8-3-2012

Fitting and Interpreting Continuous-Time Latent Markov Models for Panel Data

Jane M. Lange
University of Washington - Seattle Campus, langej@u.washington.edu

Vladimir N. Minin
University of Washington, vminin@uw.edu

Abstract

Multistate models are used to characterize disease processes within an individual. Clinical studies often observe the disease status of individuals at discrete time points, making exact times of transitions between disease states unknown. Such panel data pose considerable modeling challenges. Assuming the disease process progresses according a standard continuous-time Markov chain (CTMC) yields tractable likelihoods, but the assumption of exponential sojourn time distributions is typically unrealistic. More flexible semi-Markov models permit generic sojourn distributions yet yield intractable likelihoods for panel data in the presence of reversible transitions. One attractive alternative is to assume that the disease process is characterized by an underlying latent CTMC, with multiple latent states mapping to each disease state. These models retain analytic tractability due to the CTMC framework but allow for flexible, duration-dependent disease state sojourn distributions. We have developed a robust and efficient expectation-maximization (EM) algorithm in this context. Our complete data state space consists of the observed data and the underlying latent trajectory, yielding computationally efficient expectation and maximization steps. Our algorithm outperforms alternative methods measured in terms of time to convergence and robustness. We also examine the frequentist performance of latent CTMC point and interval estimates of disease process functionals based on simulated data. The performance of estimates depends on time, functional, and data-generating scenario. Finally, we illustrate the interpretive power of latent CTMC models for describing disease processes on a data-set of lung transplant patients. We hope our work will encourage wider use of these models in the biomedical setting.

1 Introduction

Disease processes refer to the natural history of a disease within an individual. These histories can be conceptualized as consisting of sojourns in discrete states that individuals pass through according to progressive or reversible transitions; the final transition is to the absorbing state, death. Discrete-space continuous-time multistate models are useful in describing these processes. Examples include models of HIV (Guihenneuc-Jouyaux et al., 2000), HSV-2 (Crespi et al., 2005), and multiple sclerosis (Mandel, 2010). These models allow one to characterize sojourn time distributions in each state, predict disease and mortality rates based on an individual's covariates and history, and describe population level patterns such as disease state prevalence.

Fully observed disease process trajectories present many options for model fitting (Andersen and Keiding, 2002). Panel data, consisting of snapshots of the process at discrete times on multiple individuals, present challenges for inference. We assume that the sampling frame is independent of the underlying process, except for possibly known times of death, and that observation times are not necessarily evenly spaced and may vary across subjects.

In the panel observation setting, one typically assumes that the observed data are generated by a discretely observed continuous-time Markov chain (CTMC). This family of models enjoys tractable likelihoods and has established methods of obtaining maximum likelihood estimates (MLEs) for transition intensities (Kalbfleisch and Lawless, 1985; Lange, 1995). CTMCs entail two strong assumptions: a) the Markov property indicates that transition probabilities depend on an individual's history only through the current state, and b) sojourn distributions are exponential, so that the rate of leaving a state does not depend on occupancy duration.

Ideally, we would like to fit panel data using more flexible models. Semi-Markov models present one class of alternatives, in which the sequence of states is Markov, but sojourn distributions may have any form and need not be exponential. In general, however, data from discretely observed semi-Markov processes result in likelihoods that are very difficult to compute, particularly if there are reversible transitions. Methods for fitting semi-Markov models to panel data are limited to special cases, such as progressive processes (Foucher et al., 2007) or processes in which some states have exponential sojourn distributions (Kang and Lagakos, 2007).

Titman and Sharples (2010) proposed modeling discretely observed multistate disease processes with a latent state CTMC. Each disease state maps to multiple latent states, which are traversed according to an underlying CTMC. This framework yields hazard rates of transitioning between disease states that depend on the duration spent in that state; yet likelihoods are analytically tractable, even for disease processes with reversible transitions.

A latent CTMC structure implies phase-type (PH) distributions of sojourn times in disease states. PH distributions are attractive since they can approximate generic distributions with positive support (Cumani, 1982); and PH functionals, such as hazard rates and cumulative distribution functions (CDFs), are easily expressible with matrix exponentials. Aalen (1995) reviews properties of PH distributions with applications to survival outcomes. The disadvantage of PH distributions is that model parameters may not be identifiable, compromising estimation in a frequentist setting. Fortunately, scientifically meaningful functionals describing sojourn time distributions typically are identifiable (Bladt et al., 2003). Latent CTMC models of disease processes inherit both these advantages and disadvantages.

Our focus is on parameter estimation of the latent CTMC model in the panel data setting. Titman and Sharples (2010) describe how these data fit into a hidden Markov model (HMM) framework based on an underlying discretely observed CTMC, with or without misclassification error. The observed data likelihood is obtainable from the recursive Baum-Welch forward-backward algorithm for HMMs (Baum et al., 1970). Since the transition probability matrices of the latent trajectory relate to the intensity matrix via matrix exponentials, obtaining MLEs of latent CTMC parameters is less straightforward than simply running the Baum-Welch algorithm.

Titman and Sharples (2010) suggest standard numerical optimization methods for obtaining latent model MLEs. In our experience, these methods are slow, sensitive to starting values, and exhibit poor convergence properties. Here we propose a novel expectation-maximation (EM) algorithm. EM algorithms assume a complete data space underlying the observed data whose likelihood is easy to maximize. MLEs are obtained through iterative maximizations of the expected complete data log-likelihood conditional on observed data and current parameter estimates (Dempster et al., 1977). Our complete data space consists of the underlying latent trajectory and the observed data at discrete time points. These yield exponential family score equations that can be solved easily with either an analytic maximization step (M-step) or with a few iterations of the Newton-Raphson algorithm.

Bureau et al. (2003) developed an alternative EM method for this setting that considers the complete data as the observed data plus latent CTMC states at each observation time. Their M-step is less stable and computationally more costly than our approach. We show that our EM method has better performance than both direct maximization of the observed data likelihood and the EM alogrithm of Bureau et al. (2003), particularly when we apply the EM-acceleration of Varadhan (2011).

Our EM algorithm uniquely combines computational developments derived for PH models (Asmussen et al., 1996) and discretely observed CTMCs (Hobolth and Jensen, 2005) and uses efficient methods developed for HMMs to sum over the latent states (Cappe et al., 2005). Our EM method shares a similar complete data space and E-step as the EM algorithm that Roberts and Ephraim (2008) developed for HMMs based on discretely observed CTMCs. However, our approach is considerably more general, as it accommodates known times of absorption and allows for covariates in the latent CTMC model. We also construct an exact method of calculating the Hessian matrix for model parameters using the recursive smoothing framework described by Cappe et al. (2005).

In addition to our algorithmic developments, we focus on the practical application and interpretation of latent CTMC models. Their value hinges on their ability to describe disease processes with generic sojourn distributions. Models with few latent states are more likely to result in identifiable parameters, but point estimates for disease process functionals, such as sojourn time hazard and CDFs, may be biased, and interval estimates may have poor coverage. We investigate these aspects by fitting latent CTMCs to discretely and fully observed processes simulated from known distributions. Others have investigated the use of phase-type models to approximate generic distributions (Faddy, 1998; Asmussen et al., 1996; Marshall and Zenga, 2010), but to our knowledge, no one has examined their performance with discretely observed data or investigated confidence interval coverage.

Finally, we re-analyze the bronchiolitis obliterans syndrome (BOS) dataset from Titman and Sharples (2010), both to compare performance of different fitting methods and to illustrate model interpretation, emphasizing clinically relevant functionals of the disease process (Andersen and Keiding, 2012). This application highlights the benefit of latent CTMC models for describing sojourn distributions and demonstrates the superior speed and robustness of our EM algorithm on real data against other methods for obtaining MLEs.

2 Model description

2.1 Latent CTMC parameterization

Let $W(t)$ be the disease process trajectory with disease state space $R = \{1, 2, \ldots, r\}$. Underlying $W(t)$ is a time-homogeneous CTMC, $X(t)$, with latent state space

$$S = \{1_1, 1_2, \ldots, 1_{s_1}\} \cup \{2_1, 2_2, \ldots, 2_{s_2}\} \cup \cdots \cup \{r_1, r_2, \ldots, r_{s_r}\},$$

intensity matrix $\mathbf{\Lambda}$, and initial distribution $\boldsymbol{\pi}$. We assume that S has $s = \sum_{k=1}^{r} s_k$ states. Each observable disease state maps to multiple states in the latent state space. Thus, $W(t) = j <=> X(t) \in \{j_1, j_2, \ldots, j_{s_j}\}$. For example, Figure 1A shows a latent trajectory $X(t)$ and the corresponding disease trajectory $W(t)$ for a 2-state reversible disease model.

The mapping of multiple latent states in S to a single disease state in R yields phase-type, not exponential, sojourn distributions of W(t). Generally, PH distributions characterize time-to-event variables as time to absorption in an underlying CTMC. To promote parsimony, Titman and Sharples (2010) specify the sojourn distributions of $W(t)$ to have Coxian PH structure. Coxian PH models assume the process starts in the first transient state and at each transition either proceeds forward or exits to an absorbing state (Figure 1B). These restrictions induce sparseness in $\mathbf{\Lambda}$. Figure 1C shows the allowable transitions of $X(t)$ when $W(t)$ consists of a 2-state reversible disease model with Coxian PH sojourn time distributions, corresponding to the trajectory plotted in Figure 1A. The framework can also be scaled for more complex disease models, including those where an individual in disease state $p \in R$ can transition to disease states u or v. The allowable transitions are similar; when $X(t)$ is in latent state p_k, it can proceed forward to p_{k+1} or exit to either latent state u_1 or v_1.

2.2 Observed data likelihood

The panel data with state space R may be observed with or without misclassification error. Latent states at each observation time will be denoted by $x_1, \ldots, x_n$, and observed data by $o_1, \ldots, o_n$. Observed data are conditionally independent given $W(t)$ at observation times $t_1, \ldots, t_n$. Thus, the relationship between observed and latent states is described by an emission matrix $\mathbf{E} = \{e(i, j)\}$ with entries $e(i, j) = \mathrm{P}(O_t = j | X(t) = i)$ that satisfy the identity $e(i,k) = e(j,k)$ for all latent states $i, j \in \{p_1, \ldots, p_{s_p}\}$ and observed values k.

Given the HMM formulation, the observed data likelihood is

$$\mathrm{P}(\mathbf{o}) = \sum_{x_1} \sum_{x_2} \ldots \sum_{x_n} \pi_{x_1} \prod_{i=2}^{n} \mathrm{P}_{x_i x_{i+1}}(t_{i+1} - t_i) \prod_{i=1}^{n} e(x_i, o_i), \tag{1}$$

where $\mathrm{P}_{x_i x_{i+1}}(t_{i+1} - t_i) = \mathrm{P}(X(t_{i+1}) = x_{i+1} | X(t_i) = x_i)$ and $\pi_{x_1} = \mathrm{P}(X(t_1) = i)$. For some individuals the time to absorption(death), Y, is known. When the last observation time $t_n = y$, the observed data likelihood, $\frac{\partial}{\partial y} \mathrm{P}(\mathbf{o}, Y < y)$ is similar to equation 1. The only difference is that $\mathrm{P}_{x_{n-1} x_n}(t_n - t_{n-1})$ is replaced by $f(t_n - t_{n-1} | X_{n-1} = x_{n-1})$, the density of Y given state x_{n-1} at time t_{n-1}.

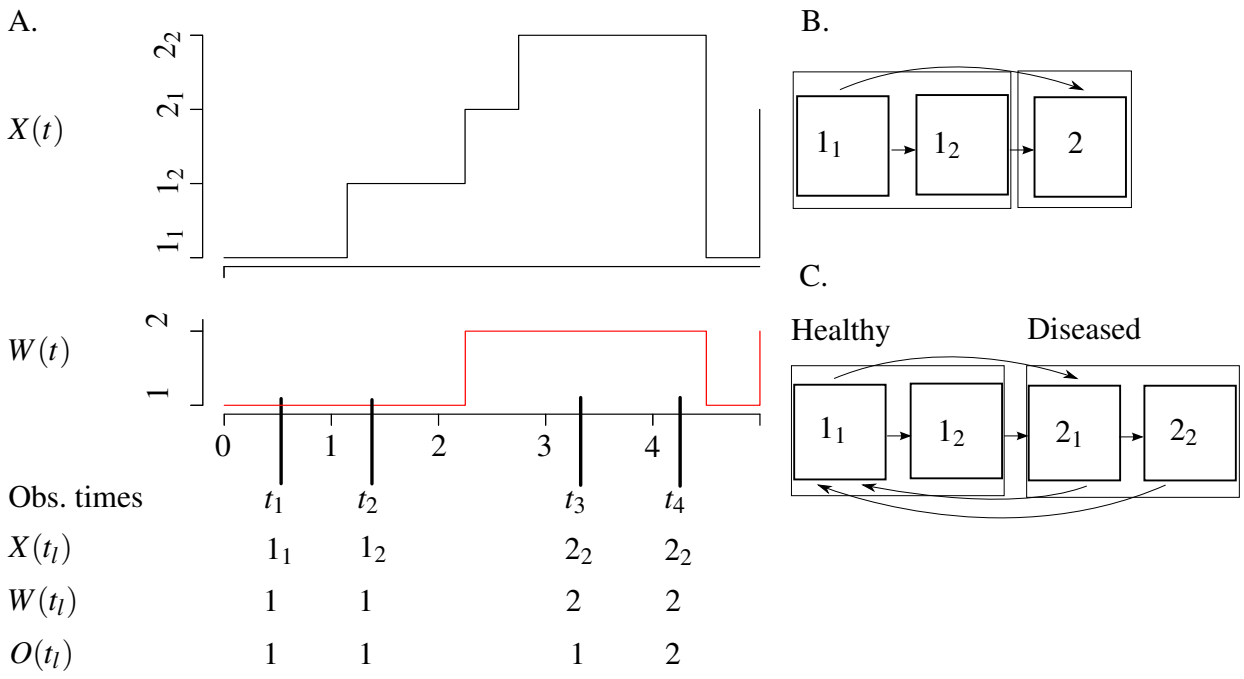

Figure 1: A. Example of latent trajectory $X(t)$, disease trajectory $W(t)$, and observed data $O(t_l)$ at discrete observation times for model in subfigure C, assuming possible misclassification error. B. 2-state survival model of W(t) assuming $R = \{1,2\}$ and $S = \{\{1_1,1_2\},\{2\}\}$, where disease state 2 is absorbing. Coxian PH structures implies X(t) starts in 1_1. C. 2-state reversible model of W(t), with state space $R = \{1 = Healthy, 2 = Diseased\}$ and $S = \{\{1_1,1_2\},\{2_1,2_2\}\}$). $X(t)$ starts in 1_1 or 2_1.

2.3 Adding covariates to the latent CTMC model

We can parameterize $\mathbf{\Lambda}$ in the latent CTMC model by the log-rates $\{\log(\lambda_{ij}) : i,j \in S; i \neq j\}$. To incorporate baseline subject-level covariates $\mathbf{w}^h$, we set $\log(\lambda_{ij}^h) = \boldsymbol{\beta}_{ij}^T \mathbf{w}^h$, where h denotes the individual. More parsimonious models equate individual covariate effects across rate parameters. In particular, the assumption that a covariate has a multiplicative effect on the sojourn time in disease state p is achieved by equating the covariate effect across all log rates $\left\{\log(\lambda_{ij}) : i \in \{p_1,\ldots,p_{s_p}\}\right\}$. Initial distributions and emission distributions are multinomial. The initial latent state is captured by an indicator vector $\mathbf{Z} = (Z_1,\ldots,Z_s)$, where $Z_i = I(X_1 = i)$. Thus $\mathbf{Z} \sim \text{Multinomial}(\boldsymbol{\pi},1)$. The initial distribution $\boldsymbol{\pi}$ has natural parameters $\left\{\eta_i = \log\left(\frac{\pi_i}{\pi_1}\right) : i = 2,\ldots,s\right\}$, and the emission distribution $\mathbf{e}_i$ has natural parameters $\left\{\eta_{ij} = \log\left(\frac{e(i,j)}{e(i,1)}\right) : j = 2,\ldots,r\right\}$. Subject-level covariates $\mathbf{w}^h$ are added to the multinomial models via a linear predictor by taking $\eta_{ij}^h = \boldsymbol{\gamma}_{ij}\mathbf{w}^h$.

2.4 Complete data likelihood

We assume m independent subjects. The vector $(\mathbf{o},\mathbf{x})$ denotes the complete data (observed data and underlying latent trajectory) for a given subject. The model parameters $\boldsymbol{\theta} = (\boldsymbol{\pi},\mathbf{\Lambda},\mathbf{E})$ characterize the initial distribution, CTMC transitions, and emission probability matrix, respectively. The complete data log-likelihood

has exponential family form and is a linear function of complete data sufficient statistics. For a subject these sufficient statistics include $n_T(i,j)$, the total counts of transitions from state i to state j; $d_T(i)$, the total duration spent in state i; z_i, the initial latent state indicator ; and $o_T(i,j) = \sum_{l=1}^{n} I(x_l = i)I(o_l = j)$, the total co-occurrences of latent state i and observed state j.

For this subject, the complete data log-likelihood (LL) has the factored form

$$\begin{aligned} l(\boldsymbol{\theta};\mathbf{o},\mathbf{x}) &= l(\boldsymbol{\pi};x_1) + l(\boldsymbol{\Lambda};\mathbf{x}|x_1) + l(\mathbf{E};\mathbf{o}|\mathbf{x},x_1) \\ &= \sum_{i}^{s} z_i \log(\pi_i) + \sum_{i=1}^{s}\sum_{j\neq i} n_T(i,j)\log(\lambda_{ij}) - \sum_{i=1}^{s} d_T(i)\left(\sum_{j\neq i}^{s}\lambda_{ij}\right) \\ &+ \sum_{i=1}^{s}\sum_{j=1}^{r} o_T(i,j)\log\{e(i,j)\}. \end{aligned} \tag{2}$$

The separation of parameters in the factored log-likelihood means that $\boldsymbol{\pi}, \boldsymbol{\Lambda}$ and $\mathbf{E}$ can be dealt with one by one. Moreover, given the independence of individual subjects, the score and information are additive, such that $\dot{l}(\boldsymbol{\theta}) = \sum_{h=1}^{m} \dot{l}_h(\boldsymbol{\theta})$ and $\ddot{l}(\boldsymbol{\theta}) = \sum_{h=1}^{m} \ddot{l}_h(\boldsymbol{\theta})$, where h indexes the score or information contribution of individual h.

3 EM algorithm

3.1 M-step

The exponential family form of the complete data log-likelihood enables a straightforward M-step in the EM algorithm. The score vectors and Hessian matrices for $\boldsymbol{\Lambda}$, $\boldsymbol{\pi}$ and $\mathbf{E}$ are provided in Web Appendix A. In the absence of covariates, the score equations solved in the M-step have closed-form solutions, namely $\hat{\lambda}_{ij} = \frac{\sum_{h=1}^{m} n_T^h(i,j)}{\sum_{h=1}^{m} d_T^h(i)}$, $\hat{e}_{ij} = \frac{\sum_{h=1}^{m} o_T^h(i,j)}{\sum_{h=1}^{m}\sum_{j=1}^{r} o_T^h(i,j)}$ and $\hat{\pi}(i) = \frac{\sum_{h=1}^{m} Z_i^h}{m}$, where h denotes an individual. With covariates, the score equations can be solved using the Newton-Raphson algorithm, which requires the Hessian as well as the score. Generally, the rth iteration of the Newton-Raphson method for parameter $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta}^{(r)} = \boldsymbol{\theta}^{(r-1)} - \ddot{l}\left(\boldsymbol{\theta}^{(r-1)}\right)^{-1} \dot{l}\left(\boldsymbol{\theta}^{(r-1)}\right)$. This procedure can be applied separately to update the parameter vectors corresponding to $\boldsymbol{\pi}, \boldsymbol{\Lambda}$ and $\mathbf{E}$. In fact, Newton-Raphson need not be run to convergence, as a single update will still yield the same EM convergence properties as full maximization (Lange, 1995).

3.2 E-step

The expectation step (E-step) requires computing the expectation of the complete data log-likelihood (2) conditional on the observed data. The log-likelihood for an individual is additive across time intervals $T_l = [t_{l-1}, t_l]$. Hence,

$$\begin{aligned} \mathrm{E}[l(\boldsymbol{\theta};\mathbf{o},\mathbf{x})] &= \sum_{i=1}^{s} \mathrm{E}[z_i|\mathbf{o}]\log(\pi_i) + \sum_{l=1}^{n}\sum_{i=1}^{s}\sum_{j\neq i} \mathrm{E}[n_{T_l}(i,j)|\mathbf{o}]\log(\lambda_{ij}) \\ &- \sum_{l=1}^{n}\sum_{i=1}^{s} \mathrm{E}[d_{T_l}(i)|\mathbf{o}]\left(\sum_{j\neq i}\lambda_{ij}\right) + \sum_{l=1}^{n}\sum_{i=1}^{s}\sum_{j=1}^{r} \mathrm{E}[o_{T_l}(i,j)|\mathbf{o}]\log\left(e(i,j)\right). \end{aligned}$$

This reduces the E-step to finding the conditional expectation of the complete data sufficient statistics across T_l.

Conditional expectations for z_i and $o_{T_l}(i,j)$ are computed as in the Baum-Welch algorithm, using the smoothing probabilities $\mathrm{P}(X_l = x_l|\mathbf{o}) = \frac{b_l(m)\alpha_l(m)}{\mathrm{P}(\mathbf{o})}$, where $\alpha_l(m)$ and $\beta_l(m)$ are HMM forward and backward probabilities (Web Appendix B) and $\mathrm{P}(\boldsymbol{o})$ refers to equation 1. Hence

$$\mathrm{E}[z_i|\mathbf{o}] = \mathrm{P}(X_1 = i|\mathbf{o}) = \frac{\beta_l(m)\alpha_l(m)}{\mathrm{P}(\mathbf{o})}$$

and

$$\mathrm{E}[o_T(j,m)|\mathbf{o}] = \sum_l I(O_l = m)\,\mathrm{P}(X_l = j|\mathbf{o}) = \sum_l \frac{\beta_l(m)\alpha_l(m)}{\mathrm{P}(\mathbf{o})}.$$

Expectations of $d_{T_l}(i)$ and $n_{T_l}(i,j)$ can be obtained by first conditioning on the latent states x_l and x_{l+1}, that is

$$\mathrm{E}[d_{T_l}|\mathbf{o}] = \mathrm{E}\left[\mathrm{E}\left(d_{T_l}|\mathbf{o}, X_l = a, X_{l+1} = b\right)\right] = \mathrm{E}\left[\mathrm{E}\left(d_{T_l}|X_l = a, X_{l+1} = b\right)|\mathbf{o}\right],$$

and likewise for $n_{T_l}(i,j)$. Thus, we break the task down into finding the "inner" expectations, $\mathrm{E}[d_{T_l}|X_l = a, X_{l+1} = b]$ and $\mathrm{E}[n_{T_l}(i,j)|X_l = a, X_{l+1} = b]$, and the "outer" expectations, which involve summing over the latent states conditional on the observed data.

3.2.1 Inner expectations: conditional moments of occupancy durations and transition counts

In a general time-homogeneous CTMC, we express conditional expectations of transition counts $n_t(i,j)$ and occupancy durations $d_t(i)$ in terms of the joint expectations $\mathrm{E}[n_t(i,j)I(X_0 = a)|X_t = b]$ and $\mathrm{E}[d_t(i)I(X_t = b)|X_0 = a]$ divided by $\mathrm{P}_{ab}(t)$, the probability of transitioning from a to b. These joint expectations are given by the integrals $\int_0^t \lambda_{ij}\,\mathrm{P}_{ai}(u)\,\mathrm{P}_{jb}(t-u)du$ and $\int_0^t \mathrm{P}_{ai}(u)\,\mathrm{P}_{ib}(t-u)du$, respectively (Hobolth and Jensen, 2005). We calculate the joint expectation integrals via the efficient matrix-based methods of Minin and Suchard (2008a) and Minin and Suchard (2008b). These methods assume $\mathbf{\Lambda}$ has no repeated eigenvalues and rely on eigen-decomposition. When $\mathbf{\Lambda}$ has repeated eigenvalues, we compute the integrals using the uniformization approach derived in Hobolth and Jensen (2011) and Bladt et al. (2011).

Our exact method of obtaining information of parameter estimates requires joint second and cross moments of $n_t(i,j)$ and $d_t(i)$. We define these quantities as $\mathrm{E}[n_t(i,j)n_t(l,m)I(X_t = c)|X_0 = a]$; $\mathrm{E}[d_t(i)d_t(j)I(X_t = c)|X_0 = a]$; and $\mathrm{E}[d_t(i)n_t(l,m))I(X_t = c)|X_0 = a]$. Details for these computations using eigen-decomposition are provided by Minin and Suchard (2008b), and using uniformization by Hobolth and Jensen (2011).

Joint first and second moments are also desired when the interval endpoint coincides with the time of absorption, Y. Let S refer to specific statistics of interest, such as $n_t(i,j)$, $d_t(i)$, $n_t(i,j)n_t(l,m)$, $d_t(i)d_t(j)$, or $d_t(i)n_t(l,m)$. We seek the differentiated joint moment $\frac{\partial}{\partial t}\mathrm{E}[S \times I(Y < t)|X_0 = a] = \mathrm{E}[S|X_0 = a, Y = t] \times f(t|X_0 = a)$. Methods for obtaining these moments are presented by Asmussen et al. (1996) and are described in detail in Web Appendix C.

3.2.2 Outer expectations: summing over latent states

To finish the E-step, we need to compute the "outer" expectations $\mathrm{E}[S_{T_l}|\mathbf{o}] = \mathrm{E}[\mathrm{E}[S_{T_l}|X_l = a, X_{l+1} = b]|\mathbf{o}]$, for the complete data sufficient statistics $S_{T_l} = d_{T_l}(i)$ or $n_{T_l}(i,j)$ on each time interval T_l. In order to integrate

over latent states x_l and x_{l+1}, we exploit the bivariate smoothing probabilities

$$\mathrm{P}(X_l = a, X_{l+1} = b|\mathbf{o}) = \frac{e(b, o_{l+1})\alpha_l(a)\beta_{l+1}(b)\,\mathrm{P}(X_{l+1} = b|X_l = a)}{\mathrm{P}(\mathbf{o})}$$

delivered by the Baum-Welch algorithm. Thus, the expression for the conditional expectation of the complete data sufficient statistic across the entire time interval $T = [t_1, t_n]$ is

$$\mathrm{E}[S_T|\mathbf{o}] = \sum_{l=1}^{n-1}\sum_{a=1}^{r}\sum_{b=1}^{r}\mathrm{E}[S_{T_l}|X_l = a, X_{l+1} = b]\,\mathrm{P}(X_l = a, X_{l+1} = b|\mathbf{o}).$$

In the case where t_n corresponds to a known time of absorption, y, the summand corresponding to the final interval is altered accordingly. The inner expectation is replaced by
$\mathrm{E}[S_{T_{n-1}}|X_{n-1} = a, Y = t_n]$, the transition probability is replaced by the density $f(t_n - t_{n-1}|X_{n-1} = a)$, and the denominator is replaced by $\frac{\partial}{\partial y}\mathrm{P}(\mathbf{o}, Y < y)$, the observed data likelihood with a known absorption time (section 2.2).

3.2.3 Recursive smoothing for complete data sufficient statistics

Our E-step calculates conditional expectations of complete data sufficient statistics via marginal and bivariate smoothing probabilities that condition on a subject's entire observed data, **o**. Another option is recursive smoothing, described by Cappe et al. (2005) for general HMMs. Recursive smoothing is an online method for computing expectations of a functional of the currently encountered latent states conditional on the currently encountered observations. We will abbreviate $x_1, \ldots, x_k$ by $\mathbf{x}_{1:k}$ and the first k observations $o_1, \ldots o_k$ by $\mathbf{o}_{1:k}$. The functional will be denoted by $t_k(\mathbf{x}_{1:k})$. The method requires that we can define the functional recursively, expressing $t_{k+1}(\mathbf{x}_{1:k+1})$ as a linear combination of $t_k(\mathbf{x}_{1:k})$ and functions of x_k and x_{k+1}, That is, the functional is initialized at $t_1(x_1)$ and is defined as

$$t_{k+1}(\mathbf{x}_{1:k+1}) = m_k(x_k, x_{k+1})t_k(\mathbf{x}_{1:k}) + s_k(x_k, x_{k+1}), \tag{3}$$

where $m_k(x_k, x_{k+1})$ and $s_k(x_k, x_{k+1})$ are sequences of possibly vector (or matrix) valued functions.

The ultimate target, $\mathrm{E}[t_n(\mathbf{x}_{1:n})|\mathbf{o}_{1:n}]$, is obtained through recursive updates of auxiliary functions $\tau_k(x_k) = \mathrm{E}[I(X_k = x_k)t_k(\mathbf{x}_{1:k})|\mathbf{o}_{1:k}]$, for $k = 1, \ldots, n$. At each step, $\mathrm{E}[t_k(\mathbf{x}_{1:k})|\mathbf{o}_{1:k}] = \sum_{x_k}\tau_k(x_k)$, with the final step enabling calculation of $\mathrm{E}[t_n(\mathbf{x}_{1:n})|\mathbf{o}_{1:n}]$. The auxiliary functions are initialized as

$$\tau_1(x_1) = t_1(x_1)\frac{e(x_1, o_1)\pi(x_1)}{\sum_a e(a, o_1)\pi(a)}.$$

Cappe et al. (2005) showed that updates to the auxiliary functions are given by

$$\begin{aligned}\tau_{k+1}(x_{k+1}) = \frac{\mathrm{P}(\mathbf{o}_{1:k})}{\mathrm{P}(\mathbf{o}_{1:k+1})}\Big\{\sum_{x_k}[\tau_k(x_k)m_k(x_k, x_{k+1}) + \mathrm{P}(X_k = x_k|\mathbf{o}_{1:k})s_n(x_k, x_{k+1})] \\ \times e(x_{k+1}, o_{k+1})\,\mathrm{P}_{x_k x_{k+1}}(t_{k+1} - t_k)\Big\}.\end{aligned} \tag{4}$$

Updates to the auxiliary functions require calculating the filtering probabilities $\mathrm{P}(X_k = x_k|\mathbf{o}_{1:k})$ and the conditional observed data likelihood $\mathrm{P}(O_k = o_k|\mathbf{o}_{1:k-1})$, described in Web Appendix B.

To apply recursive smoothing to the first moments of the complete data sufficient statistics, we define $t_k(\mathbf{x}_{1:k})$ as these moments on the interval $[t_1, t_k]$ conditional on $\mathbf{x}_{1:k}$. Let $\mathbf{S}$ be the vector of complete data sufficient statistics for a single subject and $\mathbf{S}[t_l, t_m]$ be these sufficient statistics confined to the interval $[t_l, t_m]$. Thus, the functional is $t_k(\mathbf{x}_{1:k}) = \mathrm{E}[\mathbf{S}[t_1, t_k]|\mathbf{o}_{1:k}]$. The functional is initialized $t_1(x_1) = \mathrm{E}[\mathbf{S}[t_1, t_1]|o_1]$ and expressed recursively as

$$t_{k+1}(\mathbf{x}_{1:k+1}) = \mathrm{E}[\mathbf{S}[t_1, t_{k+1}]|\mathbf{x}_{1:k+1}] = \mathrm{E}[\mathbf{S}[t_1, t_k]|x_{1:k}] + \mathrm{E}[\mathbf{S}[t_k, t_{k+1}]|x_k, x_{k+1}] = t_k(\mathbf{x}_{1:k}) + s_k(x_k, x_{k+1}).$$

Here, $m_k(x_k, x_{k+1}) = 1$. The specific values of $t_1(x_1)$ and $s_k(x_k, x_{k+1})$ for latent CTMC complete data sufficient statistics are provided in Web Appendix B.

There is no computational advantage to using recursive smoothing over our first method for first moment calculations. However, recursive smoothing can also be used to calculate second moments of complete data sufficient statistics conditional on $\mathbf{o}$, which are used in our exact method of computing the information matrix of latent CTMC parameter estimates. It excels for these calculations since it retains computational complexity $O(n)$ in the number of time intervals. Second moment recursions require the same quantities derived for first moments, motivating the introduction here.

4 Information and variance of parameter estimates and disease process functionals

We calculate the observed information matrix of parameter estimates using the formula of Louis (1982). Letting $\mathbf{o}^m$ and $(\mathbf{o}^m, \mathbf{x}^m)$ be the observed and complete data for all subjects, we can express the information matrix of parameter estimates using the formula of Louis (1982) as

$$\begin{aligned} -\ddot{l}(\boldsymbol{\theta}; \mathbf{o}^m) &= \mathrm{E}[-\ddot{l}(\boldsymbol{\theta}|\mathbf{o}^m] - \mathrm{Cov}[\dot{l}(\boldsymbol{\theta}|\mathbf{o}^m] \\ &= \mathrm{E}[-\ddot{l}(\boldsymbol{\theta})|\mathbf{o}^m] - \left\{\mathrm{E}[\dot{l}(\boldsymbol{\theta})\dot{l}(\boldsymbol{\theta})^T|\mathbf{o}^m] - \mathrm{E}[\dot{l}(\boldsymbol{\theta})|\mathbf{o}^m]\,\mathrm{E}[\dot{l}(\boldsymbol{\theta})|\mathbf{o}^m]^T\right\}. \end{aligned}$$

The expectation and covariances are taken with respect to the distribution of the complete data given the observed data for all subjects.

We can calculate $\mathrm{E}[-\ddot{l}(\boldsymbol{\theta})|\mathbf{o}^m]$ readily given the factorization of the log likelihood (2) and the relatively simple forms for Hessian functions (Web Appendix A) for $\boldsymbol{\pi}$, $\boldsymbol{\lambda}$ and $\mathbf{E}$. At the MLE, $\mathrm{E}[\dot{l}(\boldsymbol{\theta})|\mathbf{o}] = \mathbf{0}$, so we only need to calculate $\mathrm{E}[\dot{l}(\boldsymbol{\theta})\dot{l}(\boldsymbol{\theta})^T|\mathbf{o}^m]$. Given that the score functions are linear in the complete data sufficient statistics, we need second and cross moments of these statistics conditional on the observed data. These moments require the "inner" expectations defined in Section 3.2.1 and use recursive smoothing to integrate over latent states (Web Appendix B). Approximate interval estimates for disease process functionals such as hazard functions and first passage CDFs can be obtained with delta-method standard errors (Gentleman, 1994) (Web appendix D).

5 Implementation

We have implemented the EM algorithm in R (R Development Core Team, 2011), in the form of R package `cthmm` available at `http://r-forge.r-project.org/projects/multistate/`. The software accommodates panel data and exact times of absorption and allows for parameterized intensity, initial distribution,

and emission matrices. Computationally intensive E-step and information calculations are coded in C++ and rely on Rcpp (Eddelbuettel and François, 2011) and RcppArmadillo packages (Francois et al., 2011).

5.1 Speeding up the EM with acceleration methods

EM algorithms are robust but slow, displaying linear rates of the convergence in the vicinity of the maximum log-likelihood (Dempster et al., 1977). EM acceleration algorithms, such as the squared iterative method of Varadhan and Roland (2008), can substantially reduce time to convergence. This method applies to any fixed point algorithm and only requires the EM updating function. Our software uses an implementation of the method available in the R package SQUAREM (Varadhan, 2011). In our tests, SQUAREM reduces the time to convergence of our EM algorithm by a factor of 6 without notable loss of robustness.

6 Simulation study

The aim of the simulation study was to examine performance of latent CTMC estimates of disease process functionals from non-exponential sojourn distributions. Data were generated for two state survival and reversible semi-Markov models with Weibull sojourn distributions with increasing (shape=1.5, scale=1) and decreasing (shape=.75, scale=10) hazards. 100 datasets were generated for each of the 3 scenarios (survival with increasing hazard; survival with decreasing hazard; 2-state reversible semi-Markov model with increasing and decreasing sojourn distributions.) With the survival data, death times were observed exactly unless they exceeded 20, in which case they were right censored; the reversible process was observed discretely at times (0,1,..10), jittered by Uniform(-.5,.5) random deviates.

We analyzed the simulated data with 3 latent CTMC models. Models II and III fit survival data with Coxian PH models with 2 and 3 transient states, respectively; Model IV fit discretely observed data from a 2-state reversible model assuming sojourn distributions analogous to model II. All data were fit with our EM using 10 different random starting values per dataset. Hazard and CDFs of sojourn distributions were estimated for each dataset using the corresponding models. We summarized the model performance based on point-wise calculation of bias and root mean squared error (RMSE) of the estimates, as well as the point-wise coverage of 95% confidence intervals based on delta-method standard errors. We limited our analysis to datasets with more than one starting value converging to the putative maximum log-likelihood (481/500=96%). Evaluation of interval estimates based on delta-method standard errors was further limited to datasets with unique MLEs of latent CTMC parameters (449/481=93%).

The means of the point estimates from each model are shown in Figure 2A. The bias of the estimates is summarized in Figure 2B. Coxian PH models have asymptotically constant hazard functions (Aalen, 1995). As such, the estimates of the Weibull(1.5,1) hazard function were increasingly biased for $t > 1$. In contrast, bias of estimates of the Weibull(.75,10) hazard decreased with time, as the hazard flattened out. As anticipated, model II estimates were more biased than those of model III. We expected that model IV estimates would be comparable to model II, since both assume sojourn distributions characterized by 2 latent transient states. This appeared to hold for Weibull(1.5,1) functionals, but not for Weibull(.75,10) functionals; in particular, Model IV was poor at estimating the early portion of the hazard function.

RMSE (Figure 2C) provides a means of assessing the performance of models II and III accounting for bias and variability of estimates. Generally, RMSE of functional estimates were similar for II and

III, except when model II was quite biased (e.g., Weibull(1.5,1) hazard, $t > 2$). The RMSE of model IV estimates was highest overall, reflecting the loss of information due to discrete observations, and in the case of Weibull(.75,.1) functional estimates, increased bias.

Coverage of 95% confidence intervals based on delta-method standard errors are shown in Figure 2D. Poor coverage resulted when point estimates were quite biased (Weibull(1.5,1) hazards for $t > 1.5$), or when the delta-method standard errors under-estimated the true variability of the estimates (Web Appendix Figure 1), as in Weibull(75,10) CDF and hazard functions. Coverage of model IV estimates for small t was also poor for Weibull(1.5,1) functionals at t near 0, which appeared to be due to skewness in the estimates' distributions at this boundary. Nominal coverage was attained when the bias was small and the delta-method standard errors provided good approximations of the true variability of the estimates.

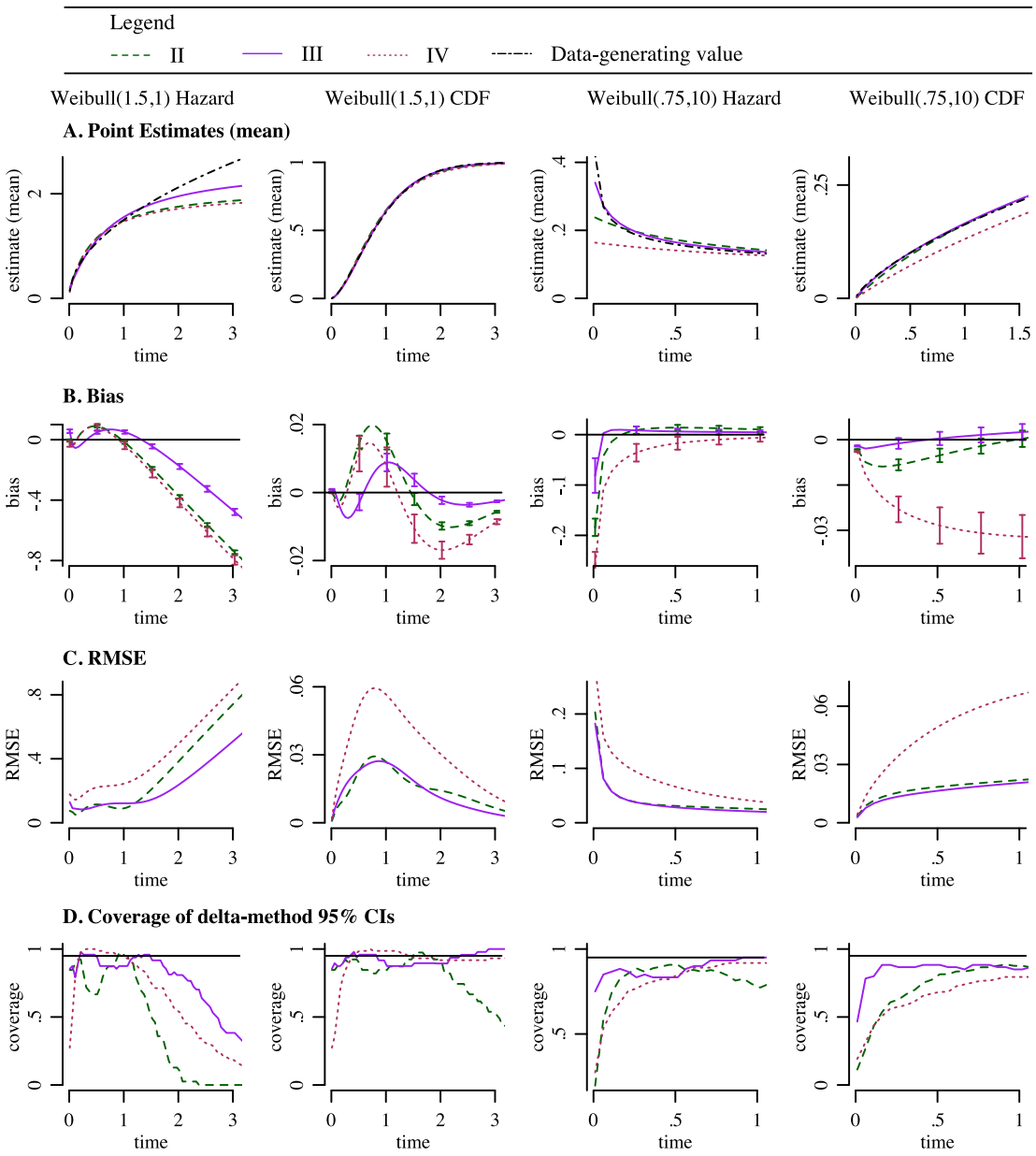

Figure 2: Summary of estimates of CDFs and hazard functions based on models fit to data generated from Weibull(1.5,1) and Weibull(.75,10) sojourn distributions. Models II and III fit survival data with Coxian PH models with 2 and 3 transient states, respectively; Model IV fit discretely observed data from a 2-state reversible model assuming sojourn distributions analogous to model II. A. Mean of point estimates from all models and the data generating value. B. Bias of estimates, with intervals representing Monte Carlo 95% confidence intervals. C. Root mean squared error of estimates. D. Coverage of nominal 95% confidence intervals based on delta-method standard errors.

7 Application

Following lung transplantation, patients are at risk of developing brochiolitis obliterans syndrome, in which bronchioles are irreversibly occluded with scar tissue. Clinically, BOS is diagnosed by >20% reduction in forced expiratory volume/second (FEV1) from post-transplant baseline (Estenne et al., 2002). Titman and Sharples (2010) use an illness-death model to characterize the disease process in a study of heart-lung and double lung transplant patients who had FEV1 monitored 6 months post-transplant and at 9 months, 12 months, and every six months thereafter (Jackson et al., 2002). Our version of the dataset consisted of 122 double lung and 244 heart lung patients. Individuals with only baseline observations were excluded.

The BOS disease process, $W(t)$, has a state space with 3 states: $R = \{1 = "healthy", 2 = "BOS", 3 = "death"\}$, where death is absorbing. The model of Titman and Sharples (2010) assumes that $W(t)$ has an underlying latent CTMC with state space $S = \{1_1, 1_2, 2_1, 2_2, 3\}$ and an intensity matrix $\mathbf{\Lambda}$ implying Coxian phase-type sojourn distributions of $W(t)$. Although the BOS disease process is irreversible, the model includes reversible transitions since they improved model fit. To promote parsimony, the intensity matrix $\mathbf{\Lambda}$ is structured, as $\lambda_{1_2 2_1} = \tau_1 \lambda_{1_1 2_1}$, $\lambda_{2_1 3} = \tau_1 \lambda_{1_1 3}$, $\lambda_{2_2 1_1} = \tau_2 \lambda_{2_1 1_1}$, and $\lambda_{2_2 3} = \tau_2 \lambda_{2_1 3}$. The parameters τ_1 and τ_2 mean that rates of exiting states 1_2 and 2_2 relative to 1_1 and 2_1 change by the same factor regardless of the destination. We expressed this parameterization using log-intensity rates and dummy covariate effects.

The model includes transplant type in the probability of misclassification of healthy patients as diseased, such that $logit(e(Healthy, BOS)) = \gamma_0 + \gamma_1 * Z_{DL}$, where Z_{DL} is an indicator of double lung transplant. Misclassification of diseased patients as healthy does not depend on covariates: $logit(e(BOS, Healthy)) = \nu_0$. Initially, individuals occupy either state 1_1 or 2_1 with a probability depending on transplant type, according to the parameterization $logit(\pi_{2_1}) = B_0 + B_1 * Z_{DL}$.

7.1 Comparison between our EM and other optimization methods

We compared the performance of our EM algorithm (denoted EM1) to a) the EM of Bureau et al. (2003) (EM2), b) the R implementation of Nelder-Mead (NM) (Nelder and Mead, 1965), and c) the box-constrained BFGS optimization algorithms (Byrd et al., 1995). The BFGS constraints assumed all model parameters fell in the interval (-50,8). We implemented the M-step of EM2 with the BFGS stopping criteria based on a relative convergence tolerance of 10^{-3}. We accelerated both EM algorithms by SQUAREM (Varadhan, 2011).

We considered scenarios in which the emission and initial probabilities were unknown or known and fixed at their MLEs. All methods were compared with the same 30 random starting values generated independently from $Normal(0, \sigma^2 = .25)$ distributions. EM convergence was declared when successive iterations of the log-likelihood differed by $< 10^{-6}$ or 200 iterations were taken, whichever came first. NM, and BFGS algorithms were run with the default relative convergence tolerance of "optim" (10^{-8}) and capped likelihood evaluations at 2,500.

The maximum log-likelihood obtained under any method was -1,248.602. Figure 3 shows the algorithm run-time and the maximum attained likelihood for each method and condition, and Table 1 summarizes the convergence results. EM1 was the clear winner in terms of run time, taking a median of 80 seconds to converge when $\boldsymbol{\pi}$ and $\mathbf{E}$ were unknown. Other methods ran between 5.5 to 18 minutes before converging or reaching the maximum iteration limit.

NM had a particularly poor record in converging to the maximum log-likelihood, reaching the iteration limit without converging for 60% of starting values. The other methods were more likely to converge, but not necessarily to the global maxima. With unknown emission and initial distributions, the two EM methods and BFGS converged to local maxima for 40-65% of starting values, though these differences were not statistically significant (chi-square p-value=.12). For these starting values, EM1 was the most stable, in that no starting value resulted in algorithm failure. The other methods failed in 16 out of 120 trials for various reasons depending on the optimization method. In general, exploration of regions of the parameter space corresponding to high rates led to numeric difficulties in calculating transition probabilities.

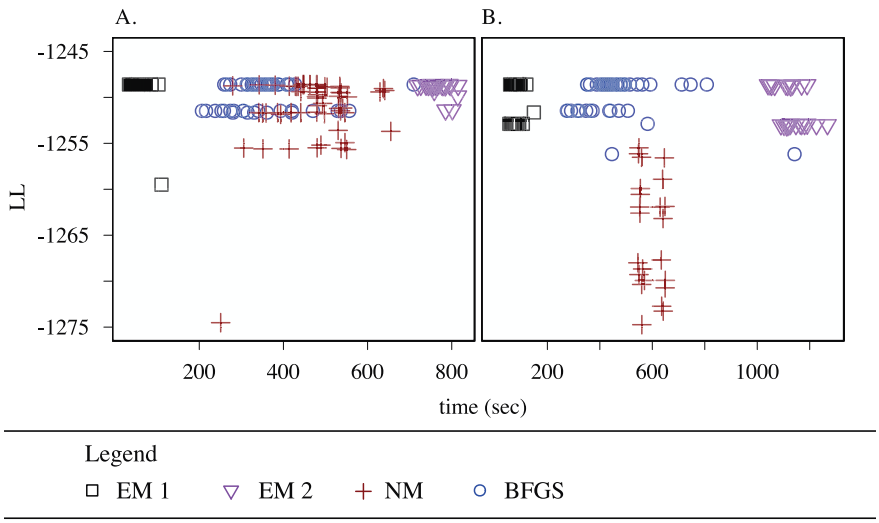

Figure 3: Runtime and attained log-likelihood (LL) when EM1 (our method), EM2, BFGS, and NM algorithms were used to fit the BOS data, using 30 random starting values and assuming either $(\mathbf{E}, \boldsymbol{\pi})$ was fixed (A), or was unknown (B). This figure appears in color in the electronic version of this article.

Table 1: Results of fitting the BOS data using different optimization methods with 30 random starting values.

	$\mathbf{E}, \pi$ fixed				$\mathbf{E}, \pi$ unkown			
	EM1	EM2	NM	BFGS	EM1	EM2	NM	BFGS
Median run-time (s)	60.6	762.4	532.6	337.0	80.3	1125.3	639.5	431.9
Converged to max. LL	30	24	11	13	12	11	0	18
Convergence to local max. or stationary point	0	3	8	10	18	16	4	10
Iteration limit reached	0	0	11	0	0	0	25	0
Algorithm failure	0	3	0	7	0	3	1	2
Total trials	30	30	30	30	30	30	30	30

7.2 BOS results

Due to differences between the dataset we used and that analyzed by Titman and Sharples (2010), model parameter estimates are similar but not identical to that of Titman and Sharples (2010). Both sets of MLEs were evidently unique, based on numeric investigations with different starting values. Estimates and 95% confidence intervals for the rate, emission, and intensity parameters on their original scales (i.e., rates, emission and initial probabilities) are shown in Web Appendix Table 2.

The first passage distribution for BOS development, depicted in Figure 4A, shows that about 30% of those starting in healthy state 1_1 will have transitioned into the disease state after 1 year and over 60% after 4 years. The rate of entry into the diseased states declines with time since transplant; disease rates are initially 35-40% and drop to 15% per year after 5 years (Figure 4B). The rates of returning to the healthy state from the BOS state also decline with time spent in the BOS state. Among those in BOS state 2_1 at time t_0, the rate of reversion to the heathy state 1_1 is initially 6%, but drops to 1.6% after a year.

Functionals of the disease process are shown in Figure 4 along with delta-method based confidence intervals. The cumulative probabilities of death conditional on starting in healthy state 1_1 versus BOS state 2_1 at t_0, is shown in Figure 4C. By 2 years post diagnosis, 15% of those initially healthy will have died. For individuals in BOS state 2_1 at time t_0, nearly 30% have died by one year, and 65% by five years. Mortality rates in individuals in healthy state 1_1 at t_0 are 9% per year at 1 year and increase slowly thereafter (Figure 4D). The increase in mortality appears solely attributable to those who transition to BOS status because mortality rates prior to BOS remain close to zero. After transitioning to BOS state 2_1, mortality rates jump dramatically ($> 50\%$ per year), and then drop to 20% after one year. The very high initial mortality rates are probably attributable to individuals with rapidly progressing versions of the disease.

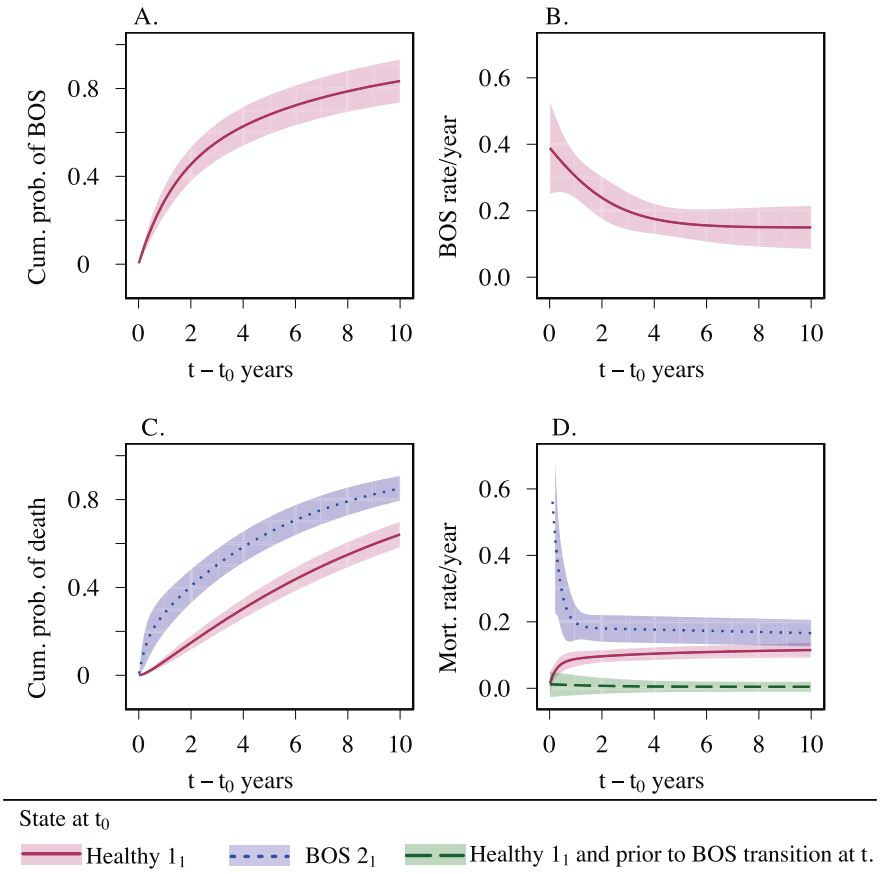

Figure 4: A. Cumulative probability of having transitioned to BOS state at least once, conditional on being in 1_1 at t_0. B. Disease rate conditional on being in healthy state 1_1 at t_0. C. Cumulative probability of death. C. Mortality rate per year, as a function of state at t_0. In all figures the shaded regions represent 95% point-wise confidence intervals for the estimates. This figure appears in color in the electronic version of this article.

8 Discussion

Multistate disease processes observed in the panel data setting pose challenges for analysis. The widely-used approach of assuming standard CTMCs leads to models that are unrealistic for processes with duration-dependent sojourn distributions. Assuming a latent CTMC framework accommodates duration-dependent sojourn distributions but yields tractable likelihoods. These models also offer interpretative advantages, as functionals describing the process are computable analytically.

Our EM algorithm provides an efficient and robust method of obtaining MLEs and standard errors of latent parameter estimates. The method considerably outperformed other optimization approaches, including those implemented in the R package msm (Jackson, 2011), Nelder-Mead, and BFGS. Our method also performed favorably relative to the EM of Bureau et al. (2003), despite our complete data space yielding a

higher fraction of missing information. Another alternative optimization method is implementing Newton-Raphson on the observed data likelihood (Lystig and Hughes, 2002). However, each Newton-Raphson step requires calculation of the observed data score and information matrix, necessitating computations similar to first and second moments of complete data sufficient statistics. On balance, it is likely the relatively computationally expensive information calculation outweighs a faster rate of convergence in the Newton-Raphson method relative to our EM algorithm.

The utility of latent CTMC models lies in their ability to approximate functionals of disease processes from generic sojourn time distributions. Our simulation studies focused on sojourn distributions with increasing or decreasing hazard functions, but latent CTMCa can also approximate non-monotonic hazards (Aalen, 1995). The quality of approximation depends on the number of latent states in the model, the data generating scenario, and time. Since latent CTMC models imply that sojourn time distributions have asymptotically constant hazard functions, substantial bias can result if the data-generating sojourn distribution has an increasing hazard. In these cases, caution should be applied to interpretation of hazard latent CTMC estimates outside of near-range time points.

Latent CTMC models appear to offer particular advantages for discretely observed reversible disease processes. However, performance of panel data estimates of disease process functionals differed from fully observed counterparts, based on models assuming the same number of latent states per disease state. Not surprisingly, estimates based on panel observations were much more variable. We also observed relatively more bias in panel data estimates of hazard and CDFs from the decreasing but not increasing hazard sojourn distributions. We suspect this bias would be reduced if observations were initially more frequent. Those designing panel studies would be well served to investigate variance and bias with their own simulations prior to committing to an observation schedule. This is especially important given that observations that are too distantly spaced may limit the estimability of latent model parameters (Bladt and Sorensen, 2005).

We also investigated the frequentist properties of interval estimates of functionals based on delta-method standard errors. Delta-method standard errors on average represented 92% of the true variability of the estimates, but performance varied by model, functional, time, and data generating distribution. Models with more latent states generally yielded better delta-method standard error estimates, which, along with reduced estimate bias, led to better coverage properties of nominal 95% confidence intervals. Given that latent CTMC models approximate generic data-generating distributions, robust variance estimates in the EM algorithm context may yield more valid standard errors (Elashoff and Ryan, 2004). Ultimately, the validity of delta-method standard errors assumes the uniqueness of latent parameter MLEs. In their absence, we recommend applying a non-parametric bootstrap. The computation time required would not be prohibitive given the increased efficiency of our fitting algorithm.

In the frequentist setting, the major weakness of latent CTMC models is that latent model parameters are potentially non-identifiable. Despite their relatively parsimonious representation, Coxian PH models of a given dimension may have multiple intensity matrices that imply the same sojourn time distribution (He and Zhang, 2007). HMMs based on discretely observed CTCMs with measurement error also may not be fully identifiable (Rosychuk and Thompson, 2004). Non-identifiable latent CTMC parameters will often still have unique disease process functional estimates. However, lack of identifiability not only affects standard error estimates of functionals but also complicates model selection, including choice of the number of latent states. Titman and Sharples (2010) describe a likelihood ratio test of exponential sojourn distributions that requires special adjustments for non-identifiable $\mathbf{\Lambda}$ parameters under the null hypothesis. Use of the Akaike and Bayesian information criteria to compare non-nested models requires that we know the number of estimable

model parameters. The increased efficiency of our fitting algorithm suggests that it may be practical to evaluate models using k-fold cross validation with a goodness of fit statistic measuring prediction error (Titman and Sharples, 2008).

Our focus has been on frequentist estimation. Bayesian methods also have a strong appeal in this setting (Bladt et al., 2003). Sensible priors may yield identifiable latent parameters, and posterior distributions provide uncertainty estimates for model functionals. Further, model selection may be possible using reversible jump MCMC (Green, 1995). McGrory et al. (2009) have implemented Bayesian model selection for PH models of length of hospital stay, and their approach might be scaled to apply to more general latent CTMC models.

9 Conclusion

Latent CTMCs provide a flexible means of describing discretely observed multistate disease processes with duration-dependent sojourn distributions. They have especial value for discretely observed processes with reversible transitions, for which few compelling analysis approaches exist. Our EM method provides an efficient and robust way to fit these models in a frequentist setting. We hope our results will encourage the wider use of latent CTMCs in the analysis of clinical studies.

Acknowledgements

The authors thank Andrew Titman, Sharples Linda, and Tsui Steven for their help in getting access to the BOS data from the Papworth Hospital, UK. Papworth Hospital data has been used to illustrate the statistical methods and that these should not be used in isolation to inform clinical practice. JML was supported by NIH grant No. T32 CA009168. VNM was supported by the NSF grant No. DMS-0856099. We thank Lurdes Inoue and Kenneth Lange for their comments on the manuscript.

References

Aalen, O. O. (1995). Phase type distributions in survival analysis. *Scandinavian Journal of Statistics*, 22(4):447–463.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115.

Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, 31:1074–1088.

Asmussen, S., Nerman, O., and Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441.

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1):164–171.

Bladt, M., Esparza, L., and Nielsen, B. (2011). Fisher information and statistical inference for phase-type distributions. *Journal of Applied Probability*, 48:277–293.

Bladt, M., Gonzalez, A., and Lauritzen, S. L. (2003). The estimation of phase-type related functionals using Markov chain Monte Carlo methods. *Scandinavian Actuarial Journal*, 2003(4):280–300.

Bladt, M. and Sorensen, M. (2005). Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):395–410.

Bureau, A., Shiboski, S., and Hughes, J. P. (2003). Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine*, 22(3):441–62.

Byrd, R., Lu, P., Noecedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing*, 16:1190–1208.

Cappe, O., Moulines, E., and Ryden, T. (2005). *Statistical Inference for Hidden Markov Models*. Springer, New York.

Crespi, C. M., Cumberland, W. G., and Blower, S. (2005). A queueing model for chronic recurrent conditions under panel observation. *Biometrics*, 61:194–199.

Cumani, A. (1982). On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectronics and Reliability*, 22(3):583–602.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Elashoff, M. and Ryan, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, 13(1):48–65.

Estenne, M., Maurer, J. R., Boehler, A., Egan, J. J., Frost, A., Hertz, M., Mallory, G. B., Snell, G. I., and Yousem, S. (2002). Bronchiolitis obliterans syndrome 2001: an update of the diagnostic criteria. *The Journal of Heart and Lung Transplantation*, 21(3):297–310.

Faddy, M. (1998). On inferring the number of phases in a Coxian phase-type distribution. *Communications in Statistics*, 14:407–417.

Foucher, Y., Giral, M., Soulillou, J. P., and Daures, J. P. (2007). A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. *Statistics in Medicine*, 26:5381–5393.

Francois, R., Eddelbuettel, D., and Bates, D. (2011). *RcppArmadillo: Rcpp integration for Armadillo templated linear algebra library*. R package version 0.2.34.

Gentleman, R. (1994). Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV Disease. *Statistics in Medicine*, 13:805–822.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Guihenneuc-Jouyaux, C., Richardson, S., and Longini, I. M. (2000). Modeling markers of disease progression process by a hidden Markov process: application to CD4 cell decline. *Biometrics*, 56(3):733–741.

He, Q.-M. and Zhang, H. (2007). An algorithm for computing minimal Coxian representations. *INFORMS Journal on Computing*, 20(2):179–190.

Hobolth, A. and Jensen, J. (2011). Summary statistics for endpoint-conditioned continuous-time Markov chains. *Journal of Applied Probability*, 48:911–924.

Hobolth, A. and Jensen, J. L. (2005). Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical Applications in Genetics & Molecular Biology*, 4(1):1–22.

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, 38(8):1–29.

Jackson, C. H., Sharples, L. D., McNeil, K., Stewart, S., and Wallwork, J. (2002). Acute and chronic onset of bronchiolitis obliterans syndrome (BOS): are they different entities? *The Journal of Heart and Lung Transplantation*, 21(6):658–66.

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

Kang, M. and Lagakos, S. W. (2007). Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics*, 8(2):252–64.

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 57(2):425–437.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, 44(2):226–233.

Lystig, T. C. and Hughes, J. P. (2002). Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics*, 11(3):678–689.

Mandel, M. (2010). Estimating disease progression using panel data. *Biostatistics*, 11(2):304–16.

Marshall, A. H. and Zenga, M. (2010). Experimenting with Coxian phase-type distributions to uncover suitable fits. *Methodology in Computational Applied Probability*, 14(1):71–86.

McGrory, C. A., Pettitt, A. N., and Faddy, M. (2009). A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. *Computational Statistics & Data Analysis*, 53(12):4311–4321.

Minin, V. N. and Suchard, M. a. (2008a). Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology*, 56(3):391–412.

Minin, V. N. and Suchard, M. a. (2008b). Fast, accurate and simulation-free stochastic mapping. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, 363(1512):3985–95.

Nelder, J. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, 7:308–313.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Roberts, W. and Ephraim, Y. (2008). An EM algorithm for ion-channel current estimation. *Signal Processing, IEEE Transactions on*, 56(1):26 –33.

Rosychuk, R. and Thompson, M. (2004). Parameter identifiability issues in a latent Markov model for misclassified binary responses. *Journal of the Iranian Statistical Society*, 3:39–57.

Titman, A. and Sharples, L. (2008). A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine*, 27:2177–2195.

Titman, A. C. and Sharples, L. D. (2010). Semi-Markov models with phase-type sojourn distributions. *Biometrics*, 66(3):742–52.

Varadhan, R. (2011). SQUAREM: Squared extrapolation methods for accelerating fixed-point iterations. R package version 2010.12-1.

Varadhan, R. and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353.

Appendix for
Fitting and Interpreting Continuous-Time Latent Markov Models for Panel Data

by Jane M. Lange and Vladimir N. Minin

Appendix A: Complete data score and Hessian

Note: All vectors are assumed to be column vectors unless otherwise noted.

CTMC parameters

The CTMC log-likelihood component is in the curved exponential family, with natural parameters $\log(\lambda_{ij})$ and $\sum_{i \neq j} \lambda_{ij}$ corresponding to sufficient statistics $n_T(i,j)$ and $d_T(i)$. Individual level baseline covariates $\mathbf{w}^h$ are added via $\log(\lambda_{ij}^h) = \boldsymbol{\beta}_{ij}^T \mathbf{w}^h$, where h denotes the individual and $\mathbf{w}^h$ and $\boldsymbol{\beta}_{ij}$ are p-dimensional vectors corresponding to p covariates. For convenience, we list the intensity parameters $\{log(\lambda_{ij}) : i, j \in S; i \neq j\}$ as a q-dimensional vector $\boldsymbol{\psi}$, indexing each i, j pair in $\boldsymbol{\psi}$ by u. This allows us to derive the score and information for all intensity parameters simultaneously, which is particularly useful if one assumes the same covariate effect for more than one transition intensity. Using the notation $i[u]$ and $j[u]$ to yield the i and j corresponding to u, the uth entry of the vector score function for $\boldsymbol{\psi}$ is

$$\dot{l}(\boldsymbol{\psi})[u] = n_T\left(i[u], j[u]\right) - d_T\left(i[u]\right) exp\left(\psi[u]\right).$$

The Hessian matrix for $\boldsymbol{\psi}$ is diagonal with non-zero entries

$$\ddot{l}(\boldsymbol{\psi})[u,u] = -d_T\left(i[u]\right) exp\left(\psi[u]\right).$$

The score function when the rate matrix is parameterized with covariates $\mathbf{w}$ is given by

$$\dot{l}(\boldsymbol{\beta}|\mathbf{w}^h) = \nabla\boldsymbol{\psi}(\boldsymbol{\beta})^T \dot{l}\{\boldsymbol{\psi}(\boldsymbol{\beta})\},$$

where $\nabla\boldsymbol{\psi}(\boldsymbol{\beta})^T$ is the $p \times q$ matrix whose m, u entry corresponds to $\frac{\partial\psi[u]}{\partial\beta[m]}$. The Hessian matrix in the presence of covariates is

$$\ddot{l}(\boldsymbol{\beta}|\mathbf{w}^h)[j,m] = -\sum_{u=1}^{q} \frac{\partial\psi[u]}{\partial\beta[j]} \frac{\partial\psi[u]}{\partial\beta[m]} d_T\left(i[u]\right) exp\left(\psi[u]\right).$$

In matrix form, this can be written as

$$\ddot{l}(\boldsymbol{\beta}|\mathbf{w}^h) = \nabla\boldsymbol{\psi}(\boldsymbol{\beta})^T (\nabla\boldsymbol{\psi}(\boldsymbol{\beta})) \circ \mathbf{D},$$

where $\mathbf{D}$ is a $q \times p$ matrix with each column consisting of column vector $\mathbf{v}$, such that entries $v[u] = -exp(\psi[u]) d_T\left(i[u]\right)$, and $\circ$ refers to the Hadamard (element-wise) product. Both the score and Hessian are additive across subjects, so the total score and Hessian are obtained by summing over corresponding subject-specific quantities.

Initial and Emission distributions parameters

We limit our attention to the score and Hessian for the emission distribution, as the initial distribution is analogous. For a single subject,
$\boldsymbol{O}_T(i) = \{O_T(i,1), \ldots, O_T(i,r)\} \sim Multinomial\{\mathbf{e}_i, N(i)\}$, where $N(i) = \sum_{j=1}^{r} O_T(i,j)$ and $\boldsymbol{e}_i = \{e(i,1), \ldots, e(i,r)\}$. Sufficient statistics include the $r-1$ length vector
$\mathbf{o_{i[-1]}} = \{o_T(i,2), ..., o_T(i,r)\}$. The natural parameters are $\left\{\eta_{ij} = \log\left(\frac{e(i,j)}{e(i,1)}\right) : j = 2, ..., r\right\}$. In the absence of covariates, the score function for the parameters $\boldsymbol{\eta}_i = (\eta_{i2}, ..., \eta_{ir})$ is

$$\dot{l}(\boldsymbol{\eta}_i) = \mathbf{o_{i[-1]}} - N(i)\mathbf{e_{i[-1]}},$$

where $\mathbf{e}_{i[-1]} = \{e(i,2), ...e(i,r)\}$ is a $r-1$ length vector of emission probabilities written in terms of $\boldsymbol{\eta}_i$.

Subject-level covariates $\mathbf{w}_i^h$ are added to the model via $\eta_{ij}^h = \boldsymbol{\gamma}_{ij}^T \mathbf{w}_i^h$, where h indexes the individual. Let $\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_{i2}, \ldots, \boldsymbol{\gamma}_{ir})$ be the vector of all p covariate parameters. The score is

$$\dot{l}(\boldsymbol{\gamma}_i|\mathbf{w}^h) = \nabla\boldsymbol{\eta}_i(\boldsymbol{\gamma}_i)^T\{\mathbf{o_{i[-1]}} - N_i\boldsymbol{e}_{i[-1])}\},$$

where $\nabla\boldsymbol{\eta}_i(\boldsymbol{\gamma}_i)^T$ is the $p \times (r-1)$ matrix of partial derivatives of $\boldsymbol{\eta}_i$ with respect to $\boldsymbol{\gamma}_i$ and $\boldsymbol{e}_{i[-1])}$ is written in terms of $\boldsymbol{\gamma}_i$. The Hessian matrix in the absence of covariates is given by

$$\ddot{l}(\boldsymbol{\eta}_i) = -\operatorname{Cov}(\mathbf{o_{i[-1]}}).$$

With covariates, the Hessian matrix is given by

$$\ddot{l}(\boldsymbol{\gamma}_i|\mathbf{w}^h) = -\left\{\nabla\boldsymbol{\eta}_i(\boldsymbol{\gamma}_i)^T \operatorname{Cov}(\mathbf{o_{i[-1]}})\nabla(\boldsymbol{\eta}_i(\boldsymbol{\gamma}_i)\right\}.$$

As before, the total score and Hessian are obtained by summing over the corresponding subject-specific quantitites.

Appendix B: Recursions for hidden Markov models

Throughout, we abbreviate $x_1, \ldots, x_k$ by $\mathbf{x}_{1:k}$ and $o_1, \ldots o_k$ by $\mathbf{o}_{1:k}$.

Forward and backward probabilities

Forward probabilities are defined as $\alpha_k(u) = \mathrm{P}(\mathbf{o}_{1:k}, X_k = u)$ and backward probabilities as $\beta_k(u) = \mathrm{P}(\mathbf{o}_{k+1:n}|X_k = u)$. When the last time coincides with the time of absorption, Y, the forward and backward probabilities are defined as before, with the exception that $\beta_k(u) = \frac{\partial}{\partial y}\mathrm{P}(\mathbf{o}_{k+1:n}, Y < y|X_k = u)$ and $\alpha_n(u) = \frac{\partial}{\partial y}\mathrm{P}(\mathbf{o}_{k+1:n}, Y < y)$, Forward and backward probabilities are calculated through recursive formulae of **?**.

Filtering and conditional likelihood calculations

Filtering probabilities, $\mathrm{P}(X_k = j|\mathbf{o}_{1:k})$ and the conditional observed data likelihood $\mathrm{P}(O_k = o_k|\mathbf{o}_{1:k-1})$ are related to modified forward probabilities, $a_k(j) = \mathrm{P}(X_k = j, O_k = o_k|\mathbf{o}_{1:k-1})$. That is, $\mathrm{P}(O_k = o_k|\mathbf{o}_{1:k-1}) = \sum_{j\in S} a_k(j)$, and $\mathrm{P}(X_k = j|\mathbf{o}_{1:k}) = \frac{a_k(j)}{\sum_{l\in S} a_k(l)}$. The modified forward probabilities can be calculated recursively. Initialize

$$a_1(j) = \mathrm{P}(O_1 = o_1, X_1 = x_1) = e(x_1, o_1)\pi(x_1),$$

and the recursion is

$$a_{k+1}(j) = \sum_{i\in S} \frac{a_k(i)}{\sum_l a_k(l)} e(x_{k+1}, o_{k+1})\, \mathrm{P}_{ij}(t_{k+1} - t_k).$$

Recursive smoothing for first moments of complete data sufficient statistics

First moment calculations define entries of $s_k(x_k, x_{k+1})$ as values of complete data sufficient statistics (section 2.4.1) on the interval $T_k = [t_k, t_{k+1}]$, conditional on x_k and x_{k+1}. Thus, $s_k(x_k, x_{k+1})$ is defined as $\mathrm{E}[d_{T_k}|X_k = x_k, X_{k+1} = x_{k+1}]$ for entries corresponding to $d_T(i)$; as $\mathrm{E}[n_{T_k}(i,j)|X_k = x_k, X_{k+1} = x_{k+1}]$ for $n_T(i,j)$; 0 for z_i; and as $I(X_{k+1} = i, O_{k+1} = j)$ for $o_T(i,j)$. Initial values for the function $t_k(\mathbf{x}_{1:k})$ are set at $t_1(x_1) = 0$ for entries corresponding to $d_T(i)$ and $n_T(i,j)$; $I(X_1 = i)$ for z_i; and $I(X_1 = i, O_1 = j)$ for $o_T(i,j)$.

Recursive smoothing for second moments of complete data sufficient statistics

The recursive smoothing method to obtain second and cross moments of complete data sufficient statistics conditional on the entirety of a subject's observed data, $\mathbf{o}$, proceeds with a similar framework and terminology as for first moments (Section 3.2.3.) First, we recursively define a functional that corresponds to $\mathrm{E}[\mathbf{S}[t_1, t_k]\mathbf{S}[t_1, t_k]^T|\mathbf{x}_{1:k}]$, the second moments of complete sufficient statistics on the interval $[t_1, t_k]$, conditional on $\mathbf{x}_{1:k}$. Next, we define the recursive updates of the auxiliary function, $\tau_k(x_k)$. Finally, we compute the auxiliary function updates for $t_1, \ldots, t_n$, enabling us to calculate the target quantity $\mathrm{E}[\mathbf{S}[t_1, t_n]\mathbf{S}[t_1, t_n]^T|\mathbf{o}_{1:n}]$.

The recursive definition of $\mathrm{E}[\mathbf{S}[t_1, t_{k+1}]\mathbf{S}[t_1, t_{k+1}]^T|\mathbf{x}_{1:k+1}]$ involves not only $\mathrm{E}[\mathbf{S}[t_1, t_k]\mathbf{S}[t_1, t_k]^T|\mathbf{x}_{1:k}]$ but also the first moment, $\mathrm{E}[\mathbf{S}[t_1, t_k]|\mathbf{x}_{1:k}]$. Thus it makes sense to consider jointly the first and second moments of complete data sufficient statistics conditional on $\boldsymbol{x}_{1:k}$. We define the joint recursive function of latent states as

$$\mathbf{t}(\mathbf{x}_{1:k+1}) = \left\{ t^{(1)}(\mathbf{x}_{1:k+1}), t^{(2)}(\mathbf{x}_{1:k+1}) \right\},$$

where

$$\begin{aligned} t^{(1)}(\mathbf{x}_{1:k+1}) &= \mathrm{E}[\mathbf{S}[t_1, t_{k+1}]|\mathbf{x}_{1:k+1}] \\ &= t^{(1)}(\mathbf{x}_{1:k}) + \mathrm{E}[\mathbf{S}[t_k, t_{k+1}]|x_k, x_{k+1}] \end{aligned}$$

and

$$\begin{aligned} t^{(2)}(\mathbf{x}_{1:k+1}) &= \mathrm{E}[\mathbf{S}[t_1, t_{k+1}]\mathbf{S}[t_1, t_{k+1}]^T|\mathbf{x}_{1:k+1}] \\ &= t^{(2)}(\mathbf{x}_{1:k}) + \mathrm{E}[\mathbf{S}[t_k, t_{k+1}]|x_k, x_{k+1}]t^{(1)}(\mathbf{x}_{1:k})^T + t^{(1)}(\mathbf{x}_{1:k})\,\mathrm{E}[\mathbf{S}[t_k, t_{k+1}]|x_k, x_{k+1}]^T \\ &\quad + \mathrm{E}[\mathbf{S}[t_k, t_{k+1}]\mathbf{S}[t_k, t_{k+1}]^T|x_k, x_{k+1}]. \end{aligned}$$

The first component is identical to first moment recursive function (eq. (3) in the main text); the second corresponds to second and cross moments of complete data sufficient statistics conditional on latent states $\mathbf{x}_{1:k}$. The calculation of $t^{(2)}(\mathbf{x}_{1:k+1})$ follows from the conditional independence of $\mathbf{S}[t_l, t_{l+1}]$ and $\mathbf{S}[t_j, t_{j+1}]$ given the endpoints $x_l, x_{l+1}, x_j, x_{j+1}$ and the fact that $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$ if X and Y are independent. We assign the function

$$\begin{aligned} \mathbf{s}_k(x_k, x_{k+1}) &= \left\{ \boldsymbol{s}_k^{(1)}(x_k, x_{k+1}), \boldsymbol{s}_k^{(2)}(x_k, x_{k+1}) \right\} \\ &= \left\{ E\left[\mathbf{S}[t_k, t_{k+1}]|x_k, x_{k+1}\right], \mathrm{E}\left[\mathbf{S}[t_k, t_{k+1}]\mathbf{S}[t_k, t_{k+1}]^T|x_k, x_{k+1}\right] \right\}. \end{aligned}$$

The specific values of $t_1^{(1)}(x_1)$ and $s_k^{(1)}(x_k, x_{k+1})$ for latent CTMC sufficient statistics were provided previously. Appendix Table 1 summarizes specific details of $s_k^{(2)}(x_k, x_{k+1})$ and $t_1^{(2)}(x_1)$ for all pairs of latent CTMC complete data sufficient statistics.

The auxiliary functions likewise have two components corresponding to first and second moments: $\boldsymbol{\tau}(x_k) = \left\{\tau^{(1)}(x_k), \tau^{(2)}(x_k)\right\}$. The updates for $\boldsymbol{\tau}(x_k)$ follow from a multivariate version of eq. (4) in the main text. The $\tau^{(1)}(x_k)$ component is defined as in eq. (4). The $\tau^{(2)}(x_k)$ component is defined recursively as

$$\begin{aligned} \tau_{k+1}^{(2)}(x_{k+1}) = \mathrm{P}(o_{k+1}|\mathbf{o}_{1:k})^{-1} \Bigg\{ \sum_{x_k} [\tau^{(2)}(x_k) + \tau_k^{(1)}(x_k) s_k^{(1)}(x_k, x_{k+1})^T \\ + s_k^{(1)}(x_k, x_{k+1})\tau_k^{(1)}(x_k)^T + \mathrm{P}(X_k = x_k|\mathbf{o}_{1:k})\,\mathrm{E}[\mathbf{S}[t_k, t_{k+1}]\mathbf{S}[t_k, t_{k+1}]^T|x_k, x_{k+1}]] \\ \times e(x_{k+1}, o_{k+1})\,\mathrm{P}_{x_k x_{k+1}}(t_{k+1} - t_k) \Bigg\}. \end{aligned}$$

The final recursion allows us to calculate $\mathrm{E}[t_n^{(2)}(\mathbf{x}_{1:n})|\mathbf{o}_{1:k}] = \sum_{x_n \in X} \tau_n^{(2)}(x_n)$, giving us the expected value of second moments of complete data sufficient statistics conditional on the observed data.

Table 1: Definition of $s_k^{(2)}(x_k, x_{k+1})$ and $t_1^{(2)}(x_1)$ for second moment calculations.

Statistics	$s^{(2)}(x_k, x_{k+1})$	$t_1^{(2)}(x_1)$
$d_T(i), d_T(j)$	$\mathrm{E}[d_{T_k}(i)d_{T_k}(j)\|x_k, x_{k+1}]$	0
$d_T(i), n_T(j,m)$	$\mathrm{E}[d_{T_k}(i)n_{T_k}(j,m)\|x_k, x_{k+1}]$	0
$d_T(i), o_T(j,m)$	$\mathrm{E}[d_{T_k}(i)I(X_{k+1} = j, O_{k+1} = m)\|x_k, x_{k+1}]$	0
$n_T(i,l), n_T(j,m)$	$\mathrm{E}[n_{T_k}(i,l)n_{T_k}(j,m)\|x_k, x_{k+1}]$	0
$o_T(j,m), o_T(l,r)$	$I(X_{k+1} = j, O_{k+1} = m, X_{k+1} = l, O_{k+1} = r)$	$I(X_1 = j, O_1 = m, X_1 = l, O_1 = r)$
$n_T(i,l), o_T(l,r)$	$\mathrm{E}[n_{T_k}(i,l)I(X_{k+1} = l, O_{k+1} = r)\|x_k, x_{k+1}]$	
z_i, z_m	0	$I(X_1 = i)I(X_1 = m)$
$z_i, o_T(j,m)$	0	$I(X_1 = i)I(X_1 = j, O_1 = m)$
$n_T(j,m), z_i$	0	0
$d_T(j), z_i$	0	0

Appendix C: Differentiated joint moments of transitions and state occupancy durations with known absorption times

We assume that the CTMC has one absorbing state g. Differentiated joint moments in the presence of known absorption times rely on the fact that if an individual is absorbed at time t, transitions to g occur only once and no time is spent in g. These joint moments formulae use the joint moments defined in Section 3.2.1, which we refer to as $M_{ij}(t)[a,b] = \mathrm{E}[n_t(i,j)I(X_0 = a)|X_t = b]$; $H_i(t)[a,b] = \mathrm{E}[d_t(i)I(X_t = b)|X_0 = a]$; $U_{ijlm}(t)[a,c] = \mathrm{E}[n_t(i,j)n_t(l,m)I(X_t = c)|X_0 = a]$; $W_{ij}(t)[a,c] = \mathrm{E}[d_t(i)d_t(j)I(X_t = c)|X_0 = a]$; and $V_{ilm}(t)[a,c] = \mathrm{E}[d_t(i)n_t(l,m))I(X_t = c)|X_0 = a]$.

When the complete-data statistic of interest is $S = d_t(i)$, the differentiated joint moment is given by

$$\frac{\partial}{\partial y}\,\mathrm{E}[d_t(i)I(Y < t)|X_0 = a] = I(i \neq g)\sum_{c\neq g} H_i(t)[a,c]\lambda_{cg}.$$

When $S = d_t(i)d_t(j)$, the differentiated joint expectation is identical, except $I(i \neq g)$ is replaced by $I(i,j \neq g)$, and $H_i(t)[a,c]$ is replaced by the duration cross moment $W_{ij}(t)[a,c]$.

For $S = n_t(i,j)$, the differentiated joint expectation is

$$\frac{\partial}{\partial y}\,\mathrm{E}[n_t(i,j)I(Y < y)|X_0 = a] = I(i,j \neq k)\sum_{c\neq k} M_{ij}(t)[a,c]\lambda_{ck} + I(i \neq k, j = k)\,\mathrm{P}_{ai}(t)\lambda_{ik}.$$

For $S = n_t(i,j)n_t(l,m)$ the differentiated joint expectation is given by

$$\begin{aligned}\frac{\partial}{\partial y}\,\mathrm{E}[n_t(i,j)n_t(l,m)I(Y < y)|X_0 = a] &= I(i,j,l,m \neq g)\sum_{c\neq g} U_{ijlm}(t)[a,c]\lambda_{cg}\\ &+ I(i,l,m \neq g, j = g)M_{lm}(t)[a,i]\lambda_{ig} + I(i,j,l \neq g, m = g)M_{ij}(t)[a,l]\lambda_{lg}\\ &+ I(i,l \neq g, i = l, j = m = g)\,\mathrm{P}_{ai}(t)\lambda_{ig}.\end{aligned}$$

For $S = n_t(l,m)d_t(i)$, the differentiated joint expectation is given by

$$\frac{\partial}{\partial y}\,\mathrm{E}[d_t(i)n_t(l,m)I(Y < y)|X_0 = a] = I(i,j,l,\neq g)\sum_{c\neq g} V_{ilm}(t)[a,c]\lambda_{cg} + I(i,l \neq g, m = g)H_i(t)[a,l]\lambda_{lm}.$$

Appendix D: Delta method standard errors of disease process functionals

Suppose $\boldsymbol{\psi}$ is a $p\times 1$ vector of latent model parameters with MLE $\hat{\boldsymbol{\psi}}$, and $F(\boldsymbol{\psi},t)$ is a one-dimensional functional. Let $\nabla F(\hat{\boldsymbol{\psi}},t)$ be the $p \times 1$ gradient of $F(\boldsymbol{\psi},t)$ with respect to $\boldsymbol{\psi}$ evaluated at $\hat{\boldsymbol{\psi}}$. The asymptotic distribution of the functional estimates $F(\hat{\boldsymbol{\psi}},t)$ is normal with mean $F(\boldsymbol{\psi},t)$ and an approximate covariance matrix given by

$$\mathrm{Cov}(F(\hat{\boldsymbol{\psi}},t)) = \nabla F(\hat{\boldsymbol{\psi}},t)^T\,\mathrm{Cov}(\hat{\boldsymbol{\psi}},t)\nabla F(\hat{\boldsymbol{\psi}},t).$$

Functionals such as CDFs, hazard functions, and transition probabilities involve the matrix exponential; thus we require the derivative of $\exp(\boldsymbol{\Lambda}(\boldsymbol{\psi})t)$ with respect to entries of $\boldsymbol{\psi}$. These derivative involve similar integrals as first moments of occupancy durations and transition counts (Section 3.2.1) and are computed with similar methods(**?**). For example, consider the functional $\mathrm{P}_{ij}(t, \boldsymbol{\psi}) = \exp(\boldsymbol{\Lambda}(\boldsymbol{\psi})t)$. Then $\frac{\partial \mathrm{P}_{ij}(t,\boldsymbol{\psi})}{\partial \psi[k]}$ is the i, j entry of the matrix given by $\int_0^t e^{\boldsymbol{\Lambda}(\boldsymbol{\psi})\tau} \boldsymbol{B}_{\psi[k]} e^{\boldsymbol{\Lambda}(\boldsymbol{\psi})(t-\tau)} d\tau$, where $\boldsymbol{B}_{\psi[k]} = \{B_{\psi[k]}(i,j)\}$ and $B_{\psi[k]}(i,j) = \frac{\partial \lambda_{ij}(\boldsymbol{\psi})}{\partial_{\psi[k]}}$.

References

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

Najfeld, I. and Havel, T. F. (1994). Derivatives of the matrix exponential and their computation. Advances in Applied Mathematics **16,** 321–375.

Table 2: Maximum likelihood estimates of BOS model intensity rates, emission probabilities, and initial probabilities.

	Transition				
Intensity rates	i	j	Point estimate	95% CI	
	1_1	1_2	0.39	0.11	1.42
	1_1	2_1	0.39	0.27	0.56
	1_1	3	0.01	0	0.29
	1_2	2_1	0.14	0.09	0.23
	1_2	3	0.004	0.00017	0.11
	2_1	1_1	0.06	0.01	0.31
	2_1	2_2	3.12	0.97	9.99
	2_1	3	0.73	0.27	1.94
	2_2	1_1	0.02	0.004	0.06
	2_2	3	0.19	0.15	0.23
Emission	e(Healthy,BOS)	Double lung	0.076	0.042	0.133
		Heart lung	0.018	0.01	0.031
	e(BOS,Healthy)		0.011	0.004	0.028
Initial Distribution	$\pi(BOS_1)$	Heart-lung	0.061	0.035	0.103
		Double lung	0.043	0.014	0.124

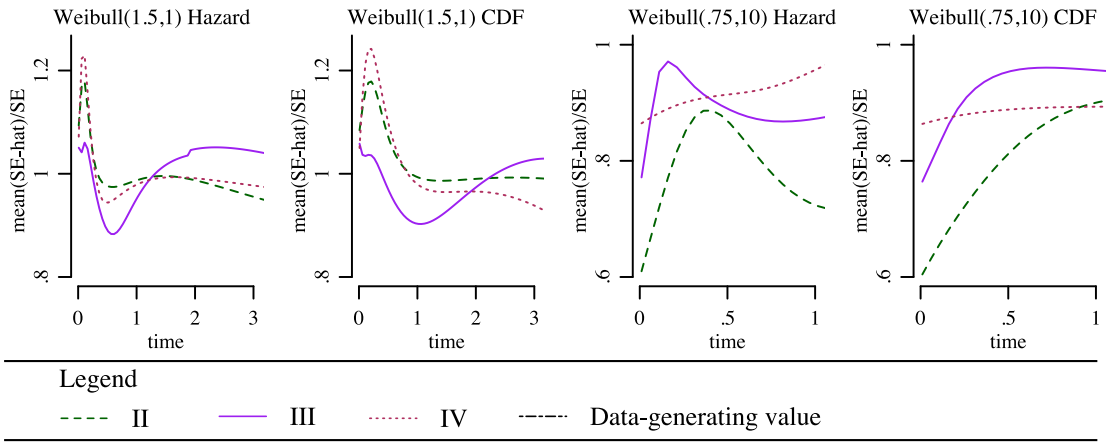

Figure 1: Ratio of average delta-method standard errors to the empirical standard errors of the estimates from simulated data. Models II and III fit survival data with Coxian PH models with 2 and 3 transient states, respectively; Model IV fits discretely observed data from a 2-state reversible model assuming sojourn distributions analogous to model II.