Research Article

Statistics in Medicine

Received 15 August 2012, Accepted 1 May 2013 Published online 5 June 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5861

Fitting and interpreting continuous-time latent Markov models for panel data‡

Jane M. Lange[a] and Vladimir N. Minin[b]*†

Multistate models 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 accordingly, 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 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 dataset of lung transplant patients. We hope our work will encourage wider use of these models in the biomedical setting.

Keywords: disease process; EM algorithm; multistate model; panel data; phase-type

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 [1], HSV-2 [2], and multiple sclerosis [3]. Our interest is estimating disease state functionals – functions of model parameters that characterize individual and population-level disease process dynamics. These functionals include disease state prevalence and hazard and cumulative distribution functions (CDFs) of disease state sojourn times.

Fully observed disease process trajectories present many options for model fitting [4]. 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 [5, 6]. CTMCs entail two strong assumptions: (i) the Markov property indicates that

[a]*Department of Biostatistics, University of Washington, Seattle, WA, U.S.A.*
[b]*Department of Statistics, University of Washington, Seattle, WA, U.S.A.*
**Correspondence to: Vladimir N. Minin, Department of Statistics, University of Washington, Seattle, WA, U.S.A.*
†*E-mail: vminin@uw.edu*
‡*Supporting information may be found in the online version of this article.*

transition probabilities depend on an individual's history only through the current state, and (ii) 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 by 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 [7] or processes in which some states have exponential sojourn distributions [8].

Titman and Sharples [9] 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 because they can approximate generic distributions with positive support [10], and PH functionals, such as hazard rates and CDFs, are easily expressible with matrix exponentials. Aalen [11] 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 [12]. Latent CTMC models of disease processes inherit these advantages and disadvantages.

Our focus is on parameter estimation of the latent CTMC model in the panel data setting. Titman and Sharples [9] 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 [13]. Because 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 [9] 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–maximization (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 (LL) conditional on observed data and current parameter estimates [14]. 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 either with an analytic maximization step (M-step) or with a few iterations of the Newton–Raphson algorithm.

Bureau *et al.* [15] 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 the direct maximization of the observed data likelihood and the EM algorithm of Bureau *et al.* [15], particularly when we apply the EM acceleration of Varadhan and Roland [16].

Our EM algorithm uniquely combines computational developments derived for PH models [17] and discretely observed CTMCs [18] and uses efficient methods developed for HMMs to sum over the latent states [19]. Our EM method shares a similar complete data space and E-step as the EM algorithm that Roberts and Ephraim [20] 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.* [19].

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 [17, 21, 22], 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 [9], both to compare performance of different fitting methods and to illustrate model interpretation, emphasizing clinically relevant functionals of the disease process [23]. 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 continuous-time Markov chain 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, $W(t) = j \iff X(t) \in \{j_1, j_2, \ldots, j_{s_j}\}$. For example, Figure 1A shows a latent trajectory $X(t)$ and the corresponding disease trajectory $W(t)$ for a two-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 [9] specify the sojourn distributions of $W(t)$ to have Coxian PH structure. Coxian PH models assume that 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 two-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

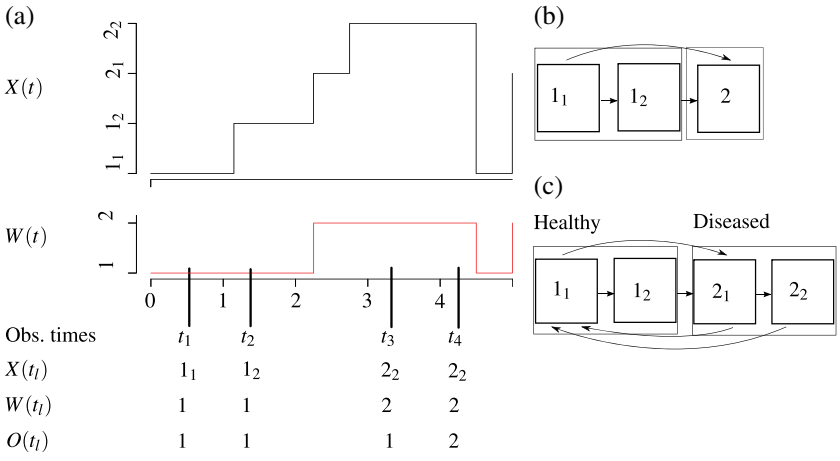

Figure 1. (a). Example of latent trajectory $X(t)$, disease trajectory $W(t)$, and observed data $O(t_l)$ at discrete observation times for model in subfigure C, assuming possible misclassification error. (b). Two-state survival model of $W(t)$ assuming $R = \{1, 2\}$ and $S = \{\{1_1, 1_2\}, \{2\}\}$,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

$e(i,j) = \mathrm{P}(O_t = j | X(t) = i)$ 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

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

where $\mathrm{P}_{x_i x_{i+1}}(t_{i+1} - t_i) = \mathrm{P}\left(X(t_{i+1}) = x_{i+1} | X(t_i) = x_1\right)$ and $\pi_{x_1} = \mathrm{P}\left(X(t_1) = x_i\right)$. 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}.

2.3. *Adding covariates to the latent continuous-time Markov chain model*

We can parameterize $\boldsymbol{\Lambda}$ 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 $\log\left(\lambda_{ij}^h\right) = \boldsymbol{\beta}_{ij}^T \mathbf{w}^h$, 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 $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\Lambda}, \mathbf{E})$ characterize the initial distribution, CTMC transitions, and emission probability matrix, respectively. The complete data log-likelihood (LL) 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 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 LL 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. Expectation–maximization algorithm

3.1. *Maximization step*

The exponential family form of the complete data LL enables a straightforward M-step in the EM algorithm. Web Appendix A provides the score vectors and Hessian matrices for $\boldsymbol{\Lambda}$, $\boldsymbol{\pi}$, and $\mathbf{E}$. 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, we can solve the score equations by 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 $\boldsymbol{\theta}^{(r)} = \boldsymbol{\theta}^{(r-1)} - \ddot{l}\left(\boldsymbol{\theta}^{(r-1)}\right)^{-1} \dot{l}\left(\boldsymbol{\theta}^{(r-1)}\right)$. We can apply this procedure 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 [6].

3.2. *Expectation step*

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

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

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 $\mathrm{P}(\mathbf{o})$ 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

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

We can obtain expectations of $d_{T_l}(i)$ and $n_{T_l}(i,j)$ 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 $\int_0^t \lambda_{ij}\mathrm{P}_{ai}(u)\mathrm{P}_{jb}(t-u)\mathrm{d}u$ and $\int_0^t \mathrm{P}_{ai}(u)\mathrm{P}_{ib}(t-u)\mathrm{d}u$, respectively [18]. We calculate the joint expectation integrals via the efficient matrix-based methods of Minin and Suchard [24,25]. These methods assume $\boldsymbol{\Lambda}$ has no repeated eigenvalues and rely on eigen-decomposition. When $\boldsymbol{\Lambda}$ has repeated eigenvalues, we compute the integrals by using the uniformization approach [26,27].

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 $\mathrm{E}[d_t(i)n_t(l,m))I(X_t = c)|X_0 = a]$. Minin and Suchard [25] and Hobolth and Jensen [26] provide details for these computations using eigen decomposition and uniformization, respectively.

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 $\frac{\partial}{\partial t}\mathrm{E}[S \times I(Y < t)|X_0 = a] = \mathrm{E}[S|X_0 = a, Y = t] \times f(t|X_0 = a)$. Asmussen *et al.* [17] present the methods for obtaining these moments, and Web Appendix C describes these methods in detail.

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. 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, $\mathbf{o}$. Another option is recursive smoothing, described by Cappe *et al.* [19] 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.

We obtain the ultimate target, $\mathrm{E}[t_n(\mathbf{x}_{1:n})|\mathbf{o}_{1:n}]$, 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.* [19] showed that updates to the auxiliary functions are given by

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

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

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

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

There is no computational advantage to using recursive smoothing over our first method for first moment calculations. However, we can also use recursive smoothing 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 because 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

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 Louis' formula [28] as

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

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). We can obtain approximate interval estimates for disease process functionals such as hazard functions and first passage CDFs with delta-method standard errors [29] (Web Appendix D).

5. Implementation

5.1. Software

We have implemented the EM algorithm in R [30], 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 [31] and RcppArmadillo packages [32].

5.2. Speeding up the expectation–maximization with acceleration methods

The EM algorithms are robust but slow, displaying linear rates of the convergence in the vicinity of the maximum LL [14]. EM acceleration algorithms, such as the squared iterative method of Varadhan and Roland [33], 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 [16]. In our tests, SQUAREM reduces the time to convergence of our EM algorithm by a factor of six without substantial loss of robustness.

5.3. Practical considerations for using the expectation–maximization algorithm

The EM algorithms will converge to local maxima, global maxima, or stationary points [34]. Latent parameter models are frequently multimodal or have local maxima, underscoring the need to use multiple starting values. Some starting values may lead to solutions corresponding to infinite values for

certain λ_{ij}, and successive EM iterations of estimates for these λ_{ij} increase without bound. These solutions are outside the parameter space for $\mathbf{\Lambda}$. Performance of the EM is also problematic given numeric inaccuracies in calculating $\exp(\mathbf{\Lambda} t)$ when certain λ_{ij} are high. For practical purposes, it may be worth bounding estimates of λ_{ij} from above. Choice of starting values for $\mathbf{\Lambda}$ is also important: They should be close enough to zero to encourage convergence to estimates with finite or zero values of λ_{ij}, but disperse enough to make it likely one detects the global maximum. In practice, we have generated random starting values for $\log(\lambda_{ij})$ from Normal($\mu = 0$, $\sigma = .25$), but it is worth experimenting with different starting distributions for specific models and datasets.

With discretely observed data, MLEs with finite entries for $\mathbf{\Lambda}$ may not exist [35]. This is more likely as observation intervals are more distantly spaced and as latent transition rates increase. Higher latent transition rates are associated with higher dimensional latent CTMCs used to approximate the data-generating distribution. Empirically, nonexistence of an MLE may be detected when multiple starting values fail to find a global maximum within the allowable parameter space. In this case, investigators should be aware when they have reached the resolution limit for their process and fit a model with fewer latent states.

5.4. Model selection

Model selection involves choosing a structure for the latent CTMC rate matrix, choosing the dimension of the latent space, and adding covariates to the rate matrix, initial distribution, and misclassification model. Although other latent structures are possible, we are advocating models with disease state sojourn distributions characterized by Coxian PH structure, because these models can represent distributions with increasing, decreasing, and nonmonotonic hazard functions and will be uniquely parameterized except in degenerate situations [10]. Choosing the number of latent states is akin to choosing the number of mixture components in a mixture model, which is challenging from a statistical perspective. Pragmatically, we recommend comparing models via the BIC because it is easy to obtain, can be used to compare nonnested models, and has been shown to be adequate in choosing the number of mixture components [36]. We suggest starting by fitting models with small latent spaces and building up to more complex models as appears warranted by the data. One can also determine the dimensionality for which adding more latent states has little effect on plotted point estimates of hazard functions, CDFs, or other functionals of interest.

6. Simulation study

Latent CTMC models can approximate disease state sojourn time distributions with arbitrary hazard functions. We used simulated data to assess the quality of such approximations under different data-generating and observation scenarios, focusing on how bias and root mean squared error of the latent CTMC estimates of hazard functions and first passage time CDFs were affected by data-generating distribution, observation scheme, and number of latent states in the model. We were also interested in coverage of confidence intervals for hazard and CDFs based on delta-method standard errors.

We generated data for two-state survival and reversible semi-Markov models with Weibull sojourn distributions with increasing (shape $= 1.5$, scale $= 1$) and decreasing (shape $= 0.75$, scale $= 10$) hazards. Sojourn time distributions with increasing and decreasing hazards are both common in actual disease models. We generated 100 datasets for each of the three scenarios (survival with increasing hazard, survival with decreasing hazard, two-state reversible semi-Markov model with increasing, and decreasing sojourn distributions). With the survival data, we exactly observed death times unless they exceeded 20, in which case they were right censored; we observed the reversible process discretely at times $(0, 1, ..., 10)$, jittered by Uniform$(-0.5, 0.5)$ random deviates.

We analyzed the simulated data using latent CTMC models with Coxian PH distributions. Models II and III fit survival data with Coxian PH models with two and three transient latent states and one absorbing state, respectively; model IV fits discretely observed data from a two-state reversible model assuming sojourn distributions with two latent states, analogous to model II. These models are able to capture sojourn time distributions with increasing or decreasing hazard functions and are less prone to identifiability or convergence problems than models with more latent states. All data were fit with our EM algorithm, accelerated by the SQUAREM method, using 10 different random starting values per dataset. We estimate hazard and CDFs of sojourn distributions for each dataset by using the corresponding models.

We limit our analysis to the 96% (481/500) of the simulated datasets that had more than one starting value that converged to the putative maximum LL. Spot checks on the remainder of datasets suggested that convergence failure was alleviated when the accelerated EM algorithm was replaced with the traditional version. Evaluation of interval estimates based on delta-method standard errors was further limited to datasets with unique MLEs of latent CTMC parameters ($449/481 = 93\%$).

Figure 2 summarizes our simulation results by reporting means of CDF and hazard function estimators (first row), biases (second row), root mean squared errors (third row), and coverages of point-wise 95% confidence intervals corresponding to these estimators (fourth row). We observe that estimating a hazard function is more difficult then estimating the corresponding CDF. Also, latent CTMC hazard estimators perform better near $t = 0$, which is expected, because latent CTMC hazard functions are asymptotically ($t \to \infty$) constant. The shape of the true hazard function and, interestingly, observation scheme (discrete

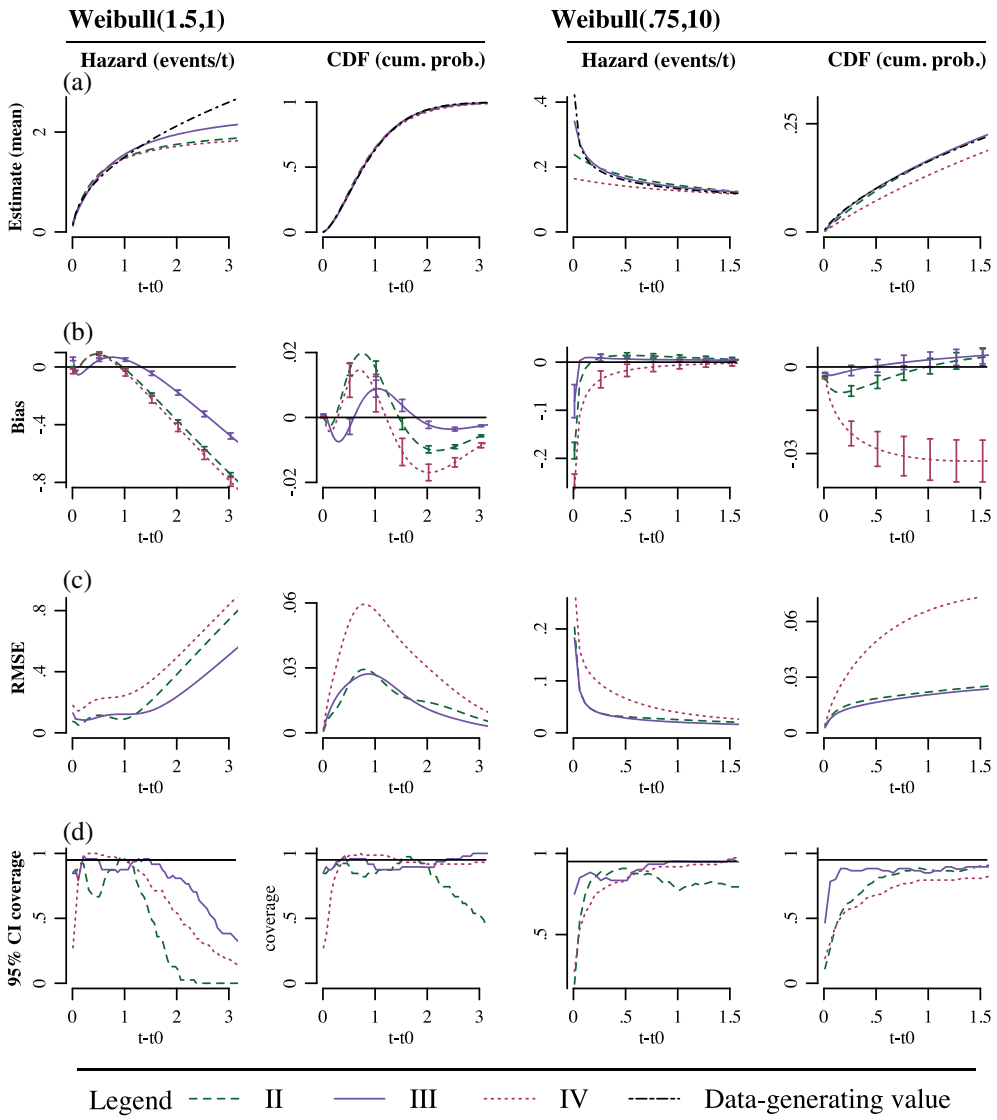

Figure 2. Summary of estimates of CDFs and hazard functions based on models fit to data generated from Weibull(1.5, 1) and Weibull(0.75, 10) sojourn distributions. Models II and III fit survival data with Coxian PH models with two and three transient states, respectively; Model IV fits discretely observed data from a two-state reversible model assuming sojourn distributions analogous to model II. The data were generated with an arbitrary time scale, and the x-axis $t - t0$ refers to time since entry into the state. (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.

vs. continuous) significantly affect bias of latent CTMC hazard estimates. We provide a more detailed discussion of the simulation results in Web Appendix E.

7. Application: Bronchiolitis Obliterans Syndrome

Following lung transplantation, patients are at risk of developing BOS, 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 [37]. Titman and Sharples [9] 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 at 6 months post-transplant and at 9, 12, and every 6 months thereafter [38]. Our version of the dataset consisted of 122 double lung and 244 heart lung patients. We excluded individuals with only baseline observations.

The BOS disease process, $W(t)$, has a state space with three states: $R = \{1 = healthy, 2 = BOS, 3 = death\}$, where death is absorbing. The model of Titman and Sharples [9] 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 $W(t)$. To promote parsimony, the intensity matrix $\mathbf{\Lambda}$ is structured, as $\lambda_{1_2 2_1} = \tau_1 \lambda_{1_1 2_1}$, $\lambda_{1_2 3} = \tau_1 \lambda_{1_1 3}$, $\lambda_{2_2 1_1} = \tau_2 \lambda_{2_1 1_1}$, and $\lambda_{2_2 3} = \tau_2 \lambda_{2_1 3}$. This parameterization says 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 expectation–maximization and other optimization methods

Using maximum likelihood to fit the model of Titman and Sharples (2010) to the BOS dataset, we compared the performance of our EM algorithm (denoted EM1) with the following: (i) the EM of Bureau *et al.* [15] (EM2), (ii) the R implementation of Nelder-Mead (NM) [39], and (iii) the box-constrained Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithms [40]. The BFGS constraints assumed that all model parameters, log-transformed if necessary, 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 [16].

We considered scenarios in which the emission and initial probabilities were unknown or known and fixed at their MLEs. All methods used the same 30 random starting values generated independently from Normal$(0, \sigma = 0.25)$. EM convergence was declared when successive iterations of the LL differed by $< 10^{-6}$, or 200 iterations were taken, whichever came first. We ran NM and BFGS algorithms with the default relative convergence tolerance of 'optim' (10^{-8}) and capped likelihood evaluations at 2500.

Table I summarizes the performance of each of the algorithms for both BOS data models. Figure 3 shows the runtime, in time either to convergence or to the maximum number of iterations, and the final value of the LL. Our method, EM1, was the clear winner in terms of runtime, taking a median of 80 s to converge when π and $\mathbf{E}$ are unknown. Other methods ran between 5.5 and 18 min before converging or reaching the maximum iteration limit. NM, BFGS, and EM2 (which used BFGS for its M-step) all had trials where the algorithm broke for specific reasons: breakdown of the simplex (NM) and entering nondifferentiable regions of the parameter space (BFGS). EM1 did not encounter issues in computing the M-step or E-steps for this model.

The maximum attained LL for the BOS data model was -1248.602. There were at least two additional local optima or stationary points. NM was particularly poor at converging to either global or local maxima, reaching the iteration limit for $11/30$ trials when $(\pi, \mathbf{E})$ were known and $25/30$ trials when $(\pi, \mathbf{E})$ were unknown. The other methods were all subject to convergence to local, rather than global, optima. When $(\boldsymbol{\pi}, \mathbf{E})$ were unknown, EM1 converged to local optima in $18/30$ trials, EM2 in $16/30$ and BFGS in $10/30$ trials. In the scenario where $(\boldsymbol{\pi}, \mathbf{E})$ were known, EM1 converged to the global maximum for all but one starting value.

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

	E, π fixed				**E**, π unknown			
	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	29	24	11	13	12	11	0	18
Convergence to local max. or stationary point	1	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

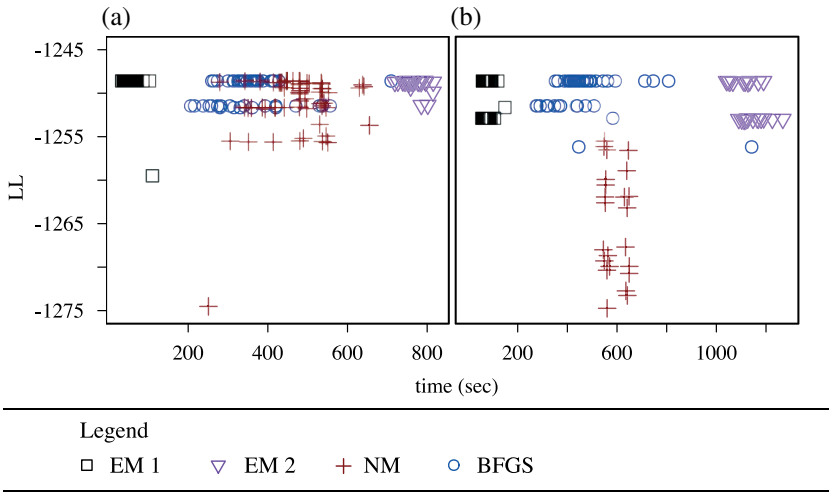

Figure 3. Runtime and attained LL when EM1 (our method), EM2, BFGS, and NM algorithms were used to fit the BOS data, using 30 random starting values and assuming either (**E**, $\boldsymbol{\pi}$) was fixed (a), or was unknown (b).

7.2. Bronchiolitis obliterans syndrome results

Our model parameter estimates are similar, but not identical to those obtained by Titman and Sharples [9], because of differences between the two datasets. Both sets of MLEs were evidently unique, on the basis of numeric investigations with different starting values. Web Appendix Table 2 shows estimates and 95% confidence intervals for the rate, emission, and intensity parameters on their original scales (i.e., rates, emission, and initial probabilities).

Figure 4a represents the first passage CDF for BOS development. The model estimates that the probability of an initial healthy individual remaining BOS free at 5 years post transplant is 34%, with a 95% confidence interval of (26%, 44%). This is consistent with estimates in the literature of a 5 year disease free probability ranging from 15% to 37% [41]. The model also predicts that 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 nonconstant disease hazard of BOS likely reflects heterogeneity in the lung transplant patient population in terms of progression to BOS. Declining BOS rates are also consistent with the initial period after transplant being a time of high risk for patients for experiencing infections or acute rejection episodes, both of which may trigger BOS development [38].

Figure 4c shows the cumulative probabilities of death conditional on starting in healthy state 1_1 versus BOS state 2_1. By 2 years post transplant, we estimate that 12% of those healthy at the start of the study will have died. After developing BOS, nearly 72% remain alive at 1 year, 50% at 2 years, and 35% at 3 years. These estimates are in agreement with the literature estimates of survival after bilateral lung transplant of 74%, 46%, and 26% at 1, 3, and 5 years after the onset of BOS, respectively, [42].

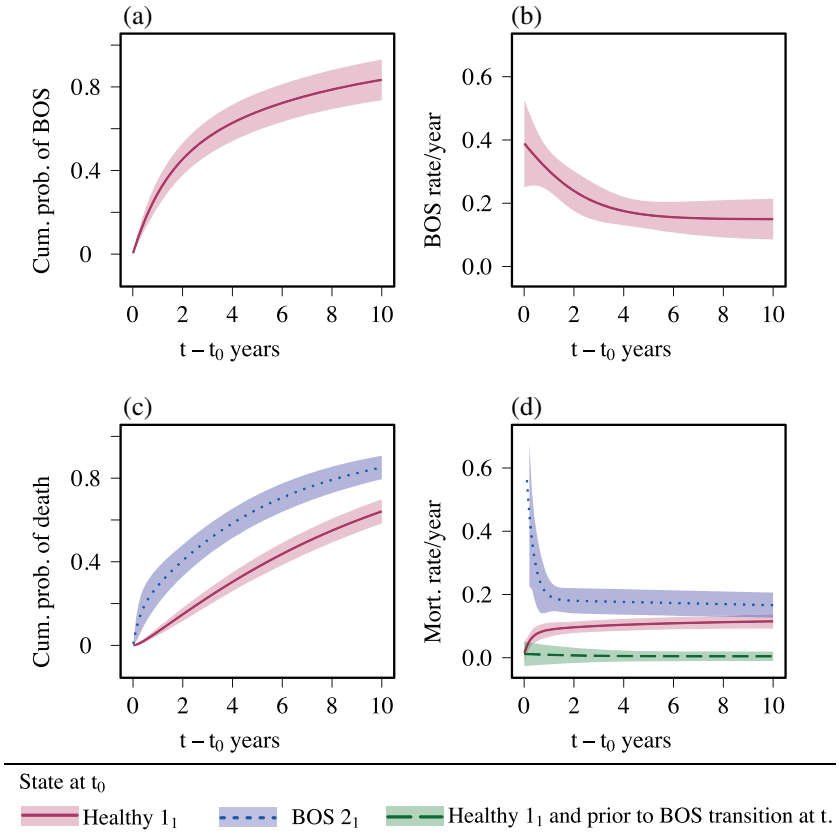

Figure 4. (a). Cumulative probability of having transitioned to BOS state at least once, conditional on being in 1_1 at t_0. (b). Disease rate conditional on being in healthy state 1_1 at t_0. (c). Cumulative probability of death. (d). Mortality rate per year, as a function of state at t_0. In all figures, the shaded regions represent 95% point-wise confidence intervals for the estimates.

Our model estimates that mortality, as with BOS onset, has declining hazard rates after an individual has developed the disease. Prior to BOS development, mortality rates are very low (Figure 4d). After transitioning to BOS state 2_1, mortality rates jump dramatically (>50% per year), and then drop to 20% after 1 year. This pattern in mortality is consistent with the identification of distinct BOS patient populations: those with acute onset and rapidly deteriorating lung function, and those with more gradual onset and slowly progressing disease [38, 43].

The latent CTMC model presented here allows for reversible transitions between BOS and healthy states despite the fact that biologically, BOS is irreversible. Initially, the rate of reversion is estimated at 6%, dropping to 1.6% after a year of having BOS. Titman and Sharples [9]) included BOS → healthy transitions based on comparing a standard HMM with and without reversible transitions via a likelihood ratio test. To investigate whether the addition of latent states made it unnecessary to include the reverse transitions, we compared our model with models with no BOS → healthy transitions and either two or three latent states per healthy state. We also found support for inclusion of the reversible transitions (Web Appendix F). It is plausible that the model's estimates of low rates of BOS→ healthy transitions near the time of disease development may reflect nonconstant misclassification probabilities with respect to BOS duration.

To assess how well the distribution of the time to BOS development, induced by the latent CTMC model, fits the observed data, we compared model-based simulations of time of first observed BOS diagnosis or death with the actual times in the data. Using MLEs for model parameters, we simulated 1000 new disease trajectories for each real individual, retaining real observation and censoring times and imputing new times if the real participant's death occurred prior to the corresponding simulated time of BOS/death. We compared the Kaplan–Meier (K-M) estimates of observed failure time distribution in the real data with analogous K-M estimates in the artificial data (Web Appendix Figure 3). Given that the observed K-M curve is within the envelope of the simulated curves, the model appears reasonable in predicting time to BOS development, particularly in the first 5 years after lung transplant.

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. The 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. On the BOS dataset, the method considerably outperformed other optimization approaches, including those implemented in the R package msm [44] – NM and BFGS – and the EM algorithm of Bureau *et al.* [15]. We suspect that results will be similar for other datasets fit with well-behaved latent CTMC models, in that the out-of-the-box numeric optimization methods will be considerably slower to converge than the accelerated version of our EM algorithm. We also suspect that our EM algorithm would be faster than a Newton–Raphson algorithm, because the latter algorithm requires the computationally expensive calculation of the observed information matrix [45] at each iteration.

The utility of latent CTMC models lies in their ability to approximate functionals of disease processes from nonexponential sojourn time distributions. Our simulation studies investigated frequentist properties of such estimates using simple survival and two-state disease models to analyze data with Weibull distributed sojourn times. In practice, many diseases will have more than two states, nonmonotonic hazard functions, and misclassification error. Although limited in scope, our investigations fill a gap in the literature of frequentist properties of latent CTMC parameter MLEs under model misspecification. We believe that the results from two-state simulation studies generalize to more complex disease models.

Overall, these simulation results suggest mixed performance of latent CTMC estimates. Latent CTMC models, while flexible, are parametric and therefore subject to model misspecification. The bias in approximations reflects the closeness of the data-generating distribution to that implied by the latent CTMC model. In particular, although CDF estimates were generally good, hazard estimates may be quite biased at the times corresponding to the distribution's tail, when latent CTMC hazards are asymptotically constant. However, estimates of the hazard function near the tail of the distribution may be of limited scientific interest and may be extrapolations to times after all events have occurred. Investigators should also be aware that estimates of hazard and CDFs may be more biased for panel data. It will be worth investigating further the sensitivity of estimate bias under different sampling frequencies and data-generating scenarios.

Confidence interval coverage for disease process functionals was sometimes poor, reflecting both the bias in the estimates and in certain cases, underestimation of the variability of the estimations using the delta-method approach. It is worth investigating the use of robust variance estimates in the EM algorithm context to yield more valid standard errors [46]. We limited our evaluation of confidence intervals to datasets with unique MLEs. In practice, the likelihood may be multimodal, in which case delta-method standard errors will not be appropriate. In the absence of unique MLEs, we recommend applying a nonparametric bootstrap. The computation time required would not be prohibitive given the increased efficiency of our fitting algorithm.

The issue of model selection for CTMC models still presents many open questions. We have advocated using the BIC to select the dimensionality of the latent state space, given its practical performance and ease of use [36]. A likelihood ratio testing framework for nested models is also possible but has accompanying challenges. It is possible to represent a latent CTMC model with p latent states within a space of $k > p$ latent states, but such parameterization is not unique. Penalized likelihood ratio tests allow for hypothesis testing in the setting of nonidentifiable parameters under the null model [47]. Currently, implementation in the latent CTMC context is limited to null models with exponential sojourn distributions [9], and extending this approach for more general testing is an area of future research. Given the increased efficiency of our fitting algorithm suggests that it may also be practical to evaluate models using k-fold cross validation with a goodness of fit statistic measuring prediction error [48].

Our focus has been on frequentist estimation. Bayesian methods also have a strong appeal in this setting [12]. 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 [49]. McGrory *et al.* [50] 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.

Acknowledgements

The authors thank Andrew Titman, Linda Sharples, and Steven Tsui for their help in obtaining access to the BOS data from the Papworth Hospital, UK. These data have been used to illustrate the statistical methods, and results should not be used in isolation to inform clinical practice. Jane M. Lange was supported by NIH grants No. T32 CA009168 and R01CA160239. Vladimir N. Minin was supported by the NSF grants No. DMS-0856099. We thank Lurdes Inoue and Kenneth Lange for their comments on the manuscript.

References

1. Guihenneuc-Jouyaux C, Richardson S, Longini IM. Modeling markers of disease progression process by a hidden Markov process: application to CD4 cell decline. *Biometrics* 2000; **56**(3):733–741.
2. Crespi CM, Cumberland WG, Blower S. A queueing model for chronic recurrent conditions under panel observation. *Biometrics* 2005; **61**:194–199.
3. Mandel M. Estimating disease progression using panel data. *Biostatistics* 2010; **11**(2):304–16.
4. Andersen PK, Keiding N. Multi-state models for event history analysis. *Statistical Methods in Medical Research* 2002; **11**(2):91–115.
5. Kalbfleisch JD, Lawless JF. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* 1985; **80**(392):863–871.
6. Lange K. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1995; **57**(2):425–437.
7. Foucher Y, Giral M, Soulillou JP, Daures JP. A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. *Statistics in Medicine* 2007; **26**:5381–5393.
8. Kang M, Lagakos SW. Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics* 2007; **8**(2):252–264.
9. Titman AC, Sharples LD. Semi-Markov models with phase-type sojourn distributions. *Biometrics* 2010; **66**(3):742–52.
10. Cumani A. On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectronics and Reliability* 1982; **22**(3):583–602.
11. Aalen OO. Phase type distributions in survival analysis. *Scandinavian Journal of Statistics* 1995; **22**(4):447–463.
12. Bladt M, Gonzalez A, Lauritzen SL. The estimation of phase-type related functionals using Markov chain Monte Carlo methods. *Scandinavian Actuarial Journal* 2003; **2003**(4):280–300.
13. Baum L, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* 1970; **41**(1):164–171.
14. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 1977; **39**(1):1–38.
15. Bureau A, Shiboski S, Hughes JP. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine* 2003; **22**(3):441–462.
16. Varadhan R. SQUAREM: Squared extrapolation methods for accelerating fixed-point iterations, 2011. R package version 2010.12-1.
17. Asmussen S, Nerman O, Olsson M. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics* 1996; **23**(4):419–441.
18. Hobolth A, Jensen JL. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical Applications in Genetics & Molecular Biology* 2005; **4**(1):1–22.
19. Cappe O, Moulines E, Ryden T. *Statistical Inference for Hidden Markov Models*. Springer: New York, 2005.
20. Roberts W, Ephraim Y. An EM algorithm for ion-channel current estimation. *IEEE Transactions on Signal Processing* 2008; **56**(1):26–33.
21. Faddy M. On inferring the number of phases in a Coxian phase-type distribution. *Communications in Statistics* 1998; **14**:407–417.
22. Marshall AH, Zenga M. Experimenting with Coxian phase-type distributions to uncover suitable fits. *Methodology in Computational Applied Probability* 2010; **14**(1):71–86.
23. Andersen PK, Keiding N. Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine* 2012; **31**:1074–1088.
24. Minin V, Suchard M. Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology* 2008; **56**(3):391–412.
25. Minin V, Suchard M. Fast, accurate and simulation-free stochastic mapping. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 2008; **363**(1512):3985–3995.
26. Hobolth A, Jensen J. Summary statistics for endpoint-conditioned continuous-time Markov chains. *Journal of Applied Probability* 2011; **48**:911–924.
27. Bladt M, Esparza L, Nielsen B. Fisher information and statistical inference for phase-type distributions. *Journal of Applied Probability* 2011; **48**:277–293.
28. Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society* 1982; **44**(2):226–233.
29. Gentleman R. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV Disease. *Statistics in Medicine* 1994; **13**:805–822.
30. R Development Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.
31. Eddelbuettel D, François R. Rcpp: seamless R and C++ integration. *Journal of Statistical Software* 2011; **40**(8):1–18.

32. Francois R, Eddelbuettel D, Bates D. RcppArmadillo: Rcpp integration for Armadillo templated linear algebra library, 2011. R package version 0.2.34.
33. Varadhan R, Roland C. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics* 2008; **35**(2):335–353.
34. Wu CFJ. On the convergence properties of the EM algorithm. *The Annals of Statistics* 1983; **11**(1):95–103. DOI: 10.2307/2240463.
35. Bladt M, Sorensen M. Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005; **67**(3):395–410.
36. Steele R, Raftery A. Performance of Bayesian model selection criteria for Gaussian mixture models. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, Chen M-H, Muller P, Sun D, Ye K, Dey D (eds). Springer: New York, 2010; 113–130.
37. Estenne M, Maurer JR, Boehler A, Egan JJ, Frost A, Hertz M, Mallory GB, Snell GI, Yousem S. Bronchiolitis obliterans syndrome 2001: an update of the diagnostic criteria. *The Journal of Heart and Lung Transplantation* 2002; **21**(3):297–310.
38. Jackson CH, Sharples LD, McNeil K, Stewart S, Wallwork J. Acute and chronic onset of bronchiolitis obliterans syndrome (BOS): Are they different entities? *The Journal of Heart and Lung Transplantation* 2002; **21**(6):658–666.
39. Nelder J, Mead R. A simplex algorithm for function minimization. *Computer Journal* 1965; **7**:308–313.
40. Byrd R, Lu P, Noecedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing* 1995; **16**:1190–1208.
41. Chan A, Allen R. Bronchiolitis obliterans: an update. *Current Opinion in Pulmonary Medicine* 2004; **10**(2):133–141.
42. Finlen Copeland CA, Snyder LD, Zaas DW, Turbyfill WJ, Davis WA, Palmer SM. Survival after bronchiolitis obliterans syndrome among bilateral lung transplant recipients. *American Journal of Respiratory and Critical Care Medicine* 2010; **182**(6):784–789.
43. Lama VN, Murray S, Lonigro RJ, Toews GB, Chang A, Lau C, Flint A, Chan KM, Martinez FJ. Course of FEV(1) after onset of bronchiolitis obliterans syndrome in lung transplant recipients. *American Journal of Respiratory and Critical Care Medicine* 2007; **175**(11):1192–1198.
44. Jackson CH. Multi-state models for panel data: the msm package for R. *Journal of Statistical Software* 2011; **38**(8):1–29.
45. Lystig TC, Hughes JP. Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics* 2002; **11**(3):678–689.
46. Elashoff M, Ryan L. An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics* 2004; **13**(1):48–65.
47. Chen H, Chen J, Kalbfleisch JD. A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2001; **63**(1):19–29. DOI: 10.1111/1467-9868.00273.
48. Titman A, Sharples L. A general goodness-of-fit test for Markov and hidden Markov models. *Statistics in Medicine* 2008; **27**:2177–2195.
49. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995; **82**(4):711–732.
50. McGrory CA, Pettitt AN, Faddy M. A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. *Computational Statistics & Data Analysis* 2009; **53**(12):4311–4321.