Stable and practical semi-Markov modelling of intermittently-observed data

Christopher Jackson

1 Abstract

Multi-state models are commonly used for intermittent observations of a state over time, but these are generally based on the Markov assumption, that transition rates are independent of the time spent in current and previous states. In a semi-Markov model, the rates can depend on the time spent in the current state, though available methods for this are either restricted to specific state structures or lack general software. This paper develops the approach of using a "phase-type" distribution for the sojourn time in a state, which expresses a semi-Markov model as a hidden Markov model, allowing the likelihood to be calculated easily for any state structure. While this approach involves a proliferation of latent parameters, identifiability can be improved by restricting the phase-type family to one which approximates a simpler distribution such as the Gamma or Weibull. This paper proposes a moment-matching method to obtain this approximation, making general semi-Markov models for intermittent data accessible in software for the first time. The method is implemented in a new R package, msmbayes, which implements Bayesian or maximum likelihood estimation for multi-state models with general state structures and covariates. The software is tested using simulation-based calibration, and an application to cognitive function decline illustrates the use of the method in a typical modelling workflow.

2 Introduction

Many scientific analyses involve studying how a categorical variable changes over time. This can be done with a *multi-state model*, a continuous-time stochastic process on a finite set of states. In biomedical and other application areas, it is common that the state of an individual is only observed at a finite set of times, giving "intermittently-observed" or "panel" data. Transition times are interval-censored, and in some models it is even unknown how many transitions took place between observations. Such data are generally modelled with Markov models, using the maximum likelihood method introduced by Kalbfleisch and Lawless [15] and

its various extensions. The msm R package [13] is widely used for this. See, e.g. Van Den Hout [28] or Cook and Lawless [8] for broad reviews.

In these models, it is particularly challenging to relax the Markov assumption, that is, to allow an individual's transition rates to depend on the previous states they have visited and their transition times. This is inherently difficult with intermittent observations, since this history is only partially known. Here we focus on semi-Markov models, in which transition rates depend on the time since entry to the current state. This time is unknown, since the time of entry is unknown. A range of approaches for semi-Markov modelling of panel data have been suggested for general transition structures.

One general approach is based on integrating over the latent transition times [1, 30, 17], though the computational cost of doing this increases substantially with the number of unobserved transitions. A nonparametric estimator was also devised by Gu et al. [12], but only for transition structures without cycles. An alternative approach which avoids explicit integration, and allows any transition structure, is to iteratively simulate pathways consistent with the data, as part of a custom MCMC scheme [4] or a stochastic EM algorithm [2], though software for this is not available.

Another approach general to any structure is the "phase-type" model proposed by Titman and Sharples [27]. This does not rely on a custom computational algorithm, hence is more amenable to a generic software implementation. This replaces the exponentially-distributed sojourn in a state by a sequence of exponential sojourns in latent "phases". This expresses a semi-Markov model as a hidden Markov model, which allows likelihoods to be evaluated easily. However it introduces many extra parameters, which can lead to poor identifiability, particularly for intermittent observations. To address this problem, Titman [24] developed a method that uses the phase-type structure to approximate simpler time-to-event distributions, specifically the Weibull and Gamma, however the "variational optimization" used to implement this relied on an intensive numerical search and ad-hoc choices of spline functions, and no code was published.

This paper describes a new computational procedure that makes stable semi-Markov modelling of intermittent data, with general transition structures, accessible in general software for the first time. This uses phase-type approximations to standard distributions, as in Titman [24], but instead of using numerical optimization, the approximation is obtained by a fast analytic procedure. This involves finding the unique phase-type distribution of a particular family whose first three moments agree with those of the Gamma or Weibull. The model from that paper is also extended here to include covariates. The method is implemented in a new R package, msmbayes, which also allows general multi-state models for intermittently-observed data to fitted by either Bayesian estimation or maximum likelihood. An advantage of a Bayesian approach is that background information can be used to stabilise computation in the situation, common for this kind of modelling, where maximum likelihood fails due to weak identifiability.

Section 3 gives an overview of the current framework for multi-state modelling of intermittently-observed data. Section 4 describes phase-type semi-Markov models, their use to approximate simpler distributions, and the novel method we introduce to implement this approximation. The computational methods and software are introduced in Section 5. Section 6 presents a simulation-based calibration study to demonstrate that the software produces the correct posterior for a wide range of model classes. Section 7 demonstrates how the models might be used in a realistically-complex example based on a dataset measuring transitions between states of cognitive function and death.

3 Overview of multi-state models

In a multi-state model, an individual moves between a set of states $1, \ldots, R$ in continuous time according to transition intensities (or rates) $q_{rs}(t, \mathcal{F}_t)$, defined as

$$lim_{\delta t \to 0} \frac{P(S(t + \delta t) = s | S(t) = r, \mathcal{F}_t)}{\delta t}$$

where the state at time t since the beginning of the process is S(t). If an individual cannot move to s immediately on exiting r, then $q_{rs} = 0$. The transition intensity matrix Q is defined to have (r, s) entry q_{rs} for $r \neq s$, and rth diagonal entry $-\sum_{s \neq r} q_{rs}$.

 \mathcal{F}_t represents the *history* of the process up to time t, that is, the states previously occupied by the individual and the times of transition between them. This paper is only concerned with Markov and semi-Markov models. In a Markov model, the transition rates are independent of the history \mathcal{F}_t . In a time-homogeneous Markov model, the $sojourn\ distribution$ in a state r (the time spent there before moving to another state) is exponential with rate $\sum_{s\neq r} q_{rs}$.

The transition intensity can be related to covariates $\mathbf{x}(t)$, usually via a proportional intensities model with $q_{rs}(\mathbf{x}(t)) = q_{rs}^{(0)} \exp(\boldsymbol{\beta}^T \mathbf{x}(t))$. If the covariates depend on time t, the model is time-inhomogeneous. A model can be time-inhomogeneous but still Markov, if covariates depend on time t since the start of the process, but not on the times spent in particular states.

An important quantity for fitting and prediction is the transition probability matrix P(t) which has (r, s) entry P(S(u + t) = s | S(u) = r). If covariates are assumed to be piecewise-constant in time, P(t) can be calculated easily (though relaxations are possible, Titman [25]). Over any interval (u, u + t) where the covariates \mathbf{x} are constant, $q_{rs}(\mathbf{x})$ is constant, hence $P(t) = Exp(tQ(\mathbf{x}))$, where Exp(t) is the matrix exponential [see, e.g. 9].

The data consist of observations of the state $S_{i,j}$ for individuals i at times $t_{i,j}$. For a Markov model, the likelihood is [15] $L(\boldsymbol{\theta}) = \prod_{i,j} p_{S_{i,j},S_{i,j+1}}(t_{i,j+1} - t_{i,j})$, where the $\boldsymbol{\theta}$ include the set of q_{rs} , or the $q_{rs}^{(0)}$ and $\boldsymbol{\beta}$. Therefore the likelihood can be calculated easily if the covariates are assumed to be constant between observations. There are simple extensions of this likelihood

to cases where the entry time to some states is known exactly, or where the state is known only to be within a particular subset of states [19].

A particularly important extension is the (continuous time) hidden Markov model, where observations of an outcome are assumed to be generated independently conditionally on the state of a Markov multi-state model. These models have been used for data from medical diagnostic tests that are subject to error, so that the observed data at time $t_{i,j}$ is a state $O_{i,j}$ that may be different from the true state $S_{i,j}$ [14]. Hidden Markov models can also be used as a mechanism for implementing semi-Markov models, as we show in the following section. The joint likelihood function for transition intensities and outcome probabilities can be evaluated easily through a standard recursive procedure (Supplementary Appendix 1).

4 Phase-type semi-Markov models

A semi-Markov model is defined by assuming that the sojourn distribution for some state follows a distribution other than the exponential, so that the rate of transition out of the state depends on the length of time spent in the state. Hence the "history" \mathcal{F}_t consists of the time since the individual last moved to the state they are in at time t, so the intensities can be written as $q_{rs}(t)$ if t is redefined as the time since last entry to state r.

In Titman and Sharples [26], a Coxian phase-type distribution is used as the sojourn distribution. An (observable) state r with this sojourn distribution is replaced by a sequence of latent states $r_1, \ldots, r_{n(r)}$, known as phases, where a larger number of phases n(r) gives a more flexible distribution. An example is illustrated in Figure 1, where observable state r=2 is given a n(2)=3-phase distribution. If an individual transitions to an observable state with a phase-type distribution, they start in the first phase, r_1 . Their subsequent progression is then governed by a Markov model on the latent state space. From each phase r_i , instantaneous transitions are allowed either to the next phase in the sequence r_{i+1} (with intensity $q_{r_i,r_{i+1}}$), or to an "exit" from the phase sequence, to any one of the next potential observable states (say s_1, s_2, \ldots), with intensity $q_{r_i,r_{exit}}$. For example, in Figure 1, on exiting the phase sequence for observable state r=2, an individual can transition to either state $s_1=3$ or state $s_2=4$. If such an exit occurs, this is to state s_j with a constant probability p_{r,s_j} (as in Titman and Sharples [26], assumed to be independent of the time spent in state r). The transition intensity from r_i to s_j is then

$$q_{r_i,s_j} = q_{r_i,r_{exit}} p_{r,s_j} \tag{1}$$

A multi-state model with a phase-type sojourn distribution is an example of a hidden Markov model, with transition intensities defined on the latent state space, and known outcome probabilities: the probability of observed state r given latent state k is 1 if k is one of the phases of state r, and 0 otherwise. The likelihood can therefore be evaluated easily using the "forward algorithm", as detailed in Supplementary Appendix 1.

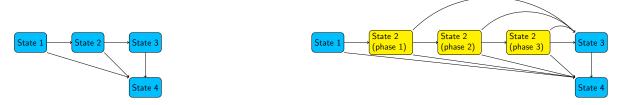


Figure 1: An example of a multi-state model where one state has a phase-type sojourn distribution with three phases. Left: observable state space. Right: latent state space.

4.1 Approximating standard distributions with phase-type models

By increasing the number of phases, any time-to-event distribution can be represented in theory (Xiangbin et al. [31]). However, with only intermittent observations of the state, there is a limit to how well the distribution of the time between state transitions can be characterised in practice. This time is interval-censored at best, and at worst, some sojourns may not be observed at all (e.g. periods spent infected, in a two state model of infection and recovery, if individuals are intermittently tested for the infection). Even two-phase sojourn distributions require three parameters (one "progression" rate and two "exit" rates). The applied interpretation of the phase transition rates is also unclear, which makes it difficult to define plausible priors in a Bayesian model. Therefore, Titman [24] developed a model based on familiar two-parameter time-to-event distributions, exploiting a phase-type approximation to enable the likelihood to be calculated. In this approach, the sojourn distribution for a "semi-Markov" state is defined as a two-parameter distribution with shape a, scale b, and probability distribution function

$$F(t|a,b) = F_p(t|\lambda = \mathbf{h}(a)/b)$$

 F_p represents a phase-type sojourn distribution with transition intensities collected in the vector λ , and defined as a function $\mathbf{h}()$ of the shape a, scaled by b. $\mathbf{h}()$ is determined so that for any (a,b), the resulting phase-type distribution closely matches, say, a Weibull or a Gamma distribution with shape a and scale b. The function need only be determined once, before observing any data, and stored in software. The model can then be fitted directly to any data, without the need to re-calculate $\mathbf{h}()$.

Titman [24] used a numerical approach to obtain $\mathbf{h}()$. For a particular shape a, this can be found by finding the rates $\lambda = \mathbf{h}(a)$ that minimise the Kullback-Leibler discrepancy $KL(\lambda)$ between the approximating phase-type model and the target Weibull or Gamma. To obtain the approximation for all possible a, $\mathbf{h}()$ was defined as a spline function of a for each component of λ , and the spline coefficients were determined by optimisation. In practice, this was a challenging optimisation problem, involving ad-hoc choices of spline knots to obtain the appropriate flexibility, and tuning of the numerical integration (over a wide range of times t) required to estimate $KL(\lambda)$.

Here, instead, we obtain $\mathbf{h}()$ using a new, fast analytic approach, which is easier to implement in software. Bobbio, Horváth, and Telek [5] showed that given values for the mean, variance and third moment, a phase-type sojourn distribution of the form in Figure 2 can be found with those moments, as long as the moments are within a set of bounds that depends on the number of phases n. As n increases, the bounds become wider, so a wider range of instances of the target distribution can be represented. Therefore, given a Gamma or Weibull distribution with a particular shape and scale, we can calculate the corresponding moments, and deduce the phase-type distribution with those moments. The phase-type parameters are moderately simple closed-form functions of the moments.

Supplementary Appendix 2 gives details of the approximation formula and required bounds, and illustrates the goodness-of-fit of the phase-type approximations to the Weibull and Gamma. For example, a Gamma distribution with shape parameters up to n can be approximated well by a phase-type model with n phases (indeed the distribution is theoretically identical when the shape equals n). Weibull distributions with shape parameters up to 2 are approximated well by a 5-phase model. For the lowest shape parameters (0.5 and less) the upper tail of the Weibull and Gamma becomes increasingly extreme compared to its centre, and the phase-type distribution with matching first three moments diverges from this. Note that for statistical modelling, we are not concerned with matching the Weibull and Gamma exactly, since we do not believe that real data are truly generated from any of these distributions. Instead the goal is to obtain usefully-flexible families of distributions, with key properties in common with the distributions that they are inspired by.

4.2 Using approximated phase-type sojourn distributions in multi-state models

To use the approximated sojourn distribution for a state r in a full multi-state model, further extensions are needed. Recall that if there is more than one state that an individual can move to on exit from state r with a phase-type distribution, parameters p_{r,s_j} describe the probability that the next state is s_j , assumed to not depend on the length of time spent in state r. We extend the models described in Titman [24] to allow covariates to influence both (a) the sojourn distribution, and (b) the next-state probability.

Covariates on the sojourn time Recall that the sojourn distribution is defined by a set of transition intensities on the latent space, $\lambda = \mathbf{h}(a)/b$. Replacing the scale parameter b by $b_0 \exp(-\beta \mathbf{x})$ (or reparameterising via a "rate" $\rho = 1/b$ and defining $\lambda = \rho_0 \exp(\beta \mathbf{x})$), defines an accelerated failure time (AFT) model for the effect of covariates \mathbf{x} on the sojourn time. In an AFT model, if P(T < t) is the CDF for scale b = 1, the CDF for scale $b \neq 1$ can be produced by scaling time, as P(T < t/b). This holds here, because the phase-type distribution has a CDF of the form $P(T < t) = 1 - \alpha \exp(St)I$, where S is a matrix comprising the transition intensities on the space of phases (see, e.g. Cumani [10]). Hence the "time scaling" can be achieved by scaling all of the transition intensities by the same scaling factor. Specifically, the progression and exit intensities (rates) between phases of a state r are related to \mathbf{x} through

the same log hazard ratio β_r :

$$q_{r_i,r_{i+1}} = q_{r_i,r_{i+1}}^{(0)} \exp(\boldsymbol{\beta}_r \mathbf{x})$$

$$q_{r_i,r_{exit}} = q_{r_i,r_{exit}}^{(0)} \exp(\boldsymbol{\beta}_r \mathbf{x})$$
(2)

Covariates on next-state probabilities Including covariates only on the sojourn time defines a model where a covariate has an equal effect on the rate of transition q_{r_i,s_j} to all exit states j (substituting Equation 2 into Equation 1). To relax this restriction, we also model the next-state probability $p_{r,s_j}(\mathbf{x})$ in terms of covariates, through a multinomial logistic regression,

$$\log(p_{r,s_i}(\mathbf{x})/p_{r,s_1}(\mathbf{x})) = \alpha + \gamma_{r_i}\mathbf{x}$$
(3)

so that γ_{rj} is the log odds ratio of transition to state s_j for a unit increase in \mathbf{x} , where the odds are respect to the first competing destination s_1 , and $1/(1 + \exp(\alpha)) = p_{r,s_1} = p_{rs}$, the "baseline" next-state probability (dropping the "1" subscript for clarity).

Then if covariates are also included on the sojourn time through Equation 2, the full model for the transition intensities on the latent Markov space (Equation 1) can be expressed as

$$q_{r_i,s_1}(\mathbf{x}) = q_{r_i,r_{exit}}^{(0)} \exp(\boldsymbol{\beta}_r \mathbf{x}) p_{rs}$$
(4)

$$q_{r_i,s_j}(\mathbf{x}) = q_{r_i,r_{exit}}^{(0)} \exp((\boldsymbol{\beta}_r + \boldsymbol{\gamma}_{rj})\mathbf{x})p_{rs}$$
 (5)

This gives another interpretation of γ_{rj} , as the relative hazard of transition to state s_j , relative to s_1 , from any of the latent phases i of state r. If $\beta_r > 0$ but all $\gamma_{rj} = 0$, covariates affect the sojourn time in state r, but do not affect which state happens next.

5 Full likelihood and computation

Consider a multi-state model with a general state-transition structure, and a state space S partitioned into (S_M, S_P) , where:

• States $r \in \mathcal{S}_M$ are "Markov", in the sense of having outward transition intensities q_{rs} that are constant, or depend only on covariates $\mathbf{x}(t)$ that are piecewise-constant with respect to the time t since the start of the process, with $q_{rs}(\mathbf{x}(t)) = q_{rs}^{(0)} \exp(\boldsymbol{\beta}^T \mathbf{x}(t))$. The parameters for these states are then $\boldsymbol{\theta}_M = \{(q_{rs}, \beta_{rs}) : r \in \mathcal{S}_M, s \in \mathcal{S}\}$, excluding any parameters for which the instantaneous r - s transition is disallowed or a covariate has no effect.

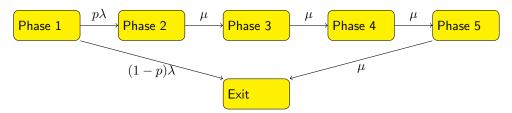


Figure 2: Coxian phase-type model used to approximate standard distributions by moment matching, illustrated for 5 phases. The phase-type distribution is the distribution of the time from entering phase 1 to entering the "Exit" state, in a continuous-time Markov model with the indicated transition intensities. The particular phase-type structure shown here is equivalent to the form in Bobbio et al. (2005), a mixture of two component distributions: (1) an Exponential(lambda) with probability 1-p, and (2) a sum of an Erlang(n-1, mu) and an Exponential(lambda) with probability p. The Erlang(n, mu) is a sum of n independent exponentials with rate mu, equivalent to a Gamma(n, mu). Hence the Gamma with integer shape n is exactly equivalent to a phase-type distribution of this class with mu = lambda and p = 1.

• States $r \in \mathcal{S}_P$ are "semi-Markov", with phase-type shape-scale sojourn distributions, with parameters $\boldsymbol{\theta}_P = \{a_r, b_r, p_{rs}, \boldsymbol{\beta}_r, \boldsymbol{\gamma}_{rs} : r \in \mathcal{S}_P\}$ comprising shape parameters a_r , baseline (i.e. for covariate values of zero) scale parameters b_r , baseline next-state probabilites p_{rs} (where permitted), and any effects of covariates on the scale $(\boldsymbol{\beta})$ or on next-state probabilities $(\boldsymbol{\gamma})$.

The full likelihood function comprises parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_M, \boldsymbol{\theta}_P\}$. The values of the parameters in $\boldsymbol{\theta}_P$ for a particular state r define transition intensities λ_{rs} between the latent phases defining state r. Therefore setting $\boldsymbol{\theta}$ (and any covariate values) defines the intensities of a Markov model on an expanded state space including the observable states \mathcal{S}_M and all latent phases. The likelihood can then be evaluated using the forward algorithm, described in Supplementary Appendix 1.

The likelihood might either be maximised or used for Bayesian inference. Both approaches are implemented in a new R package, msmbayes, available from https://chjackson.github.io/msmbayes/. This wraps around the rstan [21] interface to the Stan [22] software, which implements a quasi-Newton optimisation algorithm (L-BFGS), or the "no U-turn sampler" Hamiltonian Monte Carlo (HMC) algorithm for Bayesian inference. If the optimisation procedure is used with improper uniform priors, this is equivalent to maximum likelihood estimation, where the maximum likelihood is the posterior mode. After optimisation, to produce interval estimates, a sample can be drawn from the multivariate normal distribution defined by the

estimates and covariance matrix (obtained from the Hessian), giving a Laplace approximation of the posterior.

Each shape parameter a_r is restricted during estimation to its valid range for the moment-based phase-type approximation, which depends on the number of approximating phases n. In the Bayesian approach, normal priors are assumed for all parameters (after transformation to an unrestricted range if needed), or truncated normal distributions for the a_r . Either a Weibull or Gamma distribution may be used for the phase-type approximation in each state.

A model can be fitted with a single R command, rather than requiring the user to write Stan code. Any state transition structure can be used, with different covariate models used for different parameters via standard R linear modelling syntax. Any prior parameters can be supplied, and posteriors summarised and processed easily via the posterior (Bürkner et al. [7]) and tidybayes (Kay [18]) R packages. Vignettes giving worked examples are available on the package website.

6 Simulation-based calibration

The correctness of the computational procedures and software is assessed using simulation-based calibration [23]. The goals are to check that full MCMC (HMC) samples from the correct posterior, to assess the relative speed and accuracy of the Laplace approximation around the mode, and to compare the utility of the Bayesian approach with currently-available tools for maximum likelihood estimation of multi-state models.

A range of datasets are simulated from the prior predictive distribution for a given prior and model specification. The idea is that if the computation produces the correct posterior for the given prior and data, the posteriors produced by fitting the model to the simulated datasets a should span a range similar to the prior. Formally, the posterior $p(\boldsymbol{\theta}|\mathbf{y})$, integrated over data \mathbf{y} generated from the prior predictive distribution $p(\mathbf{y})$, is equivalent to the prior $p(\boldsymbol{\theta})$. To check this, for each $i=1,\ldots,N$, a value $\tilde{\boldsymbol{\theta}}_i$ is generated from the prior, hence a dataset $\tilde{\mathbf{y}}_i$. Then for each i, a sample is drawn from the posterior distribution of some estimand $h(\boldsymbol{\theta})$, and the rank r_i of $h(\tilde{\boldsymbol{\theta}}_i)$ among this sample is determined. The resulting ranks r_1,\ldots,r_N should then have a uniform distribution (Talts et al. [23]). If the test fails, this is evidence that either the algorithm to obtain the posterior is inaccurate, or there are bugs in its software implementation.

The model and prior tested here are motivated by a hypothetical infectious disease. The states include no infection (state 1) and infection (state 2). The time spent with the infection (assumed to be the same as the time testing positive), and the time until the next infection, may not be exponentially distributed. For example, the risk of being infected may increase since the time since the previous infection, due to waning immunity. Beyond this, we assume that the risk of being infected does not depend on the number of previous infections.

In each synthetic dataset i, there are 100 individuals, distributed in a 2/2/3/3 ratio between four age groups ("0-60", "60-70", "70-80", "80+"), and a 6/4 ratio (evenly within age groups) between men and women. Their history of infections and clearance of infections is generated from a continuous-time Markov or semi-Markov model. Each person is assumed to be tested for the infection once every month for 12 months, giving 1200 intermittent observations of the infection state at discrete times.

The priors are designed to be weakly informative, conveying that the mean length of an infection is expected to be around 14 days, but this may be as much as 30 days (note this is not the distribution of individual infection durations, but uncertainty about the mean). Similarly, the mean time to the next infection, since clearance of the last, is expected to be 6 months, but this may be as much as 18 months. The sampling models assessed are:

- (a) a Markov model, with age-sex group as a categorical covariate (with 8 levels) on the rate of infection and the rate of clearance. Priors for the baseline intensities are $\log(q_{12}) \sim N(-1.8, 0.6^2)$ and $\log(q_{21}) \sim N(0.8, 0.4^2)$, which represent the above judgements about times between events. N(0,1) priors are used for each log hazard ratio, that cover a wide range of practically-plausible hazard ratios (0.13 to 7).
- (b) a semi-Markov model implemented with a 5-phase approximation to a Weibull distribution. A $N(0,0.35^2)$ prior is used for the log shape parameter, truncated on the supported range of 0 to log(2.1) (implying a 95% credible interval of around 0.5 to 2 for the shape), and a normal prior for the scale parameter as in (a). Age-sex group is a covariate on the scale parameter, with the same priors used for the scaling factor 1/b as for the hazard ratio in the Markov model.
- (c) a model where the two-state structure is extended to three states with an absorbing "death" state, with death permitted from either uninfected or infected. This gives "competing risks" of recovery or death from infection. For the log odds of death, relative to transition to the other living state, a normal $(0, 2.3^2)$ prior is used, which gives a roughly uniform distribution for the probability that the next state is death. Age-sex group also modifies the next-state probability, with (log) standard normal priors for the relative rate parameter γ .

Each model is fitted by two methods: MCMC, or posterior mode optimisation followed by Laplace approximation. 1000 simulation replicates are used.

The computational stability of the Bayesian approach is also compared against the current standard maximum likelihood procedure for multi-state modelling, using the msm package. The Markov model (a) is fitted directly by maximum likelihood. To compare semi-Markov models, since the phase-type Weibull/Gamma approximation is unavailable in msm, the closest comparable model to (b) is fitted, that is a two-phase model where phase transition rates are estimated directly together with other parameters, rather than through a shape-scale approximation.

Results

Accuracy of different Bayesian computation methods

For the parameters of the full semi-Markov model (c), Figure 3 illustrates that the ranks are uniformly distributed, showing that the msmbayes package accurately computes the posterior distribution in each model via the MCMC algorithm from Stan. Similar results are shown in Supplementary Appendix 3 for the simpler models (a) (Figures 1-2) and (b) (Figures 3-4).

For model (c), the Laplace approximation represents the true posterior reasonably well (Figure 4), though there is bias for some parameters, in particular the log shape parameters and log odds of the next state. Taking the MCMC estimate as the true posterior, the median log shape is generally overestimated by Laplace approximation, with absolute bias having a median of 0.11 (95% CI -0.07 to 0.35) over simulated datasets. The uncertainty about this parameter is underestimated, with the posterior interquartile range having a median bias of -0.19 (95% CI 0.83 to 3.13). (Absolute rather than relative biases are presented here because the true parameter value is different in every simulation, being simulated from the prior, with a prior credible interval of -0.7 to 0.6.) The Laplace approximation is substantially faster, however, taking around 1% of the run time of MCMC. Note also that while it gives a biased median, the Laplace method is based on an exact computation of the posterior mode.

Similar biases are seen for the shape parameter when using Laplace approximation in the simpler semi-Markov model (b) (Supplementary Appendix 3, Figures 3 and 4). For the Markov model (a) the biases from using Laplace approximation are smaller (Supplementary Appendix 3, Figures 1 and 2). For the parameter whose estimates appear the most biased (Figure 2, $\beta_{male,2,1}$), the posterior median has an absolute bias with median -0.06 (95% CI -0.27 to 0.09), small in the context of the prior credible interval of -2 to 2, and the interquartile range has a bias with median -0.09 (-0.48 to 0.08).

Stability benefits of Bayesian estimation over maximum likelihood

The results from fitting Markov model (a) illustrate the benefit of Bayesian approaches in multi-state models that are weakly identifiable. For 76% of the simulated datasets, maximum likelihood estimation does not converge to a maximum, with the Hessian deemed to be non-invertible at the point where the optimisation algorithm terminated. This might be expected after summarising the data, for example in the first dataset there were only 10 infections from 8 individuals in the age 0-60, female category. Even so, Bayesian estimation produces a confidently-converged posterior, that is dominated by the weakly informative prior for the covariate effects. Figure 5 shows the posterior is not much narrower than the prior for two selected hazard ratios in the first simulated dataset, implying that the likelihood is near flat in this dimension, but the data are informative in other dimensions. Even if such a model cannot estimate these parameters from data, it still provides a useful characterisation of uncertainty that can motivate research to obtain specific further data.

Maximum likelihood estimation of the two-phase model (using \mathtt{msm}) failed to converge for 91% of simulated datasets, but Bayesian estimation of semi-Markov model (b) succeeds for all. To elucidate the problem with maximum likelihood, the two-phase model is fitted to the first simulated dataset by Bayesian estimation, with transition rates estimated directly using diffuse log normal($-2,2^2$) priors, rather than through a shape-scale approximation. Figure Figure 6 shows the prior and posterior distributions for the phase transition rates. For the phase 1-2 rate in state 1, the posterior is similar to the prior, suggesting that there is negligible information in the data for this parameter, so that maximum likelihood estimation would fail. The comparable Gamma shape-scale phase-type model has fewer parameters, that are better informed by the data, hence the posteriors are further away from the priors. The problems of eliciting prior information in phase-type models, and comparison between fitted models, are discussed in more detail in the context of an application in the following section.

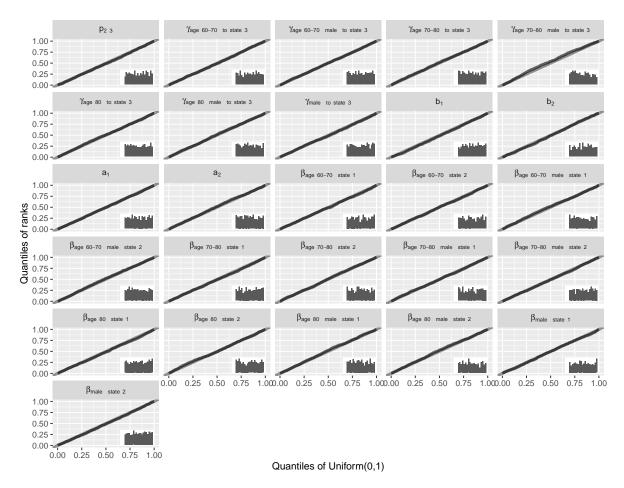


Figure 3: Simulation-based calibration of a phase-type shape/scale semi-Markov model with covariates, fitted by MCMC. The distribution of the rank statistic for two parameters (log transition intensities) over simulations is compared to a standard uniform.

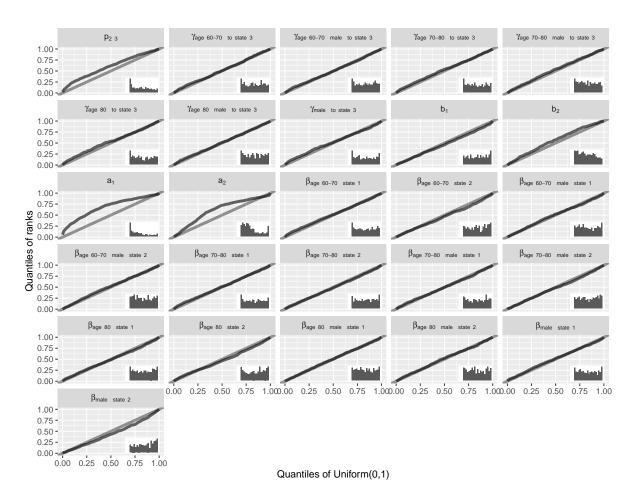


Figure 4: Simulation-based calibration of a phase-type shape/scale semi-Markov model with covariates, fitted by Laplace approximation around the posterior mode. The distribution of the rank statistic for two parameters (log transition intensities) over simulations is compared to a standard uniform.

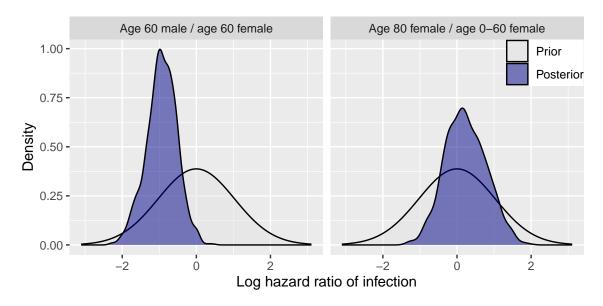


Figure 5: Prior and posterior distributions for two selected log hazard ratios in the Markov model with covariates fitted to the first simulated dataset

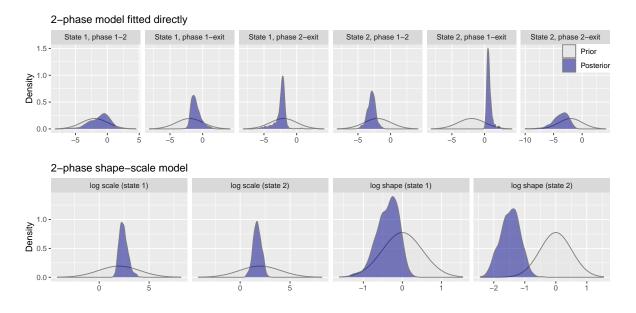


Figure 6: Prior and posterior distributions for parameters of a two-phase model with phase transition rates estimated directly (top row) or via a Gamma shape-scale approximation, for the first dataset in the simulation study

7 Application to cognitive function

To illustrate the practical capability and computational scalability of the method, we demonstrate how it might be used in a plausible application. Code to reproduce all analyses in this section using msmbayes, with a simulated dataset of the same structure, is provided at (https://chjackson.github.io/msmbayes/articles/cognitive.html).

In the English Longitudinal Study of Ageing (ELSA) [3] a cohort of people over 50 years old are surveyed around once every two years. The survey included various measures of cognitive function. The measure studied here is "delayed word recall", the number of words recalled correctly from a 10-word list, after a delay spent answering other survey questions. This outcome is categorised into one of four states, as in Van Den Hout [28], who developed a Markov multi-state model for the same dataset. The date of death is also recorded. A random sample of 1000 people is selected from the full dataset of 19802 people (giving a situation where Bayesian analysis might be beneficial due to the greater influence of the prior for a smaller sample size). The frequency of transitions between the states of cognitive function and death, observed over intervals one time point to the next, is illustrated in Table 1. Note there are only 67 deaths.

Table 1: Summary of transitions between states of cognitive function (number of words recalled in a test) and death observed in the ELSA data over intervals between successive observations

	To: 1	2	3	4	Death
From: 1 (7-10 words)	396	305	72	12	3
2 (5-6 words)	308	773	519	59	13
3 (2-4 words)	72	440	814	224	22
4 (0-1 words)	12	47	138	219	19

A multi-state model is developed to describe the dynamics of change over time in this cognitive measure, and how rates of death vary with the cognitive state. Transitions in continuous time are permitted between each adjacent cognitive state, and from each state to death (Figure 7). Covariates include year of age (50-90, median 60 at first visit), sex (54% female) and highest level of education (17% tertiary, 47% upper secondary, vs 36% less). A linear effect of year of age (since age 50) on each log transition rate is used (as in Van Den Hout [28]). In addition, the semi-Markov models will allow an relaxation of the assumption that, given year of age, and the current state, the rate of transition to the next state is constant.

The full model (with 50 parameters, from 10 transitions each with 4 covariates, excluding the semi-Markov extension) is not identifiable in practice. If maximum likelihood is attempted (using msm), while this appears to converge, it results in estimates and standard errors for several parameters that are unrealistically large if interpreted as degrees of belief, notably for the rates of death, for which the transition counts are low.

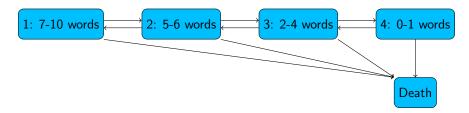


Figure 7: States and permitted continuous-time transitions in the multi-state model for the ELSA data

To stabilise estimation, we can introduce prior information as part of a Bayesian analysis. Strong priors for mortality rates, and effects of age and sex on these, are obtained from published national mortality data. Priors for other transition rates are assumed to be of a similar order of magnitude, but with wide credible intervals. Priors for other covariate effects are weakly informative, assuming (with 95% probability) that a single year of age, sex or education level will not increase (or decrease) any transition rate by more than a factor of 7, leading to a N(0, 1) prior for each $\log(\beta_{rs})$.

The Markov model is compared with a semi-Markov model, in which the sojourn distribution in each of the four living states is represented by a 5-phase approximation to the Weibull or Gamma. Priors for the parameters in the semi-Markov model that are compatible with the substantive beliefs used in the Markov model are determined by simulation.

Details of the priors and their derivation are given in Supplementary Appendix 4.

7.1 Computation

The models are implemented by posterior mode optimisation and Laplace approximation. Full MCMC, while feasible for the Markov models (with around a day of computation) was not feasible for the semi-Markov models. These computational demands come from the matrix exponential to obtain the transition probabilities in the likelihood function: $P(t) = Exp(tQ(\mathbf{x}))$ (Section 3). The matrix exponential must be evaluated for each distinct $Q(\mathbf{x})$ and time interval t. If covariates are continuous, then each individual in the dataset will typically have a different \mathbf{x} , as in this dataset, where there are 767 distinct time intervals and covariate values. The cost of evaluating the matrix exponential also strongly depends on the number of rows or columns of the matrix, which equals the number of (latent) states in the state space of the (hidden) Markov model. While there are 5 states in the Markov model, the semi-Markov model approximation here uses 5 latent phases to represent each of the four living observable states, giving 21 latent states overall.

7.2 Comparison of model fit

The overall goodness of fit of the models is compared informally via the posterior mode, i.e. the maximum penalised likelihood, by analogy with AIC which prefers the model that minimises the minus log likelihood plus the number of parameters. This neglects any influence of the priors, which are intended to be of a similar strength between the three models.

The maximised log posterior is -4906 for the Markov model, -4871.2 for the Gamma semi-Markov model and -4870.5 for the Weibull semi-Markov model. The semi-Markov models therefore give a modest improvement in fit, more than would be expected from the addition of only four extra parameters (the shape parameters for the four living states). Consistently with this improvement, the estimated shape parameters are significantly less than one, indicating a decreasing hazard of with time since state entry, for states 1, 3 and 4 (Table 2).

Note that formal cross-validation, or assessment of the *absolute* goodness of fit, would be challenging in general for this class of models, due to the unbalanced intermittent observation (Titman and Sharples [26]). Subsequent semi-Markov estimates presented are from the Weibull model.

7.3 Comparison of parameter estimates

Estimates of parameters (transformed to interpretable scales) are presented in Table 2, including the mean sojourn time and next-state probabilities for the baseline group (male, age 50, lowest level of education). Transitions forwards and backwards between the states of cognitive function are much more likely than death for the youngest ages, but the probability of death increases with age. Covariate effects on these are illustrated in Figure 8 for the Markov model and in Figure 9 for the semi-Markov model.

Figure 8 shows the effect of age on the hazard ratio of death for both men and women, and the lower risk of death for women. Age appears to increase the risk of progression, i.e. decline in cognitive function, and reduce the rate of recovery. The effect of sex on transitions between cognitive states is unclear, while higher levels of education are associated with lower risk of cognitive decline and higher chance of recovery. The posterior for the effects of age and sex on death is nearly identical to the prior from the national mortality data, indicating the ELSA data provide little information about these. All other covariate effects appear to be mainly informed by the data, since the posterior credible intervals shown in this figure are much tighter than the prior credible intervals of 1/7 to 7.

Figure 9 shows the effects of the same covariates under the semi-Markov model, which are parameterised differently. Covariates directly modify the sojourn time in a state through time acceleration factors, and separately modify the relative risk of which state the individual moves to next after this sojourn. While the message from the effects on sojourn times is harder to see, the effects on next-state probabilities are consistent with the Markov model. Age increases the chance that when a person leaves a state, this is to death rather than cognitive recovery

(third column). Age also increases the chance of progression in cognitive decline, relative to recovery. Higher levels of education reduce the risk of progression, relative to recovery.

7.4 Comparing model predictions between subpopulations

A clearer illustration of the effect of covariates can be given, under both models, by comparing the prediction of an *absolute* outcome between subgroups. As an example, we estimate the expected total length of time spent alive in the states 1 and 2 (interpreted as mild cognitive impairment or less), over a period of 10 years. This is defined as the integral, over this period, of the probability of transition to any of these states, for a person currently in state 1. Note that while, in reality, age is time-dependent, the age covariate is assumed to be fixed over this 10 year horizon (similarly to how time-dependent covariates are fixed when computing the transition probabilities over intervals between observations for the likelihood function, see Section 3).

Comparisons between subgroups are done by fixing one covariate to different values in turn, while standardising over the other covariates. For example, to compare different ages, we define a standard population whose distribution of sex and education roughly matches that of ELSA. The posterior distributions of the predicted time are computed for each member of the standard population while fixing age to 55, say, then these distributions are averaged (by concatenating posterior samples) to produce the standardised estimate for age 55. This is repeated for different age values. Comparisons of genders and education categories are done in a similar way.

These results quantify the greater time expected to be living free of cognitive impairment for younger people, women, and people with higher levels of education. The estimates agree between the Markov and semi-Markov models, suggesting the conclusions are robust to the assumptions made about the sojourn distribution.

8 Discussion

This paper has introduced a computational method that makes semi-Markov multi-state modelling of intermittently-observed data practicable in software. Analytic moment matching is used to easily construct families of phase-type sojourn distributions that approximate the Weibull and Gamma, which allows the likelihood to be evaluated as for a hidden Markov model. The phase-type approximation method is extended to handle covariates, and it is implemented in a general-purpose R package, msmbayes, that fits Bayesian multi-state models with any transition structure and covariates. The software is tested using simulation-based calibration, and the illustrative application shows the use of the methods in a typical applied workflow.

Table 2: Mean sojourn times in state of cognitive function, and next-state probabilities, under Markov and semi-Markov models. From each of the four living cognitive function states, the next state might be either progression (from states 1-3) to the next state of severity, recovery to the previous state (from states 2-4), or death (from any state). Posterior medians and 95% credible intervals for a man of age 50 years.

State	Mean	Probability of next state					
	sojourn						
	years						
		Progression	Recovery	Death			
Markov r	nodel						
1	4.97	0.99		0.01			
	(3.12, 7.78)	(0.97, 1.00)		(0.00, 0.03)			
2	1.76	0.80	0.20	0.00			
	(1.30, 2.39)	(0.69, 0.88)	(0.12, 0.31)	(0.00, 0.01)			
3	2.37	0.16	0.83	0.00			
	(1.72, 3.21)	(0.09, 0.28)	(0.72, 0.91)	(0.00, 0.01)			
4	1.14		1.00	0.00			
	(0.65, 1.96)		(0.99, 1.00)	(0.00, 0.01)			
Semi-Ma	rkov model				Shape		
1	2.97	1.00		0.00	0.64		
	(1.59, 5.48)	(0.95, 1.00)		(0.00, 0.05)	(0.57, 0.72)		
2	1.65	0.69	0.31	0.00	1.08		
	(1.22, 2.20)	(0.54, 0.81)	(0.19, 0.46)	(0.00, 0.02)	(0.95, 1.21)		
3	3.14	0.16	0.84	0.00	0.75		
	(2.12, 4.79)	(0.09, 0.26)	(0.74, 0.91)	(0.00, 0.01)	(0.63, 0.87)		
4	1.95		0.99	0.01	0.73		
	(1.06, 3.66)		(0.95, 1.00)	(0.00, 0.05)	(0.57, 0.89)		

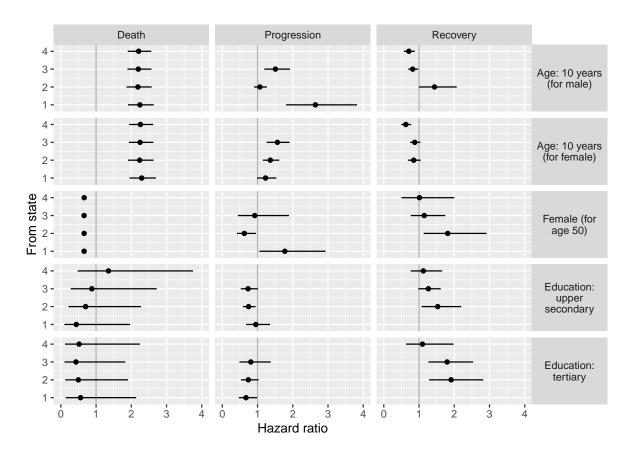


Figure 8: Covariate effects in the Markov model, as hazard ratios modifying the transition rates from each of cognitive states 1 to 4, to either progression (from states 1-3) to the next state of severity, recovery to the previous state (from states 2-4), or death (from any state). Posterior medians and 95% credible intervals.

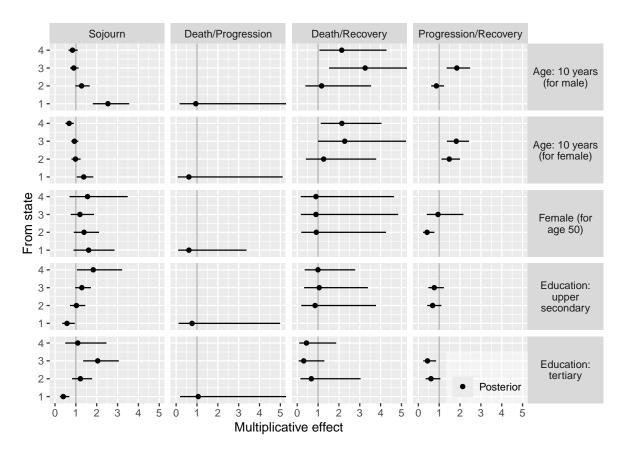


Figure 9: Covariate effects in the semi-Markov model. The first column "Sojourn" shows time acceleration factors modifying the scale of the sojourn distribution (> 1: faster time to event, or higher risk). The remaining columns show multiplicative effects on the relative risk of transition to a particular state, relative to some "baseline" state. For example, the first column "Death/Progression" shows the effect of the covariate on the relative risk of death compared to progression to state 2, for people in state 1. Posterior medians and 95% credible intervals.

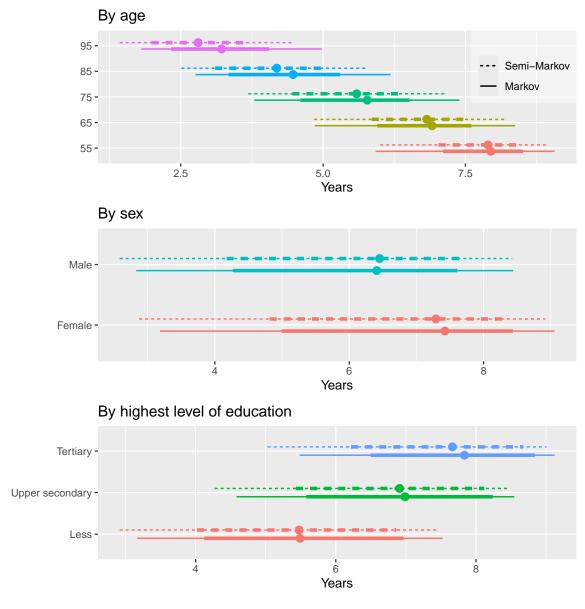


Figure 10: Predicted total time spent with no or mild cognitive impairment over 10 years (states 1 and 2, recall of 5-10 out of 10 words.) Standardised comparisions between ages, between sexes and between levels of education.

The approach has some limitations, however, in terms of flexibility and scalability. The two-parameter shape-scale sojourn distributions used here are more flexible than the exponential distributions used in these models, but are still limited. Though in practice, we would expect identifiability problems for more flexible distributions, particularly if observations are infrequent relative to the sojourn time. A stronger limitation, perhaps, is the computational scalability of the phase-type approximation. As the state space is expanded to include more latent states, the matrix exponentials (required to evaluate the likelihood) become rapidly more expensive. We have shown Laplace approximation can give a useful approximation to the posterior for larger datasets and models where MCMC is infeasible, and in practice, only two or three phases may be needed to get a usefully flexible family of sojourn distributions. By implementing the models using a probabilistic programming language (Stan), we have made it easier to extend them in the future to more complex situations (e.g. hierarchical models, informative observation, hidden Markov models with general outcomes), though problems of identifiability and computational scalability would remain.

Model checking and comparison remains a challenge for parametric multi-state models where transition times are unknown, in particular where observations are not on a regular grid [26]. It is increasingly common to assess statistical models using cross-validation, but this is difficult in situations such as this where the data structure is not exchangeable [29]. An alternative strategy is to compare estimates from a parametric model with nonparametric estimates, which is possible for restricted classes of multi-state models, e.g. acyclic transition structures [12] or with known transition times [11].

For complex Bayesian models such as these, where parameters do not all have simple interpretations, another difficulty is to obtain priors that reflect substantive information. A general strategy is to simulate from and check the prior predictive distribution, but the specific approaches we took here (Supplementary Appendix 4) were ad hoc. There is an active area of research about choosing prior parameters to minimise the discrepancy between the prior predictive distribution and information elicited about observable quantities [20, 6]. Tools for routine prior sensitivity analysis are increasingly accessible [16].

Multi-state modelling is one of many tools that can be used for longitudinal data. For example, in the ELSA application, the numerical measure of cognitive function might have been analysed directly using a parametric function of time within some generalisation of a linear model, rather than by modelling transitions between artificially discretised "states". Multi-state models are suitable for situations where the states are of direct interest, and this paper expands the range of multi-state models that can be used in practice.

9 Acknowledgements

This research was supported by the MRC, programme grant code MC UU 00040/4.

The English Longitudinal Study of Ageing was supported by the National Institute on Aging, of the National Institutes of Health, under Award Number R01AG017644, and NIHR Policy Research Programme (HEI) 198_1074_03. The content is solely the responsibility of the authors and does not necessarily represent the official views of these organisations.

Thanks to Andrew Titman for providing more details of the phase-type approximation in Titman [24].

10 Appendices

- 1. Likelihood for a hidden Markov model
- 2. Details of moment-matching procedure for approximating standard time-to-event distributions by phase-type distributions
- 3. Additional plots for the simulation-based calibration analyis
- 4. Specification of prior distributions for semi-Markov models

11 References

- [1] M. E. Aastveit, C. Cunen, and N. L. Hjort. "A new framework for semi-Markovian parametric multi-state models with interval censoring". In: *Statistical Methods in Medical Research* 32.6 (2023), pp. 1100–1123.
- [2] H. Aralis and R. Brookmeyer. "A stochastic estimation procedure for intermittently-observed semi-Markov multistate models with back transitions". In: *Statistical Methods in Medical Research* 28.3 (2019), pp. 770–787.
- [3] J. Banks et al. English Longitudinal Study of Ageing: Waves 0-10, 1998-2023. [data collection]. 40th Edition. UK Data Service. SN: 5050. 2024. URL: http://doi.org/10.5255/UKDA-SN-5050-27.
- [4] R. Barone and A. Tancredi. "Bayesian inference for discretely observed continuous time multi-state models". In: *Statistics in Medicine* 41.19 (2022), pp. 3789–3803.
- [5] A. Bobbio, A. Horváth, and M. Telek. "Matching three moments with minimal acyclic phase type distributions". In: *Stochastic Models* 21.2-3 (2005), pp. 303–326.
- [6] F. Bockting, S. T. Radev, and P.-C. Bürkner. "Simulation-based prior knowledge elicitation for parametric Bayesian models". In: *Scientific Reports* 14.1 (2024), p. 17330.
- [7] P.-C. Bürkner et al. posterior: Tools for Working with Posterior Distributions. R package version 1.6.1. 2025. URL: https://mc-stan.org/posterior/.
- [8] R. J. Cook and J. F. Lawless. Multistate models for the analysis of life history data. Chapman and Hall/CRC, 2018.

- [9] D. R. Cox and H. D. Miller. The Theory of Stochastic Processes. Routledge, 1977.
- [10] A. Cumani. "On the canonical representation of homogeneous Markov processes modelling failure-time distributions". In: *Microelectronics Reliability* 22.3 (1982), pp. 583–602.
- [11] D. Gomon and H. Putter. "Nonparametric Estimation of Transition Intensities in Interval-Censored Markov Multistate Models Without Loops". In: *Statistics in Medicine* 44.18-19 (2025), e70225.
- [12] Y. Gu et al. "Maximum likelihood estimation for semiparametric regression models with interval-censored multistate data". In: *Biometrika* 111.3 (2024), pp. 971–988.
- [13] C. H. Jackson. "Multi-state models for panel data: the msm package for R". In: *Journal of Statistical Software* 38 (2011), pp. 1–28.
- [14] C. H. Jackson et al. "Multistate Markov models for disease progression with classification error". In: *Journal of the Royal Statistical Society Series D: The Statistician* 52.2 (2003), pp. 193–209.
- [15] J. Kalbfleisch and J. F. Lawless. "The analysis of panel data under a Markov assumption". In: *Journal of the American Statistical Association* 80.392 (1985), pp. 863–871.
- [16] N. Kallioinen et al. "Detecting and diagnosing prior and likelihood sensitivity with power-scaling". In: *Statistics and Computing* 34.1 (2024), p. 57.
- [17] V. Kapetanakis, F. E. Matthews, and A. Van Den Hout. "A semi-Markov model for stroke with piecewise-constant hazards in the presence of left, right and interval censoring". In: Statistics in Medicine 32.4 (2013), pp. 697–713.
- [18] M. Kay. tidybayes: Tidy Data and Geoms for Bayesian Models. R package version 3.0.7. 2024. DOI: 10.5281/zenodo.1308151. URL: http://mjskay.github.io/tidybayes/.
- [19] R. Kay. "A Markov model for analysing cancer markers and disease states in survival studies". In: *Biometrics* (1986), pp. 855–865.
- [20] A. A. Manderson and R. J. Goudie. "Translating predictive distributions into informative priors". In: arXiv preprint arXiv:2303.08528 (2023).
- [21] Stan Development Team. RStan: the R interface to Stan. R package version 2.32.6. 2024. URL: https://mc-stan.org/.
- [22] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual. Version 2.36. 2024. URL: https://mc-stan.org.
- [23] S. Talts et al. "Validating Bayesian inference algorithms with simulation-based calibration". In: arXiv preprint arXiv:1804.06788 (2018).
- [24] A. C. Titman. "Estimating parametric semi-Markov models from panel data using phase-type approximations". In: *Statistics and Computing* 24 (2014), pp. 155–164.
- [25] A. C. Titman. "Flexible nonhomogeneous Markov models for panel observed data". In: *Biometrics* 67.3 (2011), pp. 780–787.

- [26] A. C. Titman and L. D. Sharples. "Model diagnostics for multi-state models". In: *Statistical Methods in Medical Research* 19.6 (2010), pp. 621–651.
- [27] A. C. Titman and L. D. Sharples. "Semi-Markov models with phase-type sojourn distributions". In: *Biometrics* 66.3 (2010), pp. 742–752.
- [28] A. Van Den Hout. Multi-state survival models for interval-censored data. Chapman and Hall/CRC, 2016.
- [29] A. Vehtari. Cross-validation FAQ. 2024. URL: https://mc-stan.org/loo/articles/online-only/faq.html.
- [30] S. Wei and R. J. Kryscio. "Semi-Markov models for interval censored transient cognitive states with back transitions and a competing risk". In: *Statistical Methods in Medical Research* 25.6 (2016), pp. 2909–2924.
- [31] H. Xiangbin et al. "Phase-Type Distributions for Sieve Estimation". In: *Journal of the American Statistical Association* (2025), pp. 1–12.

Appendix 1: likelihood for a continuous-time hidden Markov model

Christopher Jackson

This is Supplementary Appendix 1 of the paper "Stable and practical semi-Markov modelling of intermittently-observed data" (Jackson).

Likelihood for a continuous-time hidden Markov model

Suppose a set of individuals are generated independently from a continuous-time hidden Markov model. Consider one individual whose state is observed at times $t_j : j = 1, ..., J$. The *observed* state O_j at each time may be different from the *true* or *hidden* state S_j . We assume:

- The true states follow a continuous-time Markov process with transition intensities q_{rs} between states r and s on some state space S.
- The observed states are conditionally independent given the true states, with distribution $P(O_j|S_j=s,\mathbf{e}_s)$, where \mathbf{e}_s is a vector of parameters e.g. misclassification probabilities.
- The distribution of states at the first observation time $P(S_1 = s)$ is known.

An individual's contribution to the joint likelihood function for parameters $\theta = \{q_{rs}, e_s : r, s \in \mathcal{S}\}$ can be evaluated through the following procedure. This is sometimes termed the "forward algorithm", a common procedure for discrete-time hidden Markov models. See, e.g. Bureau, Shiboski, and Hughes [1] for its use for a continuous-time hidden Markov model.

We define:

• $\pi_{j,s}$, the probability of the data observed up to and including the jth time, and the true state being s at the jth time (marginally over all true states at time points j-1 and earlier, if j>1),

• $\mathbf{P}(t|\mathbf{q})$, the matrix of continuous-time transition probabilities over a time interval of any length t, such that the r,s entry is the probability of state s at the end of the interval, given state r at the start of the interval. When this is used, the interval is defined by a pair of successive observations, and \mathbf{q} is assumed constant over all such intervals.

The forward algorithm is then:

- 1. At the first observation time, we know $\pi_{1,s} = P(O_1|S_1 = s, \mathbf{e}_s)P(S_1 = s)$.
- 2. We can then update the marginal probabilities at the second time in terms of those at the first time, marginalising over the hidden state r at the first time:

$$\pi_{2,s} = \sum_{r} \pi_{1,r} \mathbf{P}(t_2 - t_1 | \mathbf{q})_{rs} P(O_2 | S_2 = s, \mathbf{e}_s)$$

- 3. Iterating through subsequent times j, we can then define each $\pi_{j,s}$ in the same way in terms of $\pi_{j-1,s}$.
- 4. At the final time J, we obtain the likelihood as $L(\theta) = \sum_s \pi_{J,s}$, the probability of data at all times, marginalised over all possible sequences of hidden states.

Note the same likelihood can also be expressed in a explicit form as a product of matrices [3, 2].

References

- [1] A. Bureau, S. Shiboski, and J. P. Hughes. "Applications of continuous time hidden Markov models to the study of misclassified disease outcomes". In: *Statistics in Medicine* 22.3 (2003), pp. 441–462.
- [2] C. H. Jackson et al. "Multistate Markov models for disease progression with classification error". In: *Journal of the Royal Statistical Society Series D: The Statistician* 52.2 (2003), pp. 193–209.
- [3] G. A. Satten and I. M. Longini Jr. "Markov chains with measurement error: Estimating the "true" course of a marker of the progression of human immunodeficiency virus disease". In: Journal of the Royal Statistical Society: Series C (Applied Statistics) 45.3 (1996), pp. 275–295.

Appendix 2: details of moment-matching procedure for approximating other distributions by phase-type distributions

Christopher Jackson

This is Supplementary Appendix 2 of the paper "Stable and practical semi-Markov modelling of intermittently-observed data" (Jackson).

To construct useful families of time-to-event distributions for semi-Markov multistate modelling of intermittently-observed data, the Weibull or Gamma are approximated via phase-type distributions. These approximations are constructed via algebraically matching the first three moments of the distributions, using formulae derived by Bobbio, Horváth, and Telek [1].

This appendix describes the moment matching formulae, and illustrates that the resulting distributions give good agreement with the Weibull or Gamma.

1 Moment-matching formulae

The moment-based phase-type approximations (Bobbio, Horváth, and Telek [1]) are defined in terms of the first three moments of the distribution, expressed through the mean and two further "normalised moments" (or "standardised moments"). The kth normalised (central) moment of a random variable X is defined for $k \geq 2$ as

$$n_k = \frac{E((X - \mu)^k)}{(E((X - \mu)^2)^{k/2})}$$

where μ is the mean.

For the Gamma distribution with shape a and scale b (or rate 1/b), the mean is $\mu = ab$ and the normalised moments are $n_2 = (a+1)/a$, $n_3 = (a+2)/a$.

For the Weibull distribution with shape a and scale b, $\mu = b\Gamma(1+1/a)$, $n_2 = \Gamma(1+2/a)/\Gamma(1+1/a)^2$ and $n_3 = \Gamma(1+3/a)/(\Gamma(1+1/a)\Gamma(1+2/a))$.

1.1 Restrictions on distributions that can be represented

Theorem 3.1 of Bobbio, Horváth, and Telek [1] says that if n_2 and n_3 satisfy a particular set of bounds (a moderately complex function depending on n, given in the theorem) then they can be the second and third normalised moments of a phase-type distribution of the form in Figure 1 with n phases.

For the Gamma distribution, a necessary condition for these bounds to be satisfied is a < n. This can also be shown to numerically to be sufficient (demonstrated at least for $n \le 10$ and shape values on a fine grid of 0.01 between 0 and n). For the Weibull distribution, an upper bound on the shape parameter for each n can be found by numerical root-finding (for 2, 5, 10 phases, this is 1.2, 2.0 and 3.1, respectively), and again the lower bound can be shown numerically to be zero. Code to obtain these bounds, and these demonstrations, are given in Section 3 of this document.

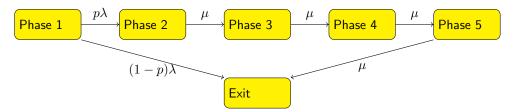


Figure 1: Coxian phase-type model used to approximate standard distributions by moment matching, illustrated for 5 phases. The phase-type distribution is the distribution of the time from entering phase 1 to entering the "Exit" state, in a continuous-time Markov model with the indicated transition intensities. The particular phase-type structure shown here is equivalent to the form in Bobbio et al. (2005), a mixture of two component distributions: (1) an Exponential(λ) with probability 1-p, and (2) a sum of an Erlang($n-1,\mu$) and an Exponential(λ) with probability p. The Erlang(n,μ) is a sum of n independent exponentials with rate μ , equivalent to a Gamma(n,μ). Hence the Gamma with an integer shape parameter is exactly equivalent to a phase-type distribution of this class with $\mu = \lambda$ and p = 1.

1.2 Definition of phase-type distribution with matching moments

Bobbio, Horváth, and Telek [1] then gave expressions for the parameters of the unique phase-type distribution of the "Erlang-Exp" form in Figure 1 which matches the first three moments when the required bounds are satisfied.

This is equivalent to a mixture of two component distributions: (1) an Exponential(λ) with probability 1-p, and (2) a sum of an $\operatorname{Erlang}(n-1,\mu)$ and an $\operatorname{Exponential}(\lambda)$ with probability p. The $\operatorname{Erlang}(n,\mu)$ is a sum of n independent exponentials with rate μ , equivalent to a $\operatorname{Gamma}(n,\mu)$. This is shown for 5 phases in Figure 1.

Expressions for p, λ and μ are given in Bobbio, Horváth, and Telek [1] as the solution to a pair of simultaneous polynomial equations. The solutions given below were derived for this paper using the SymPy symbolic algebra software (Meurer et al. [2]). They are essentially a robustified version of the solution given in Bobbio, Horváth, and Telek [1], including a check for which of the alternative solutions give a valid probability $0 \le p \le 1$ and $a \ge 0$, and handling special cases.

Defining

$$A = nn_2(12nn_2^2 + nn_2n_3^2 - 14nn_2n_3 - 15nn_2 + 16nn_3 + 12n_2^2 - 8n_2n_3 - 24n_2 + 16n_3)$$

$$B = \sqrt{A}$$

$$C = nn_2 - n + n_2 - 4$$

$$D = -nn_3 + n + 4$$

$$E = nn_2 - 2n + n_2 - 2$$

$$F = nn_2 - nn_3 + n_2$$

$$G = 8nn_2^2(n+1)(n_2 - 2)$$

the solution is either

$$a = -(n-1)(n_2D - B)(2n_2(n_2D - B)C - 8n_2FE + (n_2D - B)^2)/(GF^2)$$

$$b = (-n_2n_3 + n_2 + 4n_2 - B)/(2n_2F)$$

or

$$a = -(n-1)(n_2D + B)(2n_2(n_2D + B)C - 8n_2FE + (n_2D + B)^2)/(GF^2)$$

$$b = (n_2D + B)/(2n_2F)$$

or if $n(n_3 - n_2) = n_2$, the simpler form:

$$a = (2n_2 - n_3)(3n_2^2 - 4n_3)(-3n_2^2 + 2n_3(n_2 - 2) + 4n_3)/(n_2^2n_3(n_2 - 2)(n_2n_3 + 3n_2 - 4n_3))$$

$$b = (3n_2^2 - 4n_3)/(n_2(n_2n_3 + 3n_2 - 4n_3))$$

Then p = (b-1)/a, $\lambda = (pa+1)/m_1$, and $\mu = \lambda(n-1)/a$. The Exponential(λ) distribution has p = 0.

2 Illustration of phase-type approximation to Gamma and Weibull

The 5-phase approximation to the Gamma distribution is illustrated here for shape parameters between 0.2 and 3, showing the median, interquartile range and 95% quantiles. It is

verified that each 5-phase distribution has the same mean, variance and third moment as the corresponding Gamma, though there are slight differences in the precise quantiles. Note the extremely skewed shape of the distribution for the lowest shape parameters, where the phase-type approximation captures the pattern, though to a less precise degree.

This code uses internal functions in the msmbayes package.

```
library(msmbayes)
library(dplyr)
library(ggplot2)
library(ggdist)
shapes \leftarrow seq(0.2, 3, by=0.1)
qs \leftarrow c(0.025, 0.25, 0.5, 0.75, 0.975)
qmatg <- qmatp <- matrix(nrow=length(shapes), ncol=length(qs))</pre>
for (i in seq_along(shapes)){
  r <- shapescale to rates(shapes[i], family="gamma", list=TRUE, nphase = 5)
  qmatg[i,] <- qgamma(qs, shapes[i])</pre>
  qmatp[i,] <- qnphase(qs, r$p, r$a)</pre>
  if (!isTRUE(all.equal(shapes[i],1))){
    mo <- unlist(msmbayes:::gamma_nmo(shapes[i]))</pre>
    stopifnot(msmbayes:::in_moment_bounds(mo[["n2"]], mo[["n3"]], n=5))
    mop <- c(mean_nphase(r$p, r$a), var_nphase(r$p, r$a),</pre>
             skewness_nphase(r$p,r$a))
    stopifnot(all.equal(as.numeric(mo[c("mean","var","skew")]),mop))
  }
}
qmatg <- as.data.frame(qmatg) |> setNames(qs) |> mutate(model="Gamma")
qmatp <- as.data.frame(qmatp) |> setNames(qs) |> mutate(model="5-phase")
qmat <- rbind(qmatg, qmatp) |>
  mutate(shape=rep(shapes,2),
         shapey = ifelse(model=="Gamma", shape+0.02, shape-0.02))
ggplot(qmat, aes(x=`0.5`, y=shapey, col=model)) +
  geom pointinterval(aes(xmin=`0.025`,xmax=`0.975`), linewidth=2, size=5) +
  geom_pointinterval(aes(xmin=`0.25`,xmax=`0.75`), linewidth=10) +
  scale_x_continuous(breaks=seq(0, 15, by=2)) +
  scale_y_continuous(breaks=unique(qmat$shape)) +
  xlab("") + ylab("Shape") +
  guides(col=guide_legend(position="inside", title="")) +
  theme minimal() +
  scale_color_manual(breaks=c("5-phase", "Gamma"),
                      values=c("lightgreen","black")) +
  theme(panel.grid.major.y = element_blank(),
```

```
panel.grid.minor.y = element_blank(),
panel.grid.minor = element_blank(),
legend.position.inside=c(1, 0),
legend.justification.inside=c(1,0))
```

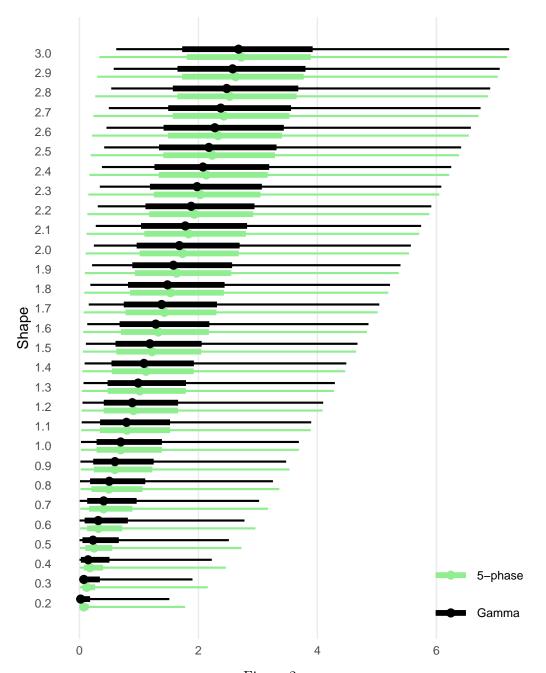
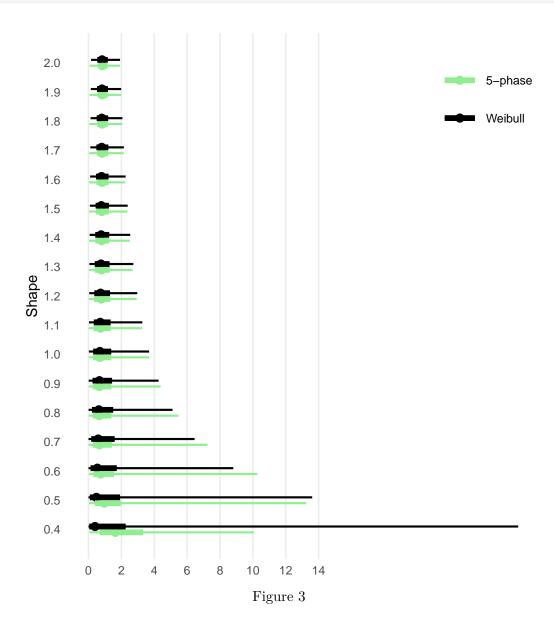


Figure 2

The 5-phase approximation to the Weibull distribution is illustrated here for shape parameters up to the maximum allowable 2, and a scale parameter of 1. The phase-type distributions with matching first three moments also give a nearly exact match to the quantiles for shape parameters > 1 (in which the hazard is increasing over time and skewness is low). For shape parameters < 1, where the hazard is decreasing, the phase-type approximations capture the general characteristics of the distribution, though the upper tail quantile is less well-matched. There is a limit to how well the phase-type distributions can capture the extremely long tails of the Weibull for shape parameters ≤ 0.5 , though we expect this to be an uncommon distribution in real data. Similar behaviour for these parameter ranges was also found by Titman (2014, supplementary Figure 1) who used a numerical method of achieving a phase-type approximation.

```
shapes \leftarrow seq(0.4, 2.0, by=0.1)
qs <-c(0.025, 0.25, 0.5, 0.75, 0.975)
qmatw <- qmatp <- matrix(nrow=length(shapes), ncol=length(qs))</pre>
for (i in seq_along(shapes)){
 r <- shapescale to rates(shapes[i], family="weibull", list=TRUE, nphase = 5)
  qmatw[i,] <- qweibull(qs, shapes[i])</pre>
 qmatp[i,] <- qnphase(qs, r$p, r$a)</pre>
  if (!isTRUE(all.equal(shapes[i],1))){
    mo <- unlist(msmbayes:::weibull_nmo(shapes[i]))</pre>
    stopifnot(msmbayes:::in moment bounds(mo[["n2"]], mo[["n3"]], n=5))
    mop <- c(mean_nphase(r$p, r$a), var_nphase(r$p, r$a),</pre>
             skewness nphase(r$p,r$a))
    stopifnot(all.equal(as.numeric(mo[c("mean","var","skew")]),mop))
 }
}
qmatw <- as.data.frame(qmatw) |> setNames(qs) |> mutate(model="Weibull")
qmatp <- as.data.frame(qmatp) |> setNames(qs) |> mutate(model="5-phase")
qmat <- rbind(qmatw, qmatp) |>
 mutate(shape=rep(shapes,2),
         shapey = ifelse(model=="Weibull", shape+0.01, shape-0.01))
ggplot(qmat, aes(x=`0.5`, y=shapey, col=model)) +
  geom_pointinterval(aes(xmin=`0.025`,xmax=`0.975`), linewidth=2, size=5) +
 geom_pointinterval(aes(xmin=`0.25`,xmax=`0.75`), linewidth=10) +
  scale_x_continuous(breaks=seq(0, 15, by=2)) +
 scale_y_continuous(breaks=unique(qmat$shape)) +
 xlab("") + ylab("Shape") +
  guides(col=guide legend(position="inside", title="")) +
  theme minimal() +
  scale color manual(breaks=c("5-phase", "Weibull"),
                     values=c("lightgreen","black")) +
  theme(panel.grid.major.y = element_blank(),
```

```
panel.grid.minor = element_blank(),
legend.position.inside=c(1, 1),
legend.justification.inside=c(1, 1))
```



Note that the class of phase-type distribution used for the moment approximation (Figure 1) is defined as a mixture of two component distributions: (1) an Exponential(λ) with probability 1-p, and (2) a sum of an Erlang($n-1,\mu$) and an Exponential(λ) with probability p. Therefore the Gamma with integer shape parameter n is exactly equivalent to an n-phase phase-type

distribution of this class with $\mu = \lambda$ and p = 1, since the Erlang (n, μ) is a sum of n independent exponentials with rate μ , equivalent to a Gamma (n, μ) .

Note also that the approximation for a given shape would not improve with more phases n, since the first three moments match exactly for any number of phases, and the fourth or higher moments are not considered. The purpose of adding more phases is to define families that represent a range of shape parameters with a higher upper bound (e.g. here shapes > 2 would be allowed for the Weibull with more than 5 phases).

3 Verification of constraints on shape parameter required for a matching phase-type distribution

For a phase-type distribution with given first three moments to exist, a necessary and sufficent condition is that the normalised moments lie within a particular set of bounds. The width of these bounds increases with the number of phases n, and they have a moderately complicated analytic formula, given in Bobbio, Horváth, and Telek [1].

Below we show numerically that for the Gamma and Weibull distribution, this condition is equivalent to the shape parameter of the distribution lying within a particular range. Hence when the shape is in this range, the distribution can be approximated well with a phase-type distribution. This holds for any scale or rate parameter, since the condition involves only the normalised second and third moments.

3.1 Gamma distribution

For the Gamma with shape a, the normalised second and third moments are $n_2 = (a+1)/a$, $n_3 = (a+2)/a$.

A necessary condition for a matching phase-type distribution (Bobbio, Horváth, and Telek [1]) is $n_2 \geq (n+1)/n$, where n is the number of phases. This is simply equivalent to a < n. Below it is shown numerically that the moment bounds are satisfied for any shape parameter a on a fine grid of values between 0 and n, for each of n between 2 and 10. The phase-type distribution for each shape is then obtained, and it is verified that the first three moments of this distribution match those of the Gamma.

```
library(msmbayes)
for (nphase in 2:10){
    shapes <- seq(0.01, nphase-0.01, by=0.01)
    shapes <- setdiff(shapes, 1)
    inb <- msmbayes::gamma_shape_in_bounds(shapes, nphase)
    stopifnot(all(inb))
    for (i in seq_along(shapes)){</pre>
```

3.2 Weibull distribution

For the Weibull distribution with shape a, the normalised moments involve Gamma functions (see the function weibull_nmo below). Unlike for the Gamma, the maximum shape a, for which the Weibull can be moment-matched to to a phase-type distribution, is not available analytically. Instead it can be found by a numerical search for a point at which the bounds are satisfied on one side, but not on the other. This is done below for the number of phases ranging from 2 to n.

```
weibull_nm <- function(shape){</pre>
  n2 \leftarrow gamma(1+2/shape) / gamma(1+1/shape)^2
  n3 \leftarrow gamma(1+3/shape) / (gamma(1+1/shape)*gamma(1+2/shape))
  list(n2=n2, n3=n3)
}
fn <- function(shape, n){</pre>
  nmo <- weibull_nm(shape)</pre>
  mb <- msmbayes::n3_moment_bounds(nmo$n2, n=n)
  nmo$n3 - mb[["lower"]]
weibull_ubounds <- c(NA, # 1 phase
                      uniroot(fn, interval=c(1.001, 1.2), n=2)$root,
                      uniroot(fn, interval=c(1, 1.7), n=3)$root,
                      uniroot(fn, interval=c(1, 2), n=4)$root,
                      uniroot(fn, interval=c(1, 2.3), n=5)$root,
                      uniroot(fn, interval=c(1, 2.4), n=6)$root,
                      uniroot(fn, interval=c(1, 2.6), n=7)$root,
                      uniroot(fn, interval=c(1, 2.8), n=8)$root,
                      uniroot(fn, interval=c(1, 3), n=9)$root,
                      uniroot(fn, interval=c(1, 3.5), n=10)$root
```

It is then verified numerically that for shapes a on a fine grid between 0 and this upper bound, the bounds are satisfied. Hence a moment-matching phase-type distribution exists for the

Weibull with this range of shape parameters. As before, this distribution is obtained and we check that the moments match. Note this procedure does not work for shapes around 0.05 or less, for which the mean and variance of the Weibull are so large that they numerically overflow.

References

- [1] A. Bobbio, A. Horváth, and M. Telek. "Matching three moments with minimal acyclic phase type distributions". In: *Stochastic Models* 21.2-3 (2005), pp. 303–326.
- [2] A. Meurer et al. "SymPy: symbolic computing in Python". In: *PeerJ Computer Science* 3 (Jan. 2017), e103. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.103. URL: https://doi.org/10.7717/peerj-cs.103.

Appendix 3: additional plots for simulation-based calibration

Christopher Jackson

This is Supplementary Appendix 3 of the paper "Stable and practical semi-Markov modelling of intermittently-observed data" (Jackson).

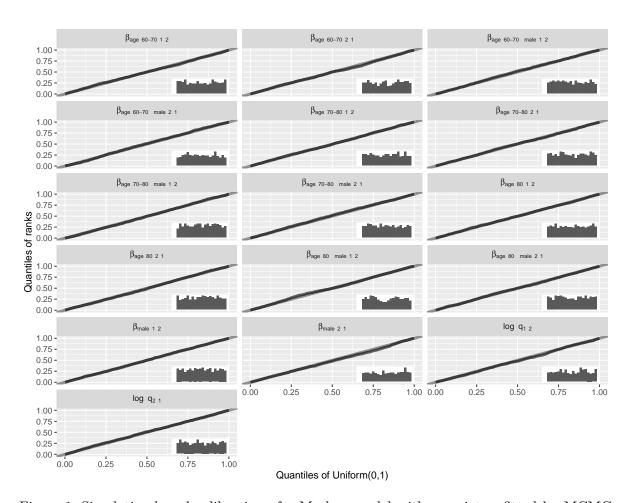


Figure 1: Simulation-based calibration of a Markov model with covariates, fitted by MCMC. The distribution of the rank statistic for the model parameters (log transition intensities q and log hazard ratios β) over simulations is compared to a standard uniform.

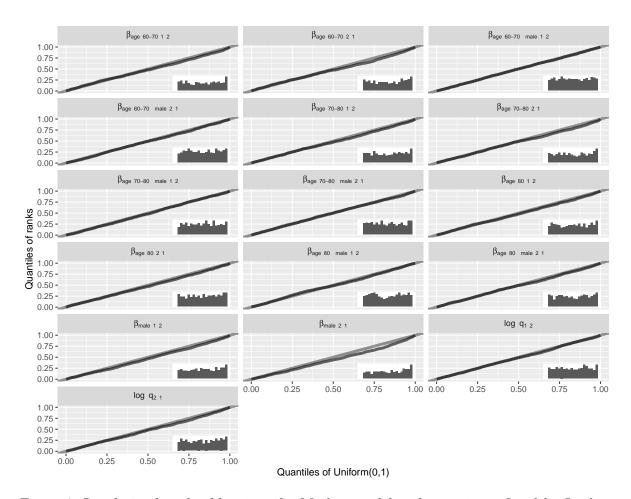


Figure 2: Simulation-based calibration of a Markov model with covariates, fitted by Laplace approximation around the posterior mode. The distribution of the rank statistic for the model parameters (log transition intensities q and log hazard ratios β) over simulations is compared to a standard uniform.

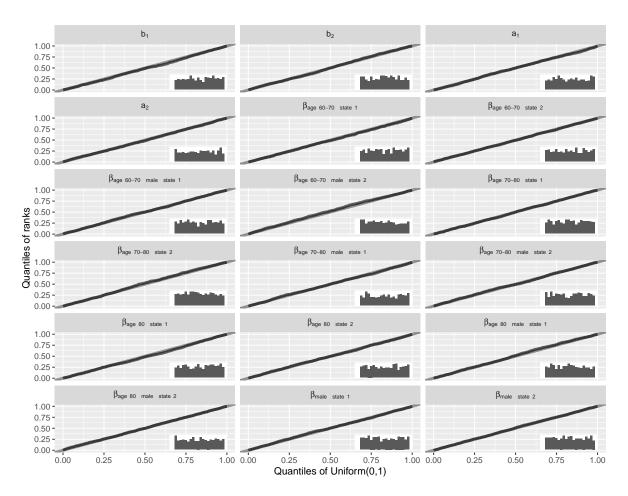


Figure 3: Simulation-based calibration of a semi-Markov model with covariates and no competing risks, fitted by MCMC. The distribution of the rank statistic for each model parameter (notation as in Section 4-5) over simulations is compared to a standard uniform.

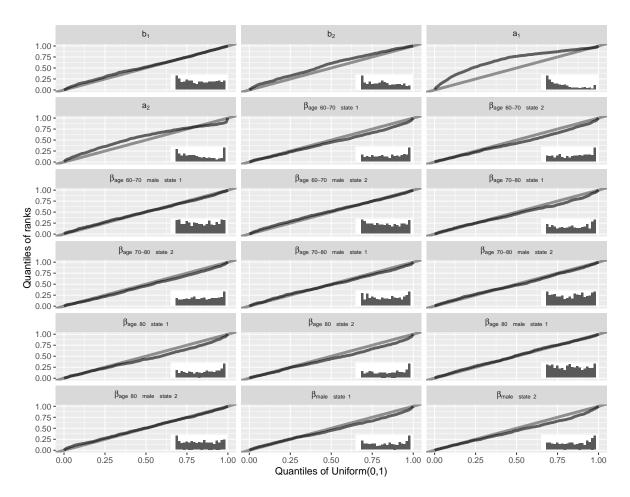


Figure 4: Simulation-based calibration of a semi-Markov model with covariates and no competing risks, fitted by Laplace approximation around the posterior mode. The distribution of the rank statistic for each model parameter over simulations is compared to a standard uniform.

Appendix 4: specification of prior distributions for semi-Markov models

Christopher Jackson

This is Supplementary Appendix 4 of the paper "Stable and practical semi-Markov modelling of intermittently-observed data" (Jackson).

- Section 1 describes some general principles that can be used to obtain priors in shapescale semi-Markov multistate models.
- Section 2 details how those principles were used to obtain priors in the ELSA application in the main manuscript, including R code for reproducibility.

1 Prior distributions in semi-Markov phase-type approximation models

In the Markov model, suppose we have obtained log-normal prior distributions for the transition rates $q_{r,s}$, and log hazard ratios $\boldsymbol{\beta}_{r,s}$ describing effects of covariates on the rates. Here we describe how to obtain priors for the analogous parameters of the semi-Markov model that agree substantively with these judgements.

To describe transitions from a given state r, the semi-Markov model includes the shape a and scale b for the sojourn distribution, next-state probabilities p_{r,s_j} , and effects of covariates on the sojourn time (via parameters β) and the next-state probabilities (via parameters γ). See Section 4.2 of the main manuscript. The dependence of these parameters on the state r is dropped in the notation here for clarity.

1.1 Shape and scale parameter

We suppose we do not have substantive beliefs about the shape parameter a, and specify vague normal priors truncated within the range of permitted values. This depends on the number of phases in the approximation, see Appendix 2, e.g. 5 phases supports Weibull shapes up to 2 and Gamma shapes up to 5. The prior variance is chosen to balance numerical stability with

the flexibility of the resulting distribution family. Specifically, for the Gamma distribution, a SD of 0.5, supporting shapes around 0.4 to 2.7 with 95% probability, and for the Weibull a SD of 0.25, supporting shapes around 0.6 to 1.6. Below these lower limits, the distributions become increasingly skewed (see Figures 2 and 3 in Appendix 2).

A prior for the scale parameter b is then deduced from the shape prior via our beliefs about the mean sojourn time, as follows:

- 1. Given our choice of log-normal priors for the q_{rs} , we can draw a sample from the implied prior for the mean sojourn time $T_r = 1/\sum_s q_{rs}$ in the Markov model, which we suppose represents our judgement about the mean sojourn time.
- 2. In the semi-Markov model, since the phase-type sojourn distribution is obtained by moment matching, its mean is the same as the mean of the distribution that it is designed to approximate. Hence the mean sojourn time under the Weibull model is $b\Gamma(1+1/a)$, and under the Gamma model a/b.
- 3. Therefore given any choice of prior for b, and our previously-specified prior for a, we can sample from the implied prior for the mean sojourn time in the semi-Markov model. Hence a prior for b that approximates our the judgement about T_r in step (1) can be deduced by a numerical search.

1.2 Next-state probabilities

In the Markov model, the probability that the next state on exit from state r is state s_j is $p_{r,s_j} = q_{r,s_j}/T_r$. In the semi-Markov model, we specify a log-normal prior for the odds $p_{r,s_j}/p_{r,s_1}$ of the next state being s_j (relative to the first of the potential destination states s_1).

Note that in the Markov model, this odds equals the relative rate $q_{r,s_j}/q_{r,s_1}$, since the mean sojourn time T_r cancels. Therefore a normal prior representing our belief about the log odds in the semi-Markov model can be determined by summarising a sample from the distribution of the log relative rate implied by our choice of priors for q_{r,s_j}, q_{r,s_1} .

1.3 Covariate effects on the scale parameter

The semi-Markov model has the accelerated failure time property (see Section 4.2 of the main manuscript), so the multiplicative covariate effect on the scale parameter can be interpreted as an effect on the mean sojourn time. Hence the prior for this effect in the semi-Markov model can be derived by simulating the distribution of this effect implied by the priors in the Markov model, as follows.

Extending the notation to allow the sojourn time $T_r(\mathbf{x})$ to depend on covariates, we draw a sample from the distribution of $\log(T_r(\mathbf{x}=1)) - \log(T_r(\mathbf{x}=0))$ implied by the priors for

the rates and hazard ratios in the Markov model, and set the effect β on the scale in the semi-Markov model to have the same mean and variance. This approach of contrasting two arbitrary values of \mathbf{x} also assumes that the effect of \mathbf{x} on $\log(T_r)$ in the semi-Markov model is linear within the range of \mathbf{x} used in the data (which is not implied exactly by linearity of the effect on the log rate in the Markov model).

1.4 Covariate effects on next-state probabilities

The covariate effects on the next-state probabilities are quantified by a parameter $\gamma_{r,j}$, the multiplicative effect of a unit increase in the covariate on the odds $p_{r,s_j}/p_{r,s_1}$ of transition to s_j , relative to the first competing risk s_1 .

Extending the notation to allow the rates in the Markov model to depend on covariates as $q_{r,s_j}(\mathbf{x}) = \exp(\beta_{r,s_j}\mathbf{x})q_{r,s_j}$, the additive effect of a unit increase in the covariate value on the log relative rate (equal to the log odds, as explained above) is

$$\left\{ \log(q_{r,s_j}(\mathbf{x}=1)) - \log(q_{r,s_1}(\mathbf{x}=1)) \right\} - \left\{ \log(q_{r,s_j}(\mathbf{x}=0)) - \log(q_{r,s_1}(\mathbf{x}=0))) \right\}$$

which equals $\beta_{r,s_j} - \beta_{r,s_1}$. Therefore a normal prior for $\gamma_{r,j}$ in the semi-Markov model can be defined with mean and variance taken from the sum of the means and the sum of the variances of these two β terms.

2 Priors used in the application to the English Longitudinal Study of Ageing

The R code in this section is included to aid reproducibility of this analysis. The text is intended to be understandable without the code.

2.1 Priors used in Markov model

Priors for the rates of transition to death, and the effect of age and sex on these, are informed by published mortality rates for 2010-2012 in England, by year of age and sex (Office for National Statistics (2015)). Specifically, the prior means of $\log(q_{rs})$ and $\log(\beta_{rs})$ are derived from the intercept (representing men aged 50 years) and the coefficients of a linear regression of the published log rate against year of age, sex and their interaction. We suppose that we are confident in this background data to a degree represented by a prior standard deviation of 0.1 times the magnitude of the prior mean.

```
nsim <- 1000000
lqbase <- -5.449
lhr_y10 <- 0.836
lhr_female <- -0.411
lhr_y10_female <- 0.022
d_confidence <- 0.1
drate_sam <- exp(rnorm(nsim, lqbase, d_confidence*abs(lqbase)))</pre>
```

We then suppose the transition rates between states of cognitive function are of a similar order of magnitude to the death rates, hence use the same prior mean, but with a weaker confidence, and represented by a standard deviation of 0.5 times the mean. The intention of this is to bound the times to events roughly within the expected lifetimes of the study population. (A more advanced approach would be to consider previous studies of cognitive decline.)

Mean sojourn times

These priors imply a prior for the rate governing the sojourn distribution in each state r. This is given by a sum of two or three log-normal distributions, one for each potential next state for state r. See the transition structure in Figure 7 of the main manuscript: there are two potential next-states following states 1 (progression and death) and 4 (recovery and death), and three potential next states (progression, recovery and death) following states 2 and 3. A sample from the sojourn rate can be obtained as a sum of these log-normal samples. The reciprocal of the sojourn rate is the mean sojourn time. We verify that the priors for the mean sojourn times are within a plausible order of magnitude (i.e. in the sense of closer to 100 than 1000) for years to death or other events in a human population, giving weakly informative priors.

2.2 Priors used in semi-Markov model

Shape and scale

For the semi-Markov model, we then need to deduce a prior for the scale parameter of the sojourn distribution, and for the next-state probability. First we define a vague prior for the shape parameter. The prior on the scale parameter is obtained by a numerical search to match the above judgement about the mean sojourn time (Weibull in this case).

Specifically, we write a function to compute the median and interquartile range (med,IQR) of the prior on the mean sojourn time that are implied by a given mean and SD for the prior on the scale. The scale prior mean and SD are then chosen by numerically minimising the sum of squared differences from the originally-judged median and IQR: $(med - med_0)^2 + (IQR - IQR_0)^2$. It is verified that the quantiles of the resulting sojourn distribution roughly match the belief used in the Markov model.

75% 93.18876 160.86784

```
scaleprior_to_soj_quantiles(scale23[1], scale23[2])
```

75% 34.15081 80.89037

Next-state probability

Then for the next-state probability, we need to deduce the prior on the parameter $\gamma = q_{r,d}/q_{r,1}$, the relative rate of transition to death, relative to the rate of transition to the "baseline" competing risk in state r, e.g. if this is state 1, $\log(q_{r,d}) - \log(q_{r,1})$. Under the Markov model, we have a $N(\mu_d, (0.1\mu_d)^2)$ for $\log(q_{r,d})$ and a $N(\mu_d, (0.5\mu_d)^2)$ prior for $\log(q_{r,1})$, where μ_d is the prior mean used for all transition rates (Section 2.1). The prior for $\log(\gamma)$ in the semi-Markov model is then the sum of these, $N(0, (0.1^2 + 0.5^2)\mu_d^2)$.

```
sd_loggam <- sqrt(d_confidence^2 + c_confidence^2)*abs(lqbase)</pre>
```

Covariate effects on scale parameter

The principles described in Section 1.3 are used here, by drawing samples from the priors of the transition rates and sojourn rates for covariate values of $\mathbf{x} = 0$ and $\mathbf{x} = 1$, transforming to a sample for the difference in mean sojourn times, which is summarised to get the prior mean and variance. This is done for each covariate (age, sex, age/sex interaction, education level).

```
lhr_to_soj_msd <- function(lhr,lhrsd,ncrisks=1){</pre>
  drate_x1 <- exp(rnorm(nsim, lhrsd))*drate_sam</pre>
  crrate x1 <- exp(rnorm(nsim, 0, 1))*crrate sam</pre>
  srate_x1 <- drate_x1 + crrate_x1</pre>
  srate x0 <- drate sam + crrate sam</pre>
  if (ncrisks==2) {
    crrate2_x1 <- exp(rnorm(nsim, 0, 1))*crrate2_sam</pre>
    srate_x1 <- srate_x1 + crrate2_x1</pre>
    srate_x0 <- srate_x0 + crrate2_sam</pre>
  beta_soj <- log(srate_x1) - log(srate_x0)</pre>
  c(mean(beta_soj), sd(beta_soj))
lta_y10_14 <- lhr_to_soj_msd(lhr_y10, d_confidence*abs(lhr_y10))</pre>
lta_y10_23 <- lhr_to_soj_msd(lhr_y10, d_confidence*abs(lhr_y10), ncrisks=2)</pre>
## compromise between N(0,1) and informative prior on death rate
lta_female_14 <- lhr_to_soj_msd(lhr_female, d_confidence*abs(lhr_female))</pre>
lta_female_23 <- lhr_to_soj_msd(lhr_female, d_confidence*abs(lhr_female),</pre>
                                    ncrisks=2)
lta_inter_14 <- lhr_to_soj_msd(lhr_y10_female, d_confidence*abs(lhr_y10_female))</pre>
lta_inter_23 <- lhr_to_soj_msd(lhr_y10_female, d_confidence*abs(lhr_y10_female),</pre>
                                    ncrisks=2)
lta_educ_14 <- lhr_to_soj_msd(0, 1, ncrisks=1)</pre>
lta_educ_23 <- lhr_to_soj_msd(0, 1, ncrisks=2)</pre>
dump(c(
"scale14", "scale23", "sd_loggam",
"lta_y10_14","lta_y10_23","lta_female_14","lta_female_23",
"lta_inter_14", "lta_inter_23", "lta_educ_14", "lta_educ_23"), file="")
scale14 <-
c(4.499999999999991, 1.149999999999999)
scale23 <-
c(3.4968750000000002, 1.484375)
sd loggam <-
2.7784557329567083
lta_y10_14 <-
c(0.15774303168248024, 0.86368348280138285)
lta_y10_23 <-
c(0.1843789541425275, 0.80994370968281904)
```

```
lta_female_14 <-
c(0.13716341485806524, 0.86313132548364113)
lta_female_23 <-
c(0.1742754534025506, 0.80867313394779472)
lta_inter_14 <-
c(0.11712907281177055, 0.86320851049878722)
lta_inter_23 <-
c(0.15977533896978316, 0.8095176160634735)
lta_educ_14 <-
c(0.67030802612214158, 0.92747010097200067)
lta_educ_23 <-
c(0.50286314437654134, 0.85826648213044643)</pre>
```

Similarly for the effects on the next-state probabilities, the prior can be deduced analytically from the means and variances of the appropriate Markov hazard ratios, following the principles in Section 1.4.

Though in this case, these are practically equivalent to the vague N(0,1) priors placed on the effects on transition rates between living states, since the "informative" prior variances for the effects on death are very small by comparison with the "vague" variance of 1. Therefore for convenience, in this illustrative example, these are given weakly informative priors with variances of 1, in the same way as the hazard ratios in the Markov model.

```
vbdeath_y10 <- d_confidence^2*abs(lhr_y10)^2
vbdeath_female <- d_confidence^2*abs(lhr_y10_female)^2
vbdeath_y10_female <- d_confidence^2*abs(lhr_y10_female)^2
vbdeath_y10</pre>
```

[1] 0.00698896

```
vbdeath_female
```

[1] 4.84e-06

```
vbdeath_y10_female
```

[1] 4.84e-06

Office for National Statistics. 2015. "English Life Tables No.17: 2010 to 2012." URL: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/englishlifetablesno17/2015-09-01.