ij})\right].$$

This will be specified, in the correct position, as part of the list

```
formula <- list(...,
                t ~ s(t) + x1 + s(x2), ti(t, x2), # r -> r' trans.
                ...)
```

where `...` represent other possible transition-specific equations or `0`s for transitions not allowed by the process. The model specified here is only an example and many types of effects are supported. For instance, as the above example shows, time-dependent effects are modelled by using a tensor interaction function `ti()` on the covariate of interest and time.

Functions `summary()` and `plot()` can be used in the usual way to obtain post-estimation summaries for each non-zero transition intensity and the plots of the smooths. In the example above there is a two-dimensional spline, thus `plot()` will also automatically produce a three-dimensional plot of the surface representing this time-dependent effect.

Function `conv.check()` allows the user to check the convergence of the fitted model by providing information on whether the gradient is zero and the Hessian is positive definite. It also provides information on the values taken by the **Q** matrix since, in practice, we have found that particularly large values are red flags for ill-defined problems, for instance.

Prediction and plotting of the **P** and the **Q** matrices can be carried out through the functions `P.pred()` and `Q.pred()`, respectively. For instance, the specification

```
P.hat <- P.pred(out, newdata = newdata, plot.P = TRUE
                get.CI = TRUE, prob.lev = 0.05)
```

will provide an object `P.hat` containing the estimated transition probability matrix corresponding to the time interval and profile of interest, specified through argument `newdata`. The intermediate transition probabilities corresponding to each sub-interval specified in `newdata` are also provided. The $100(1-$`prob.lev`$)\%$ confidence intervals can be obtained by setting `get.CI = TRUE`. When `plot.P = TRUE` the transition probabilities are also plotted as function of time over the interval considered, otherwise the plots are suppressed. The analogous output can be obtained for the **Q** matrix through function `Q.pred()` with similar syntax.

To exemplify the usage of the software, we report the code used to fit the models presented in Section 6. We recall that the IDM specified in Section 6.1 is given by

$$q^{(rr')}(t_{ij}) = \exp\left[\beta_0^{(rr')} + s_1^{(rr')}(t_{ij}) + \beta_2^{(rr')}\texttt{dage}_{ij} + \beta_3^{(rr')}\texttt{pdiag}_{ij}\right].$$

This can be fitted in the following way:

```
formula <- list(t ~ s(t, bs = 'cr', k = 10) + dage + pdiag, # 1-2
                t ~ s(t, bs = 'cr', k = 10) + dage + pdiag, # 1-3
                0,                                          # 2-1
                t ~ s(t, bs = 'cr', k = 10) + dage + pdiag, # 2-3
                0,                                          # 3-1
                0)                                          # 3-2

fmsm.out <- fmsm(formula = formula, data = Data,
                 id = PTNUM, state = state, death = TRUE,
                 sp.method = 'perf', parallel = TRUE)
```

Here `bs = 'cr'` and `k = 10` imply that the smooths of time are specified through cubic regression splines with ten basis functions. We will omit this in the following to avoid redundancies. To obtain the two-dimensional spline based model, it suffices to swap the `formula` reported above with the following

```
formula <- list(t ~ s(t) + s(dage) + ti(t, dage) + pdiag, # 1-2
                t ~ s(t) + s(dage) + ti(t, dage) + pdiag, # 1-3
                0,                                        # 2-1
                t ~ s(t) + s(dage) + ti(t, dage) + pdiag, # 2-3
                0,                                        # 3-1
                0)                                        # 3-2
```

For the five-state model described in Section 6.2, the first model explored was

$$q^{(rr')}(t_{ij}) = \exp\left[\beta_0^{(rr')} + s_1^{(rr')}(t_{ij})\right].$$

This can be implemented in the following way:

```
formula <- list(t ~ s(t) + sex + edu, # 1-2
                0,                    # 1-3
                0,                    # 1-4
                t ~ s(t) + sex + edu, # 1-5
                t ~ s(t) + sex + edu, # 2-1
                t ~ s(t) + sex + edu, # 2-3
                0,                    # 2-4
                t ~ s(t) + sex + edu, # 2-5
                0,                    # 3-1
                t ~ s(t) + sex + edu, # 3-2
                t ~ s(t) + sex + edu, # 3-4
```

```
                t ~ s(t) + sex + edu, # 3-5
                0,                    # 4-1
                0,                    # 4-2
                t ~ s(t) + sex + edu, # 4-3
                t ~ s(t) + sex + edu, # 4-5
                0,                    # 5-1
                0,                    # 5-2
                0,                    # 5-3
                0)                    # 5-4

fmsm.out <- fmsm(formula = formula, data = ELSA.df,
              id = idauniq, state = state, death = TRUE,
              sp.method = 'efs')
```

C Parameter estimation

The algorithm employed for model fitting is characterised by two steps. In the first step, $\boldsymbol{\lambda}$ is held fixed at a vector of values and for a given $\boldsymbol{\theta}^{[a]}$, where a is an iteration index, equation (4) is maximised using

$$\boldsymbol{\theta}^{[a+1]} = \boldsymbol{\theta}^{[a]} + \underset{\mathbf{e}:\|\mathbf{e}\|\leq\Delta^{[a]}}{\arg\min}\ \breve{\ell}_p(\boldsymbol{\theta}^{[a]}), \tag{9}$$

where $\breve{\ell}_p(\boldsymbol{\theta}^{[a]}) = -\left\{\ell_p(\boldsymbol{\theta}^{[a]}) + \mathbf{e}^\mathsf{T}\mathbf{g}_p(\boldsymbol{\theta}^{[a]}) + \frac{1}{2}\mathbf{e}^\mathsf{T}\mathbf{H}_p(\boldsymbol{\theta}^{[a]})\mathbf{e}\right\}$, $\mathbf{g}_p(\boldsymbol{\theta}^{[a]}) = \mathbf{g}(\boldsymbol{\theta}^{[a]}) - \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\theta}^{[a]}$, and $\mathbf{H}_p(\boldsymbol{\theta}^{[a]}) = \mathbf{H}(\boldsymbol{\theta}^{[a]}) - \mathbf{S}_{\boldsymbol{\lambda}}$. $\mathbf{g}(\boldsymbol{\theta}^{[a]}) = \partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[a]}}$ and $\mathbf{H}(\boldsymbol{\theta}^{[a]}) = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[a]}}$ are given in Section 4, $\|\cdot\|$ denotes the Euclidean norm, and $\Delta^{[a]}$ is the radius of the trust region which is adjusted through the iterations. The first line of (9) uses a quadratic approximation of $-\ell_p$ about $\boldsymbol{\theta}^{[a]}$ (the so-called model function) to choose the best $\mathbf{e}^{[a+1]}$ within the ball centered in $\boldsymbol{\theta}^{[a]}$ of radius $\Delta^{[a]}$, the trust-region. Throughout the iterations, a proposed solution is accepted or rejected and the trust region adjusted (i.e., expanded or shrunken) based on the ratio between the improvement in the objective function when going from $\boldsymbol{\theta}^{[a]}$ to $\boldsymbol{\theta}^{[a+1]}$ and that predicted by the approximation. The use of the observed information matrix gives global convergence guarantees due to Moré and Sorensen (1983). Importantly, convergence to a point satisfying the second-order sufficient conditions (i.e., a local strict minimiser) is super-linear. Near the solution, the algorithm proposals become asymptotically similar to Newton-Raphson steps, hence benefitting from the resulting fast convergence rate. Trust region algorithms are also generally more stable and faster compared to in-line search methods. See Nocedal and Wright (Chapter 4, 2006) for proofs and further details. The starting values $\boldsymbol{\theta}^{[0]}$ are set automatically to small positive values, except for the transition-specific intercepts which are given by the maximum likelihood estimates one would obtain when assuming that the data represent the exact transition times of the corresponding covariate-free time-homogeneous Markov process. Vector $\boldsymbol{\theta}^{[0]}$ can, alternatively, be provided by the user. Importantly, through extensive experimentation, we have found that the algorithm is not particularly sensitive to the choice of starting values.

In the second step, at $\boldsymbol{\theta}^{[a+1]}$, there are two options to estimate the smoothing parameter vector: the stable and efficient multiple smoothing parameter approach adopted by Marra and Radice (2020), and the generalised Fellner-Schall method of Wood and Fasiolo (2017). Both techniques can be employed for fitting penalised likelihood-based models, and require the availability of the analytical score and information matrix. In the former, the following problem is solved

$$\boldsymbol{\lambda}^{[a+1]} = \underset{\boldsymbol{\lambda}}{\arg\min}\ \|\mathbf{M}^{[a+1]} - \mathbf{O}^{[a+1]}\mathbf{M}^{[a+1]}\|^2 - \check{n} + 2\text{tr}(\mathbf{O}^{[a+1]}). \tag{10}$$

The idea is to estimate $\boldsymbol{\lambda}$ so that the complexity of the smooth terms not supported by the data is suppressed. This is formalised as $\mathbb{E}\left(\|\boldsymbol{\mu}_\mathbf{M} - \widehat{\boldsymbol{\mu}}_\mathbf{M}\|^2\right) = \mathbb{E}\left(\|\mathbf{M} - \mathbf{OM}\|^2\right) - \check{n} + 2\text{tr}(\mathbf{O})$, where $\mathbf{M} =$

$\boldsymbol{\mu}_{\mathbf{M}} + \boldsymbol{\epsilon}$, $\boldsymbol{\mu}_{\mathbf{M}} = \sqrt{-\mathbf{H}(\boldsymbol{\theta})}\boldsymbol{\theta}$, $\boldsymbol{\epsilon} = \sqrt{-\mathbf{H}(\boldsymbol{\theta})}^{-1}\mathbf{g}(\boldsymbol{\theta})$, $\mathbf{O} = \sqrt{-\mathbf{H}(\boldsymbol{\theta})}\left(-\mathbf{H}(\boldsymbol{\theta}) + \mathbf{S}_{\boldsymbol{\lambda}}\right)^{-1}\sqrt{-\mathbf{H}(\boldsymbol{\theta})}$, and $\mathrm{tr}(\mathbf{O})$ is defined in Section 5 of the main paper. It can be proved that (10) is approximately equivalent to the AIC with number of parameters given by $\mathrm{tr}(\mathbf{O})$. Iteration (10) is implemented via the routine by Wood (2004), which is based on the Newton method and can evaluate in an efficient and stable manner the terms in (10), their scores and Hessians, with respect to $\log(\boldsymbol{\lambda})$.

The approach proposed in Wood and Fasiolo (2017) is based on a different principle. The starting point is the well established stance that smoothing penalties can be viewed as resulting from improper Gaussian prior distributions on the spline coefficients. This is also the Bayesian viewpoint taken for the inferential result discussed in Section 5, and implies the following improper joint log-density, where the dependence on the smoothing parameter has been made explicit,

$$\log L(\boldsymbol{\theta}; \boldsymbol{\lambda}) = \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\theta} + \frac{1}{2}\log|\mathbf{S}_{\boldsymbol{\lambda}}|.$$

The idea is to develop an update for $\boldsymbol{\lambda}$ that maximises the restricted marginal likelihood $L(\boldsymbol{\lambda})$, obtained integrating $\boldsymbol{\theta}$ out of the likelihood $L(\boldsymbol{\theta}; \boldsymbol{\lambda})$. It is, however, more computationally efficient and equally theoretically founded to maximise the log Laplace approximation

$$\ell_{LA}(\boldsymbol{\lambda}) = \ell(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\hat{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\hat{\boldsymbol{\theta}} + \frac{1}{2}\log|\mathbf{S}_{\boldsymbol{\lambda}}| - \frac{1}{2}\log|-\mathbf{H}(\hat{\boldsymbol{\theta}}) + \mathbf{S}_{\boldsymbol{\lambda}}|,$$

where $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \boldsymbol{\lambda})$ for a given $\boldsymbol{\lambda}$. At $\boldsymbol{\theta}^{[a+1]}$, the update for the k^{th} element of $\boldsymbol{\lambda}^{(rr')}$ for all $(r, r') \in \mathcal{A}$ is

$$\lambda_k^{(rr')[a+1]} = \lambda_k^{(rr')[a]} \times \frac{\mathrm{tr}\left\{\mathbf{S}_{\boldsymbol{\lambda}^{[a]}}^{-1}\dfrac{\partial \mathbf{S}_{\boldsymbol{\lambda}}}{\lambda_k^{(rr')}}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[a]}}\right\} - \mathrm{tr}\left\{\left[-\mathbf{H}(\hat{\boldsymbol{\theta}}) + \mathbf{S}_{\boldsymbol{\lambda}^{[a]}}\right]^{-1}\dfrac{\partial \mathbf{S}_{\boldsymbol{\lambda}}}{\partial \lambda_k^{(rr')}}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[a]}}\right\}}{\hat{\boldsymbol{\theta}}^{\top}\left(\dfrac{\partial \mathbf{S}_{\boldsymbol{\lambda}}}{\partial \lambda_k^{(rr')}}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[a]}}\right)\hat{\boldsymbol{\theta}}}, \qquad (11)$$

with $k = 1, \ldots, K^{(rr')}$. The two steps, (9) and either (10) or (11), are iterated until the algorithm satisfies the stopping rule $\frac{\left|\ell(\boldsymbol{\theta}^{[a+1]}) - \ell(\boldsymbol{\theta}^{[a]})\right|}{0.1 + \left|\ell(\boldsymbol{\theta}^{[a+1]})\right|} < 1e-07$, and convergence is assessed by checking that the maximum of the absolute value of the gradient vector is numerically equivalent to 0 and that the observed information matrix is positive definite. In practice, we found the two smoothing methods to yield similar smooth term estimates.

As with any estimation algorithm, convergence failures may occur. With multi-state models, we mainly found this to be the case when the information provided by the data is insufficient to support the model specified. For instance, when a transition is characterised by a low number of observations, empirical identification of a non-trivial model may not possible. And this tends to be independent of the starting values and of the smoothing method chosen. Such pathological behaviour can often already be spotted in the first few iterations of the optimisation algorithm, with the proposed estimates leading to very large transition intensity values ($> 10^5$).

D Simulation study

To exemplify the empirical effectiveness of the proposed approach in recovering the true values of key quantities of interest (e.g., transition intensity curves), we carried out two simulation studies. The first one replicates that designed in Machado et al. (2021) and uses an IDM set-up. The second study is about a five-state Markov process and serves to illustrate the performance of the proposal in a setting that is more complex than those supported by the methods available in the literature.

D.1 IDM based simulation

We consider a progressive IDM, assuming a different time-dependent shape for each of the three allowed transitions. The time-to-events relating to transition $1 \to 2$ are simulated from a log-normal distribution with location 1.25 and scale 1. This implies that the hazard increases first and then decreases at a later time. For $1 \to 3$, an exponential distribution with rate $\exp(-2.5)$ is employed. For $2 \to 3$, we assume a strictly increasing hazard by simulating the time-to-events from a conditional Gompertz distribution with rate $\exp(-2.5)$ and shape 0.1. For this transition, we have to condition on the event that the individual transitions to state 2 to ensure that the simulated time is larger than the $1 \to 2$ transition time. As in Machado et al. (2021), we simulate $N = 500$ trajectories (i.e., individuals) $\mathcal{M} = 100$ times. Tests with larger $\mathcal{M}$ confirmed the results reported below, thus we kept $\mathcal{M} = 100$ to retain the comparability with Machado et al. (2021).

More specifically, let $T_{rs} = T_{rs|u}$ represent the time of the transition to state r' conditional on being in state *r* at time $u > 0$. If the state at *u* is 1 then the time of transition to the next state can be obtained by taking $T = \min\{T_{12}, T_{13}\}$. If $T = T_{12}$ then the next state is 2, otherwise the next state is 3. If the state is 2 then the time of the next state is T_{23}. Censoring needs to be imposed to render the data intermittently observed; we assume a yearly time-grid spanning over 15 years, i.e. $(t_{i0}, t_{i1}, \ldots, \min\{t_{i15}, T_{13}\}) = (0, 1, \ldots, \min\{t_{i15}, T_{13}\})$ for $i = 1, \ldots, N$. The reader is referred to Van Den Hout (2017) for further details on how to simulate intermittently-observed multi-state survival data. The transition intensities are specified as $q^{(rr')}(t) = \exp\left[\beta_0^{(rr')} + s_1^{(rr')}(t)\right]$ for $(r, r') \in \{(1,2), (1,3), (2,3)\}$, where $s_1^{(rr')}(t)$ is represented using a cubic regression spline with $J_1^{(rr')} = 10$ and second order penalty.

In line with Machado et al. (2021), Figure 7 shows the estimated median and true hazards as well as all the $\mathcal{M}$ estimated hazards. Note that the large variation observed towards the end of the study time is due to scarceness of data at later years. Overall, the plots show that the proposed approach is able to recover well the true transition intensity curves for each allowed transition, and that the performance is similar across the two methods. The discrepancy between fitted median and true hazards for transition $1 \to 2$ is due to definition of interval censoring adopted in the simulation study: the sampling design implies that the living states are observed at intervals of one year; for the first two years after baseline, this design does not work well.

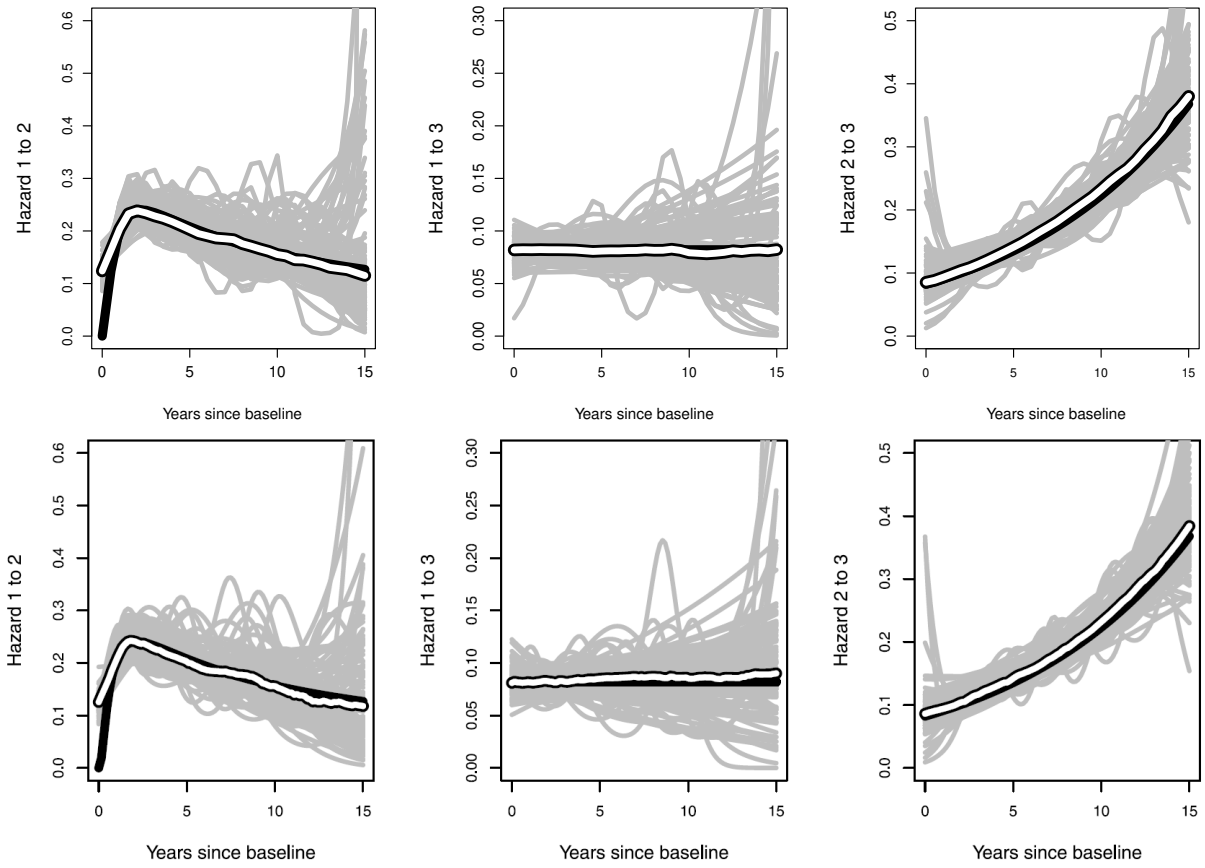


Figure 7: True (black), estimated (grey, $\mathcal{M} = 100$ replicates) and median estimated (white) hazard functions for transitions $1 \rightarrow 2$ (left), $1 \rightarrow 3$ (middle) and $2 \rightarrow 3$ (right) obtained by `flexmsm` (top row) and Machado et al. (2021) (bottom row).

We also evaluated our approach on the transition probability scale. In particular, Table 4 reports the true, average and median ten-year estimated transition probabilities, where the average is taken over the $\mathcal{M}$ simulations. The biases are also reported and are defined as $\text{Bias}^{(rr')}(t) = \frac{1}{\mathcal{M}}\left(\sum_{\nu=1}^{\mathcal{M}} p^{(\nu,rr')}(0,10) - p^{(rr')}(0,10)\right)$, where $p^{(\nu,rr')}(0,10)$ denotes the estimated ten-year probability of transitioning from state r to state r' for the ν^{th} simulated dataset. Our methodology recovers well the true ten-year transition probabilities and consistently outperforms the approach of Machado et al. (2021).

True	`flexmsm`		M. et al. (2021)	
	Mean	Bias	Mean	Bias
$p^{(11)}(0,10) = 0.065$	0.063	−0.002	0.060	0.004
$p^{(12)}(0,10) = 0.231$	0.232	0.001	0.222	0.009
$p^{(13)}(0,10) = 0.704$	0.705	0.001	0.718	−0.014
$p^{(22)}(0,10) = 0.245$	0.242	−0.003	0.231	0.014
$p^{(23)}(0,10) = 0.755$	0.758	0.003	0.769	−0.014

Table 4: Ten-year true and average estimated transition probabilities, and bias for $\mathcal{M} = 100$ replicates.

Finally, we explored the effect that the length of the gap occurring between two successive observations has on estimation performance; it is known that when such gap is large, identifiability issues may arise. To this end, we additionally considered two-, three-, four- and five-yearly time-grids. As expected, the performance deteriorated as the gap increased, with reasonable results (not reported here, but available upon request) still attainable for two- and three-yearly time-grids.

D.2 Five-state process based simulation

We consider a progressive five-state survival process with an absorbing state, and seven transitions whose parameters were chosen to produce intensities similar to those found in the ELSA case study described in Section 6.2. In particular, we simulate the time-to-events from (conditional) Gompertz distributions with rates and shapes provided for each transition in Table 5. We simulate $N = 500$ trajectories $\mathcal{M} = 100$ times, which are observed for 40 semesters. An intermittently observation scheme is imposed by assuming that individuals are visited every 4 semesters. The time is then brought back to the year scale. This gives counts of pairs of consecutively observed states that are similar to those found in the ELSA case study.

	$1 \to 2$	$1 \to 5$	$2 \to 3$	$2 \to 5$	$3 \to 4$	$3 \to 5$	$4 \to 5$
rate	$\exp(-2.25)$	$\exp(-5)$	$\exp(-2.20)$	$\exp(-5)$	$\exp(-2)$	$\exp(-5)$	$\exp(-3)$
shape	0.06	0.02	0.05	0.09	0.01	0.02	0.04

Table 5: Rates and shapes for the (conditional) Gompertz distributions generating the transition times in the five-state process based simulation.

The transition intensities are specified as $q^{(rr')}(t) = \exp\left[\beta_0^{(rr')} + s_1^{(rr')}(t)\right]$ for $(r, r') \in \{(1,2),(1,5),(2,3),(2,5),(3,4),(3,5),(4,5)\}$, where $s_1^{(rr')}(t)$ is represented using a cubic regression spline with $J_1^{rs} = 10$ and second order penalty.

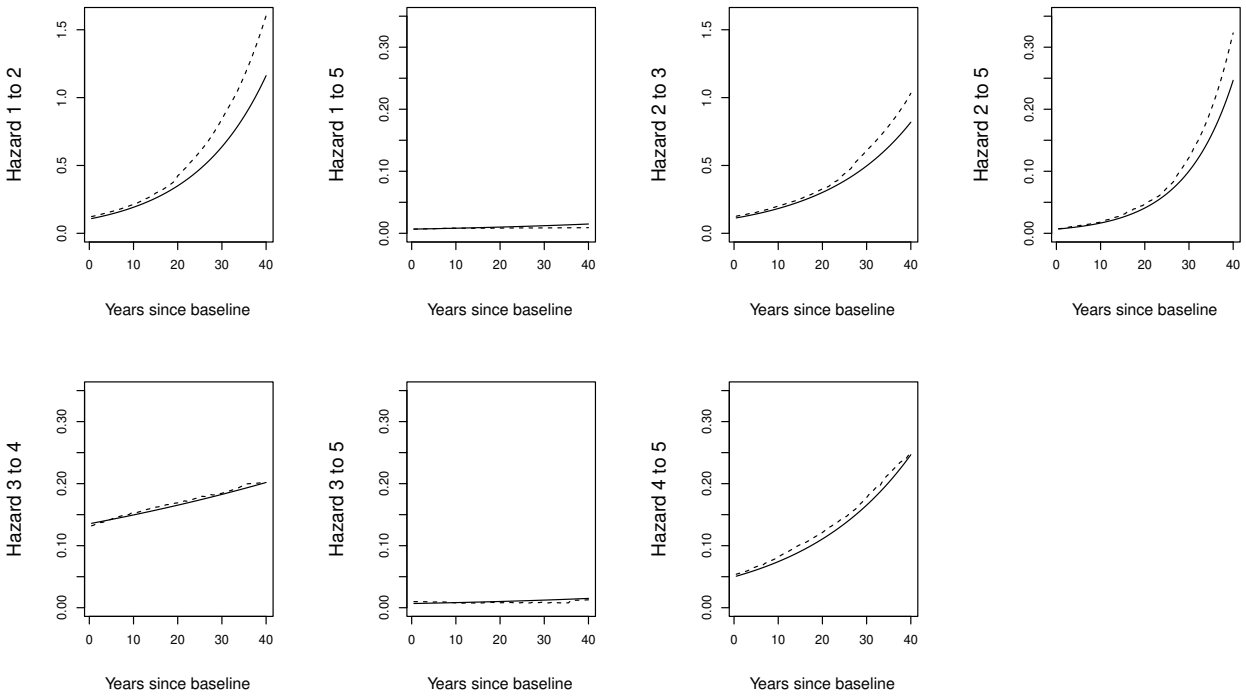


Figure 8: True (black) and median estimated (dashed) hazard functions for each transition in the simulated five-state process.

In Figure 8, we report the median estimated transition intensities obtained for the $\mathcal{M}$ simula-

tions with our framework, alongside the true curve $q^{(rr')}(t)$, for each of the seven allowed transitions. Overall, the proposed approach recovers adequately the true transition intensity curves.

True	Mean	Bias
$p^{(11)}(0,10) = 0.229$	0.192	-0.037
$p^{(12)}(0,10) = 0.318$	0.300	-0.018
$p^{(13)}(0,10) = 0.230$	0.255	0.025
$p^{(14)}(0,10) = 0.121$	0.137	0.016
$p^{(15)}(0,10) = 0.102$	0.116	0.014
$p^{(22)}(0,10) = 0.222$	0.186	-0.036
$p^{(23)}(0,10) = 0.330$	0.333	0.003
$p^{(24)}(0,10) = 0.294$	0.299	0.006
$p^{(25)}(0,10) = 0.154$	0.181	0.027
$p^{(33)}(0,10) = 0.225$	0.222	-0.003
$p^{(34)}(0,10) = 0.508$	0.481	-0.027
$p^{(35)}(0,10) = 0.267$	0.297	0.03
$p^{(44)}(0,10) = 0.549$	0.527	-0.021
$p^{(45)}(0,10) = 0.451$	0.473	0.021

Table 6: Ten-year true, average and median transition probabilities for our framework. The order is that found when reading the transition probability matrix row-wise.

As done for the three-state simulated process, we also evaluate our approach on the transition probabilities scale. In Table 6, we report the true and average ten-year estimated transition probabilities, where the average is taken over the $\mathcal{M}$ simulations, and the corresponding biases. The method is able to recover the true ten-year transition probabilities reasonably well, exhibiting consistently small biases. This is reassuring considering the multi-state process adopted here, which is more involved and complex that those commonly explored and used in the literature.

E List of symbols

Covariates and functions or longer terms

- age_ι covariate in model
- Bias$^{(rr')}(t)$ bias relating to the $r \to r'$ transition at time t in the simulation study
- $\texttt{dage}_{ij}$ covariate in CAV model
- $\texttt{pdiag}_{ij}$ covariate in CAV model
- sex_ι covariate in model
- $\texttt{sex}_{ij}$ covariate in ELSA model
- $\texttt{higherEdu}_{ij}$ covariate in ELSA model
- *edf* for effective degrees of freedom
- tr$(\cdot)$ trace function
- $\mathbf{1}_{\check{n}}$ vector of 1s of length $\check{n}$.

Latin letters

- a estimation algorithm iteration index.
- **A** matrix of eigenvectors.
- $\mathcal{A}$ set of allowed transitions
- $\mathbf{b}_k^{(rr')}(\tilde{\mathbf{x}}_{k\iota})$ bases function vector for the k^{th} term in the (r, r') transition intensity.
- c indexing for likelihood contributions (censored state contribution and for exactly observed absorbing state).
- C total number of states.
- d_υ difference of knots in the construction of the cubic regression spline.
- $\mathbf{D}_k^{(rr')}$ penalty matrix for the k^{th} term in the (r, r') transition intensity.
- **e** vector in the Taylor approximation.
- $\mathbb{E}$ expectation function.

- $f_{\boldsymbol{\theta}}$ prior on the model parameter $\boldsymbol{\theta}$.
- $G_{lm}^{(w)}$ the (l, m) element of $\mathbf{G}^{(w)}$.
- $G_{lm}^{(ww')}$ the (l, m) element of $\mathbf{G}^{(ww')}$.
- $\mathbf{g}(\boldsymbol{\theta})$ gradient vector.
- $\mathbf{G}^{(w)}$ matrix needed for the closed form expression of $\partial^2\mathbf{P}$ (transformation of first derivative of Q matrix).
- $\mathbf{G}^{(ww')}$ matrix needed for the closed form expression of $\partial^2\mathbf{P}$ (transformation of second derivative of Q matrix).
- h infinitesimal time in the limit-based definition of the transition intensity.
- $\mathbf{H}(\boldsymbol{\theta})$ hessian matrix.
- $\mathbf{H}_p(\boldsymbol{\theta})$ penalised hessian matrix.
- i indexing for the statistical units when defining the likelihood. Here $i = 1, \ldots, N$.
- j indexing for the observations of a specific statistical unit.
- $J_k^{(rr')}$ number of basis functions for the k^{th} term in (r, r') transition intensity.
- k indexing for overall covariate/parameter vector, with $k = 1, \ldots, K^{(rr')}$.
- $K^{(rr')}$ total number of terms in additive predictor $\eta_i^{(rr')}(t_i, \mathbf{x}_i; \boldsymbol{\beta}^{(rr')})$, excluding the intercept.
- l indexing for the (l, m) element of the matrices needed for the closed form expression of $\partial^2\mathbf{P}$.
- ℓ_{LA} log Laplace approximation of $L(\boldsymbol{\lambda})$.
- $\ell(\boldsymbol{\theta})$ model log-likelihood.
- $\ell_p(\boldsymbol{\theta})$ penalised log-likelihood.
- $\breve{\ell}_p(\boldsymbol{\theta})$ second order approximation of the model log-likelihood.
- $L_{ij}(\boldsymbol{\theta})$ likelihood contribution for j^{th} observation of i^{th} individual.
- $L(\boldsymbol{\theta}; \boldsymbol{\lambda})$ joint log density (used to explain efs smoothing approach).
- $L(\boldsymbol{\lambda})$ joint log density when integrating out $\boldsymbol{\theta}$ (used to explain efs smoothing approach).

- m indexing for the (l, m) element of the matrices needed for the closed form expression of $\partial^2\mathbf{P}$.
- $\mathcal{M}$ number of simulations in the simulation study.
- $\mathbf{M}$ matrix appearing in the update of the smoothing parameter.
- N total number of statistical units.
- $\check{n}$ total number of observations in the dataset.
- n_i number of observations for the i^{th} statistical unit with $i = 1, \ldots, N$.
- n_{sim} number of simulations used to obtain confidence intervals.
- $\mathbf{O}$ quantity appearing in the smoothing parameter update and *edf* definition.
- $p^{(rr')}(t, t')$ transition probabilities referring to time interval (t, t').
- $p^{(\nu,rr')}(t, t')$ the ν^{th} simulated transition probability referring to time interval (t, t'), with $\nu = 1, \ldots, \mathcal{M}$.
- $\mathbf{P}(t, t')$ transition probability matrix referring to time interval (t, t').
- $\hat{\mathbf{P}}(t, t')$ estimated transition probability matrix referring to time interval (t, t').
- $q^{(rr')}(t)$ transition intensity at time t.
- $q^{(n_{sim},rr')}$ the n_{sim}^{th} simulated transition intensity (for confidence interval construction).
- $\mathbf{Q}(t)$ transition intensity matrix at time t.
- $\hat{\mathbf{Q}}(t)$ estimated transition intensity matrix at time t.
- $\mathbf{Q}_j(\boldsymbol{\theta})$ transition intensity matrix at the j^{th} observation of a generic individual.
- r starting state.
- r' arrival state.
- $\mathbb{R}$ real numbers set.
- $s_k^{(rr')}(\tilde{\mathbf{x}}_{k\iota})$ k^{th} smooth for the (r, r') transition intensity.
- $\mathcal{S}$ state space of process.
- $\mathbf{S}^{(rr')}_{\boldsymbol{\lambda}^{(rr')}}$ penalty term for the (r, r') transition intensity.
- $\mathbf{S}_{\boldsymbol{\lambda}}$ overall penalty term.

- t and t' generic time.
- t_{ij} with $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$ is the j^{th} observed time for the i^{th} statistical unit.
- t_j used as shorthand of t_{ij} for the generic statistical unit (i.e. when dropping i for simplicity).
- δt time interval in the definition of the closed form expression of $\mathbf{P}$
- T_{rs} time of the $r \to r'$ transition
- $T_{rs|u}$ time of the $r \to r'$ transition conditional on being in state r at time u
- u integration variable when integrating transition intensity.
- u_υ knot for the example in the (cubic regression) smooth of time.
- $\check{\mathbf{U}}_{ww'}$ one of the matrices of the closed form expression of $\dfrac{\partial^2}{\partial\theta_w \partial\theta_{w'}}\mathbf{P}$.
- $\dot{\mathbf{U}}_w$ one of the matrices of the closed form expression of $\dfrac{\partial}{\partial\theta_w}\mathbf{P}$.
- $\dot{\mathbf{U}}_{ww'}$ one of the matrices of the closed form expression of $\dfrac{\partial^2}{\partial\theta_w \partial\theta_{w'}}\mathbf{P}$.
- $\mathbf{V}_{\boldsymbol{\theta}}$ estimated negative inverse penalised Hessian.
- w and w' indexing for gradient vector and Hessian, with $w, w' = 1, \ldots, W$.
- W total number of parameters
- $\mathbf{x}_i$ covariate vector (without time).
- $\tilde{\mathbf{x}}_\iota$ overall covariate vector (with time).
- $\tilde{\mathbf{x}}_{k\iota}$ is the k^{th} sub-vector of the overall covariate vector $\mathbf{z}_i$.
- $\tilde{\mathbf{X}}_k^{(rr')}$ the design matrix corresponding to the k^{th} term in the (r, r') transition intensity.
- $\tilde{\mathbf{X}}^{(rr')}$ overall design matrix for the (r, r') transition intensity.
- y indexing of the eigenvalues.
- Y number of eigenvalues.
- z_{ij} with $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$ is the j^{th} state occupied by the i^{th} statistical unit.
- $Z(t)$ multi-state process.

Greek letters

- α confidence level.
- $\beta_0^{(rr')}$ intercept parameter for (r, r') transition intensity.
- $\boldsymbol{\beta}_k^{(rr')}$ parameter vector for the k^{th} term in the (r, r') transition intensity. Its length is $J_k^{(rr')}$.
- $\boldsymbol{\beta}^{(rr')}$ parameter vector for (r, r') transition intensity. Its length is $\sum_{k=1}^{K^{(rr')}} J_k^{(rr')}$.
- $\hat{\boldsymbol{\beta}}^{(rr')}$ estimated parameter vector of $\boldsymbol{\beta}^{(rr')}$.
- $\boldsymbol{\beta}^{(n_{sim},rr')}$ the n_{sim}^{th} simulated parameter vector for the (r, r') transition intensity.
- γ_y the y^{th} eigenvalue, with $y = 1, \ldots, Y$.
- $\boldsymbol{\Gamma}$ matrix of eigenvalues.
- δt time interval in the definition of the closed form expression of the transition probability matrix (and its derivatives).
- $\Delta^{[a]}$ radius of the trust region at the a^{th} iteration.
- $\boldsymbol{\epsilon}$ quantity appearing in the smoothing parameter update.
- ζ indexing for the series representing the exponential.
- $\eta_\iota^{(rr')}(t_\iota, \mathbf{x}_\iota; \boldsymbol{\beta}^{(rr')})$ additive predictor.
- $\boldsymbol{\eta}^{(rr')}$ overall additive predictor for the (r, r') transition intensity.
- $\boldsymbol{\theta}$ overall parameter vector.
- $\hat{\boldsymbol{\theta}}$ estimated overall parameter vector.
- $\boldsymbol{\theta}^{[a]}$ overall parameter vector at the a^{th} iteration of the estimation algorithm.
- ι indexing of the observations when defining the additive predictor. Here $i = 1, \ldots, \check{n}$.
- κ indexing for the summations appearing in the proof of the $\partial^2\mathbf{P}$ expression.
- $\lambda_k^{(rr')}$ smoothing parameter for the k^{th} term in the (r, r') transition intensity.
- $\boldsymbol{\lambda}^{(rr')}$ smoothing parameter vector in the (r, r') transition intensity. It's length is $K^{(rr')}$.
- $\boldsymbol{\lambda}$ overall smoothing parameter vector.
- $\boldsymbol{\mu}_\mathbf{M}$ and $\hat{\boldsymbol{\mu}}_\mathbf{M}$ quantity appearing in the smoothing parameter update.
- ν indexing for simulated probabilities to compute the bias in the simulation study.
- ρ indexing for the summations appearing in the proof of the $\partial^2\mathbf{P}$ expression.