Latent Continuous Time Markov Chains for Partially-Observed Multistate Disease Processes

Jane M. Lange

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Vladimir Minin, Chair

Rebecca Hubbard

Lurdes Inoue

Program Authorized to Offer Degree:

Biotatistics

University of Washington

Abstract

Latent Continuous Time Markov Chains for Partially-Observed Multistate Disease Processes

Jane M. Lange

Chair of the Supervisory Committee:
Vladimir Minin
Department of Statistics, University of Washington

A disease process refers to a patient's traversal over time through a disease with multiple discrete states. Multistate models are powerful tools used to describe the dynamics of disease processes. Clinical study settings present modeling challenges, as patients' disease trajectories are only partially observed, and patients' disease statuses are only assessed at discrete clinic visit times. Furthermore, imperfect diagnostic tests may yield misclassification error in disease observations. Observational data, such as that available in electronic medical records (EMR), present additional challenges, since patients initiate visits based on symptoms, and these times are informative about patients' disease histories. Many of the flexible modeling methods suited for fully observed trajectories are no longer tractable with partially observed data. A typical approach is to assume a standard continuous time Markov chain for the disease process, due to its computational tractability. This assumption means that disease state sojourn times have constant hazard functions, which is frequently unrealistic. Our approach is to model the disease process via a latent continuous time Markov chain, enabling greater flexibility yet retaining tractability. We devise a novel expectation-maximization algorithm (EM) for fitting these models in a panel data setting in which observation times are non-informative. We then extend the model and the EM algorithm to accommodate observation times that are patient-initiated and informative about the disease process. We apply our model to a study of secondary breast cancer events using an EMR dataset of mammography and biopsy records.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

My advisor, Vladimir Minin, has provided invaluable mentoring throughout my graduate career. He has provided insight and appropriate doses of skepticism, and has enabled my research integrating stochastic processes with more traditional epidemiology. His enthusiasm, passion for his own work, and active engagement with the scientific community have been inspiring and motivating during this whole journey. I would also like to thank Lurdes Inoue and Rebecca Hubbard for their financial support of my research via their Semi-Markov Process grant. Their kind support and intelligent and on-point comments both on this thesis and on other manuscripts on which we have collaborated have greatly improved the quality of my efforts.

Rebecca was also instrumental in providing the idea for the secondary breast cancer application study and in guiding me through the many steps of the process for obtaining the data from the Group Health Research Institute. Correspondence with Andrew Titman was key to obtaining access to the bronchiolitis obliterans syndrome data, and I also thank Papworth General Hospital for its generosity in providing access to this dataset.

Throughout my career, I have been blessed with many excellent scientific mentors who have dedicated their careers both to the pursuit of knowledge and to instilling scientific values in the next generation. Norm Breslow, whom I worked with as an research assistant, is one such person. I am lucky to have had the chance to work with him on the Wilms Tumor Project and to learn from his compelling mix of curiosity about both theoretical and applied aspects of our field.

Finally, my parents have provided endless support, both emotionally and intellectually. My father's gifts as a mathematician and his willingness to share them with me throughout my life have been a great gift. My mother's literary talents have helped hone my writing. Both of their patience and love have kept me going in tough times.

DEDICATION

To my parents

Chapter 1

OVERVIEW OF CONTRIBUTIONS

1.1 Multistate models for diseases

Many diseases can be conceptualized as consisting of a finite set of discrete states through which patients transition over time. A disease process is the stochastic progression through these states over the course of a patient's disease history. Depending on the context, states may refer to an underlying biological condition assessed by biomarkers, to the presence or absence of clinical symptoms, or to ordinal measures of disease progression or severity. Multistate models (MSMs) are powerful tools to characterize disease processes, enabling joint modeling of the disease events (Andersen and Keiding, 2002; Meira-Machodo et al., 2009). From a scientific standpoint, MSMs provide a framework for investigating disease process dynamics and how covariates affect transition rates. MSMs can also be used to estimate how disease state prevalence in a population varies over time, which is relevant to public health strategies (Commenges, 1999). Finally, MSMs can be used to make individual level predictions based on observed histories, which may be useful in a clinical setting (Putter et al., 2006; Touraine et al., 2013).

Figure 1.1 shows some commonly encountered MSMs in biostatistics. States may either be transient, which individuals can enter and exit, or absorbing, which individuals never exit once they enter. In disease models, death is an absorbing state. Transitions between the disease states may either be reversible or uni-directional. The simplest MSMs are survival models, which consist of a single transient state and single absorbing state. Other common models include competing risks models, consisting of a single transient state and multiple absorbing states, such as different causes of death (Andersen et al., 2002); and illness-death models consisting of a healthy state, a diseased state, and death (Meira-Machado et al., 2006). MSMs with reversible transitions are appropriate for diseases with recurrent states, such as symptomatic or acute symptomatic episodes in multiple

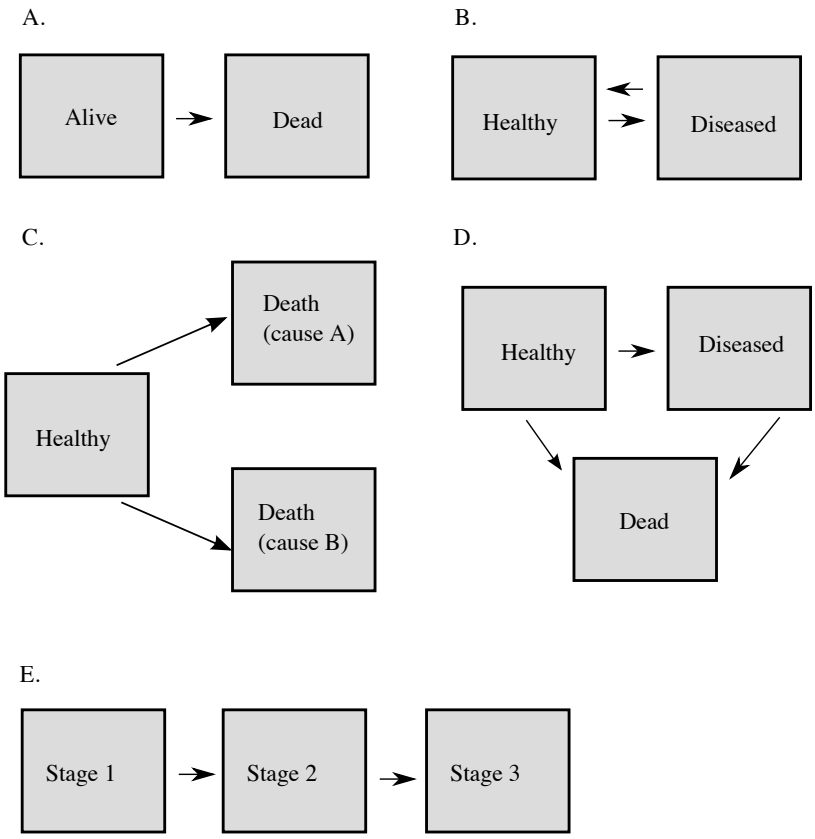

Figure 1.1: Multistate models common in biostatistics. A. Survival model. B. Reversible disease model with two states. C. Competing risks model with two causes of death. D. Illness-death model. E. Progressive, staged disease model.

sclerosis, or infectious and non-infectious periods in diseases such as herpes simplex virus-2 (Crespi et al., 2005; Mandel, 2010). MSMs consisting of forward transitions through underlying discrete states, can be used to characterize progressive diseases such as HIV (Guihenneuc-Jouyaux et al., 2000). More complex MSMs can capture diseases with more complex relationships between events, although the specific model useful in a particular context relies both on the disease and on the model's scientific purpose.

1.2 Challenges of modeling partially observed disease processes

The goal of our research is to develop tools for fitting multistate disease models based on partially observed realizations of the disease process. Rather than observing the actual transition times and states, scientists observe snapshots of the process at discrete time points. Furthermore, observations may be subject to misclassification error (Jackson et al., 2003). Such data are common in clinical research, both in designed and observational studies. In these studies individuals may only have records of their disease status at clinic visits, and their disease status is measured through screening exams or diagnostic tests that may have false positives or negatives. Observational studies present additional challenges, since the sampling times may be informative about the underlying disease status (Gruger et al., 1991). There is particular need for methodology development in the observational data setting, given the potential for exploiting increasingly available electronic medical records (EMR) data sources that capture information from clinic visits (Dean et al., 2009).

This thesis develops two complementary methods for modeling data consisting of discrete, possibly misclassified observations of the disease process, one appropriate for a designed study with non-informative sampling times and the other in an observational setting with informative observation times. The primary challenge for modeling in this setting is that methods that are most tractable make unrealistic assumptions. A common approach for analyzing discretely observed disease process data is to assume that observations are a discretely observed continuous time Markov chain (CTMC); these models are computationally amenable yet assume that an individual's hazard rate for transitioning between states is constant with respect to state sojourn time (Gentleman, 1994). In contrast, more flexible approaches for partially observed MSMs are only available in limited contexts and are unfeasible when the disease models contain reversible transitions.

The approach we take here is a compromise between standard CTMCs and fully non-parametric models. In particular, we view our observed process as partial observations of a CTMC with an expanded state space, where multiple latent states map onto each observed disease state. Such models are referred to as latent, or aggregated, CTMCs. These models exploit the computational tractability of standard CTMCs but allow for disease state sojourn hazard functions to vary with respect to the duration spent in the state. Latent CTMC models are popular in engineering contexts, with

applications in systems repair (Montoro-Cazorla and Perez-Ocorn, 2014), but have seen relatively little use in disease modeling contexts, despite their attractive features. Titman and Sharples (2010) advocated for using latent CTMCs in the setting of panel observations of disease processes with possible misclassification error; throughout, we assume their disease modeling framework.

1.3 Methodology contributions

Our first main contribution is the development of an efficient method for fitting latent CTMC models with discretely observed multi-state data. In particular, we develop a novel way of obtaining maximum likelihood estimates for these models based on an expectation-maximization (EM) algorithm, providing substantial improvements in speed over the out-of-the box numerical optimization methods used by Titman and Sharples (2010). Our estimation method allows for covariates in the parameterization. We also develop an efficient way of approximating the sampling variance of model parameter estimators. Finally, we examine the statistical performance of estimates based on latent CTMCs when they are used in the context of approximating arbitrary sojourn time distributions.

Our second contribution is extending the latent CTMC disease modeling framework to the setting where observation times depend on the individual's disease status. This situation is common with observational clinical data, since patients will be seen at the clinic when they are sick. Because ignoring informative observation times leads to bias in estimates of disease process parameters, it is necessary to jointly model the observation and disease process (Gruger et al., 1991). Our novel mulitstate-disease-driven observation (multistate-DDO) model assumes that the informative observation times follow a Markov-modulated Poisson process with rates that depend on the underlying disease state. The multistate-DDO model can be viewed as a partially observed bivariate CTMC and is therefore computationally tractable. Moreover, with relatively few alterations, we can extend our latent CTMC EM algorithm for estimating parameters in the multistate-DDO model.

We offer an R package (R Core Team, 2013), `cthmm`, available at `http://r-forge.r-project.org/projects/multistate/` that provides software implementations of the EM algorithms for fitting data using latent CTMCs with panel data, both with and without informative observation schemes. The package enables flexible covariate parameterizations for the models and can be used

for a broad variety of multistate disease models.

1.4 Applications

Along with model development and estimation techniques, we focus on application of the latent CTMC and multistate-DDO models to real data situations, with an emphasis on scientific interpretation. Rather than assigning inherent meaning to parameters of the latent model, we view them as a means for capturing functionals of the disease process, such as a hazard rates and sojourn time cumulative distribution functions (Andersen and Keiding, 2012).

We re-examine the application in Titman and Sharples (2010), an analysis of bronchiolitis obliterans syndrome in lung transplant patients (Scott et al., 2005) as an illustration of the latent CTMC model for panel data with non-informative visit times. To illustrate the multistate-DDO model, we apply it to a study of the development of secondary breast cancer events (SBCEs) in women who have experienced primary breast cancer (BC), using an EMR resource consisting of mammogram and biopsy records that reflect both non-informative screening visits and informative, patient-initiated visits. Our multistate-DDO model enables us to investigate time to development of mammographically detectable cancer rather than merely diagnosed disease, providing information about the fraction of women with undiagnosed but mammographically detectable disease over time since the original BC diagnosis. The model also allows us to estimate how baseline covariates affect rates of SBCEs, and how these covariates differentially affect rates of SBCEs that occur on the same or opposite side as the original BC. Finally, the multistate-DDO model allows us to estimate mammography sensitivity and specificity in this population.

Chapter 2

PRELIMINARIES ON MULTISTATE PROCESSES FOR COMPLETE AND PARTIALLY OBSERVED DATA

This chapter introduces terminology and a generic modeling framework for multistate models. We provide a brief overview of key quantities used to characterize MSMs. We also describe Markov and semi-Markov modeling frameworks and their assumptions of how the future of a disease process depends on its history. We then consider forms of the likelihood for fully observed data and the partially observed data scenarios we consider in this work. These scenarios include panel observations with non-informative sampling times, panel observations with informative sampling times, and both situations in the presence of disease misclassification error. To provide background for our introduction of the latent CTMC disease modeling approach, we discuss the existing options for modeling multistate disease in these partial observation settings.

2.1 Multistate model preliminaries

A multistate process $W(t)$ is a stochastic process with a finite state space $R = \{1,2,\ldots,r\}$ and right-continuous sample paths ($W(t+) = W(t)$), for $t \in [0,T]$ (Andersen and Keiding, 2002). In the disease modeling framework, the state space corresponds to the set of possible events in the disease process. The initial state is distributed multinomially with probability vector $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_r)$, where $\pi_i = \mathrm{P}[W(0) = i]$, $i = 1,\ldots,r$.

A single realization of a multistate process across a given time interval is characterized by a trajectory of state occupancies during the time interval (Figure 2.1A). Jumps in the process correspond to state transitions; the random length of time the process remains in the state is referred to as the sojourn time. The history of the process $\mathscr{H}_t$ is an individual's realized trajectory of the disease process on the interval $[0,t)$.

For a given history $\mathscr{H}_t$, the probability of transition between state i and state j between t and $t+s$ is denoted by

$$\mathrm{P}_{ij}(t,t+s|\mathscr{H}_t) = \mathrm{P}[W(t+s) = j|W(t) = i, \mathscr{H}_t].$$

Conditional on the $\mathscr{H}_t$, instantaneous transition probability (also known as the intensity or hazard rate) between states i and j is given by

$$h_{ij}(t|\mathscr{H}_t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}_{ij}(t,t+\Delta t|\mathscr{H}_t)}{\Delta t},$$

assuming the limit exists. If state k is absorbing, $h_{kj}(t) = 0$ for all t and $k \neq j$; otherwise it is transient.

We refer to any function of model parameters and time that characterizes individual and disease process dynamics as a functional of the process, such as hazard functions, and cumulative distribution functions for disease state sojourn times. Another example of a functional is disease state prevalence, characterizing the proportion of individuals in a given disease state at a particular time, given by

$$\mathrm{P}[W(t) = j] = \sum_{i \in R} \pi_i \mathrm{P}_{ij}[0,t|W(0) = i].$$

Functionals are frequently the target of inference in disease modeling frameworks (Andersen and Keiding, 2012).

2.1.1 Model specification: Markov and semi-Markov models

Often, it is convenient to restrict the history dependence of the transition probabilities and intensities, for model parsimony or computational convenience. The strictest form of history dependence is a Markov assumption, specifying that transition probabilities and intensities depend on the process history only through the last occupied state. Markov models based on processes that occur in continuous time are called continuous time Markov chains (CTMCs). Supposing that $W(t)$ is in state i at time t, for $s > t$, the CTMC assumption specifies that $\mathrm{P}_{ij}(t,s|\mathscr{H}_t) = \mathrm{P}_{ij}[t,s|W(t) = i]$. Similarly, the Markov assumption implies that the intensity $h_{ij}(s|\mathscr{H}_t) = h_{ij}[s|W(s) = i]$.

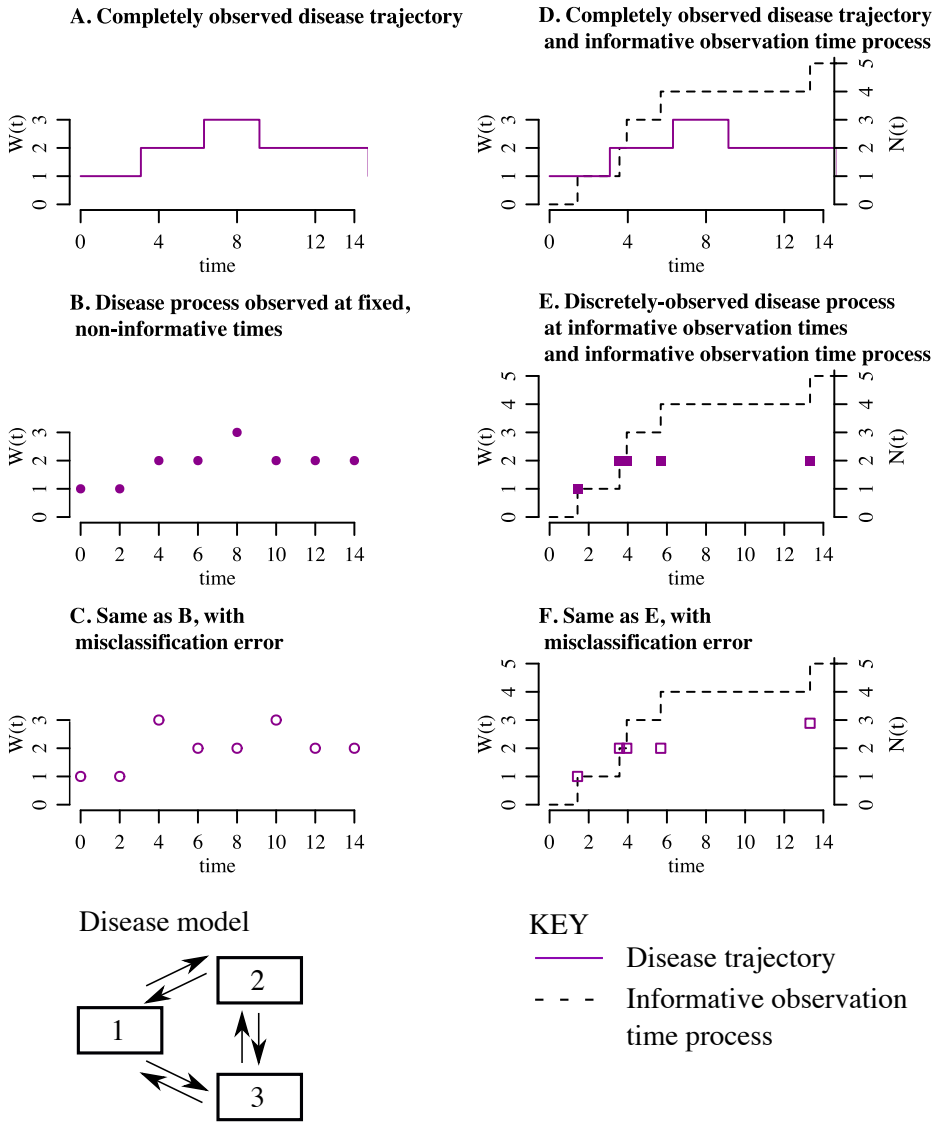

Figure 2.1: Complete and partial data scenarios for multistate models. A. A realization of the full disease process trajectory, W(t). B. Panel observations of the disease process with non-informative sample times. C. Same as B, with misclassification error. D. Completely observed disease process, W(t) and information counting process, N(t). E. Discrete observations at informative sampling times. F. Same as E, with misclassification error.

With time-homogeneous Markov models, discussed in more detail in Chapter 3, the transition probabilities and intensities are independent of s; that is, $P_{ij}(t,s|\mathscr{H}_t) = P_{ij}[0,s-t|W(0)=i]$, and the hazard is a constant $h_{ij}(s|\mathscr{H}_s) = \lambda_{ij}$. In contrast, time-inhomogeneous CTMCs allow for transitions

and intensities to depend on the current time, s. In this case, the hazard is a time-dependent intensity function, $h_{ij}(s|\mathscr{H}_s) = \lambda_{ij}(s)$. Particular time-inhomogeneous models may assume that $\lambda_{ij}(s)$ has a specific functional form, such as a piecewise constant (Kay, 1986) or a spline function (Titman, 2011).

Semi-Markov models provide a modeling framework for specifying hazard functions that depend on the sojourn time duration in a particular state (Lagakos et al., 1978). That is, semi-Markov models specify that the future evolution of the process depends not just on the current state, but also on time since entry into the current state. Suppose that W(t) entered into state i at time t_i. Under the semi-Markov assumption, for $s > t \geq t_i$, the transition probability is

$$\mathrm{P}_{ij}(t,s|\mathscr{H}_t) = \mathrm{P}_{ij}[t,s|W(t) = i, t - t_i = u],$$

and the intensity function is

$$h_{ij}(s|\mathscr{H}_s) = h_{ij}[s|W(s) = i, s - t_i = u].$$

Semi-Markov models, as with Markov models, can be either time-homogeneous or inhomogeneous, reflecting additional dependence on an external time scale. We note also that the semi-Markov assumption for MSMs with one transient and one or more absorbing states–that is, a survival or competing risks model– is equivalent to a time inhomogeneous CTMC whose time origin is time of entry in the transient state.

2.2 Fully observed data

In a complete data situation, we observe an individual's trajectory of $W(t)$ on the interval $[0,T]$ (Figure 2.1A). Assume that $W(t)$ starts at w_0 and there are N_T total transitions occurring at $(\tau_1, \tau_2, \ldots, \tau_{N_T})$, and corresponding states, $(w_1, \ldots, w_{N_T})$. A competing risks interpretation for state sojourns provides intuition of the form of the complete data-likelihood (Andersen et al., 2002). When $W(t)$ enters state w_k at τ_k, each of the possible destination states out of state w_k is a competing event. Conditional on the process history, $\exp\left(-\int_{\tau_k}^{\tau_{k+1}} \sum_{j \neq w_k} h_{w_k,j}(u|\mathscr{H}_u)du\right)$ is the probability of staying in state w_k

between τ_k and τ_{k+1}, and $h_{w_k,w_{k+1}}(\tau_{k+1}|\mathscr{H}_{\tau_{k+1}})$ is the instantaneous probability of transitioning to w_{k+1} at time τ_{k+1}. Thus the likelihood for the fully observed trajectory is

$$L(\tau_1,\tau_2,\ldots,\tau_{N_T},w_0,w_1,\ldots,w_{N_T}) = \pi_{w_0}\exp\left(-\int_{\tau_0}^{\tau_1}\sum_{j\neq w_0}h_{w_0,j}(u|\mathscr{H}_u)du\right)h_{w_0,w_1}(\tau_1|H_{\tau_1})\times\ldots$$
$$\times\exp\left(-\int_{\tau_{n_T-1}}^{\tau_{N_T}}\sum_{j\neq w_{N_T-1}}h_{w_{N_T-1},j}(u|\mathscr{H}_u)du\right)h_{w_{N_T-1},w_{N_T}}(\tau_{N_T})\exp\left(-\int_{\tau_{N_T}}^{T}\sum_{j\neq w_{N_T}}h_{w_{N_t},j}(u|\mathscr{H}_u)du\right).$$

An alternate, equivalent likelihood formulation for complete trajectories can be constructed using notation for marked point processes (Andersen and Keiding, 2002). Consider the trajectory of a single individual on time $[0,T]$. Let $N_T(ij)$ be the total number of direct transitions of $i\to j$ in $[0,T]$, occurring at times

$$\tau_{ij}^1<\cdots<\tau_{ij}^{N_T(ij)}.$$

Denote $Y_i(t)=I\{W(t-)=i\}$ as an indicator whether the individual is in state i just prior to time t. An individual's complete data likelihood is

$$L(\tau_1,\tau_2,\ldots,\tau_{N_T},w_0,w_1,\ldots,w_{N_T}) = \pi_{w_0}\prod_{i\neq j}\prod_{k=1}^{N_T(ij)}h_{ij}(\tau_{ij}^k|\mathscr{H}_{\tau_{ij}^k})\exp\left(-\int_0^T h_{ij}(t|\mathscr{H}_t)Y_i(t)dt\right).$$

When $W(t)$ is a time-homogeneous CTMC, the counting process version of the likelihood has a particularly simple form, reducing to

$$L(\tau_1,\tau_2,\ldots,\tau_{N_T},w_0,w_1,\ldots,w_{N_T}) = \pi_{w_1}\prod_{i\neq j}\lambda_{ij}^{N_T(ij)}\exp[\lambda_{ij}d_T(i)],$$

where $d_T(i)$is the duration the individual spent in state i on $[0,T]$.

Fully observed multistate-model trajectories allow for multiple modeling options (see Meira-Machodo et al. (2009) for a review). In addition to fitting parametric Markov and semi-Markov models, it is possible to fit separate Cox models for each transition, which allows for flexible baseline hazard functions and proportional hazards covariate representations. This type of model is an inhomogeneous Markov model given an external time origin, or a semi-Markov model if the time origin is time since entry into a state. In limited contexts, such as illness-death models, it is also pos-

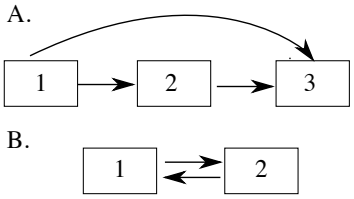

Figure 2.2: A. Example of model with forward transitions. Calculating $P_{13}(t,s|\mathscr{H}_t)$ requires integrating across trajectories that go through states $1 \to 2 \to 3$ and through states $1 \to 3$ B. Example of reversible model. Calculating $P_{12}(t,s|\mathscr{H}_t)$ requires integrating over an infinite number of possible state sequences, i.e., $1 \to 2$, $1 \to 2 \to 1 \to 2$, $1 \to 2 \to 1 \to 2 \to 1 \to 2$, etc.

sible to fit non-parametric non-Markov models to fully-observed data trajectories (Meira-Machado et al., 2006).

2.3 Partially observed data

2.3.1 Panel data with ignorable observation times

We now consider the partial observation scenario in Figure 2.1B, consisting of panel data of the disease process $W(t)$ observed at non-informative visit times. We assume the process is observed at discrete times $t_1, t_2, \ldots, t_n$ on $[t_1, t_n]$, with corresponding disease states $(w_1, \ldots w_n)$. The likelihood for these data is

$$L(w_1, \ldots w_n) = \pi_{w_1} \prod_{k=2}^{n} P_{w_{k-1}w_k}(t_{k-1}, t_k|\mathscr{H}_{t_{k-1}}). \tag{2.1}$$

Likelihood computations therefore require calculating transition probabilities $P_{ij}(t_{k-1}, t_k|\mathscr{H}_{t_{k-1}})$, which may be computationally challenging. The most tractable, but most restrictive, models for panel data assume that $W(t)$ follows a time-homogeneous CTMC. Such models have well-developed means of estimating parameters (Kalbfleisch and Lawless, 1985). To provide more flexibility by allowing hazards to vary with an external time scale, it is feasible to fit panel data with inhomogeneous CTMCs (Kay, 1986; Hubbard et al., 2008; Titman, 2011).

More difficulties arise when one wishes to assume a semi-Markov framework, allowing hazard

functions to vary with state occupancy duration. In general, options for these models are limited by the structure of the model and completeness of the observations. Data that are interval censored, such that transitions are observed but the times are unknown, permit semi-Markov model approaches, both parametric (Foucher et al., 2007) and non-parametric (Frydman and Szarek, 2009). True panel data, for which both transition times and states are unknown in the inter-observation interval, present more difficulties, since calculating transition probabilities requires integrating over all of the possible trajectories connecting states i and j on $[t_{k-1}, t_k]$. When the model consists only of forward transitions (Figure 2.2A), one must integrate over a finite set of potential sample paths, which is possible–for example, via Gaussian quadrature–if the model has a relatively small number of states (Foucher et al., 2010). Fitting Semi-Markov models with reversible transitions presents the most problems, since calculating transition probabilities requires integrating over an infinite number of possible state sequences (Figure 2.2B). Semi-Markov models for panel data for a general MSM with reversible transitions are only feasible if one assumes that some of the transitions are Markov (Kang and Lagakos, 2007).

Likelihood for panel data with misclassified disease observations

The next partial observation scenario we consider is panel data of the disease process $W(t)$ with non-informative visit times, but with potentially misclassified disease observations (Figure 2.1C). At the discrete time points, $(t_1, \ldots, t_n)$, we observe possibly misclassified states $o_1, \ldots, o_n$ rather than the underlying disease states $w_1, \ldots, w_n$. We will assume that the observed data has discrete state space R', which may or may not coincide with the state space of $W(t)$, R.

Models for multistate disease processes in the presence of misclassification error typically assume that observed data are independent conditional on the underlying disease process, and that o_k only depends on $W(t)$ through the disease state at time t_k, w_k. To compute the likelihood for misclassified panel data, one sums over the unobserved disease states. Therefore, the likelihood for panel data with non-informative sampling times and misclassifIcation error is

$$L(o_1, \ldots o_n) = \sum_{w_1} \ldots \sum_{w_n} \pi_{w_1} \prod_{k=2}^{n} \mathrm{P}_{w_{k-1} w_k}(t_{k-1}, t_k | \mathscr{H}_{t_{k-1}}) \prod_{k=1}^{n} \mathrm{P}(O_k = o_k | W_k = w_k). \tag{2.2}$$

When $W(t)$ is a Markov process, the joint model of $(w_1,\ldots,w_n,o_1,\ldots o_n)$ is a hidden Markov model (HMM) (Bureau et al., 2003); likewise, it is a hidden semi-Markov model when $W(t)$ is semi-Markov (Kang and Lagakos, 2007). We will discuss efficient methods for computing this likelihood via summing over the hidden states in Chapter 3, Section 3.4.

Panel data with non-ignorable sampling times

Finally, we turn to the partial observation settings consisting of discrete observations of the disease process that occur at observation times that depend on the individual's disease process, either without or with misclassification error (Figure 2.1 E,F). We assume the visit times are random, corresponding to observational data with symptom-based patient-initiated visits. The presence of informative observation times necessitates a joint modeling approach of the disease and observation process to avoid bias in estimating disease process parameters (Gruger et al., 1991).

The random visit time process can be characterized as a point process, a sequence of binary events that occur in continuous time (Daley and Vere-Jones, 2003). Equivalently, the visit time process can be represented as a counting process, $N(T)$, that counts the total number of informative observations in $[0,T]$. $N(t)$ has state space $\{0,1,2,\ldots,\infty\}$. We will assume that the disease process $W(t)$ and the observation process $N(t)$ cannot change states simultaneously. Jointly, the disease and observation time counting process $Y(t)=[W(t),N(t)]$ is a multistate process with a state space that is the Cartesian product of the disease states space R and the state space for $N(t)$,

$$\{(1,0),(2,0),\ldots,(r,0),(1,1)\ldots(r,1)\ldots(1,\infty),\ldots(r,\infty)\}.$$

Figure 2.1D provides an example of a complete joint disease and observation process trajectory.

In our partial observation scenario, the observed data consist of the informative observation process trajectory N(t) and the discrete observations of the disease process (Figure 2.1 E). We assume that the process is observed on $[0,T]$ with N_T total informative observation times, occurring at $(\tau_1,\ldots\tau_{N_T})$. Assuming the disease process is also observed at $t=0$ and $t=T$, the discrete observations of the disease process are $(w_0,w_1,\ldots,w_{N_T})$ We use the notation $\mathscr{H}_t$ to represent the history of

observations in the joint process $Y(t)$ up to time t. The joint likelihood for the observed data is

$$L(\tau_1,\tau_2,\ldots,\tau_{N_T},w_0,w_1,\ldots,w_{N_T},w_T) = \pi_{w_0,0}\,\mathrm{P}_{(w_0,0),(w_1,0)}(0,\tau_1|\mathscr{H}_0)h_{(w_1,0),(w_1,1)}(\tau_1|\mathscr{H}_{\tau_1})\times \left[\prod_{i=1}^{N_T-1}\mathrm{P}_{(w_i,i),(w_{i+1},i)}(\tau_i,\tau_{i+1}|\mathscr{H}_{\tau_i})h_{(w_{i+1},i),(w_{i+1},i+1)}(\tau_{i+1}|\mathscr{H}_{\tau_{i+1}})\right] \times \mathrm{P}_{(w_{N_T},N_T),(w_T,N_T)}(\tau_{N_T},T|\mathscr{H}_{\tau_{N_T}}), \tag{2.3}$$

where $\mathrm{P}_{(w_i,i),(w_{i+1},i)}(\tau_i,\tau_{i+1}|\mathscr{H}_{\tau_i})h_{(w_{i+1},i),(w_{i+1},i+1)}(\tau_{i+1}|\mathscr{H}_{\tau_{i+1}})$ corresponds to $W(t)$ transitioning between w_i and w_{i+1} between τ_i and τ_{i+1} and an observation event occurring at τ_{i+1}.

Should it be necessary to expand the model to allow for misclassification error in disease observations (Figure 2.1F), such that we observe $(o_0,o_1,\ldots,o_{N_T},o_T)$ rather than $(w_0,w_1,\ldots,w_{N_T},w_T)$, we can use an analogous approach to that for panel data with non-informative times (Section 2.3.1), making similar conditional independence assumptions about the relationship between observed and underlying disease trajectories. This allows us to express the the joint likelihood of the observed data and observation time process by summing the likelihood in Eq. 2.3 over the unobserved disease states, $w_0,w_1,\ldots,w_T$, such that

$$\begin{aligned} L(\tau_1,\tau_2,\ldots,\tau_{N_T},o_0,\ldots,o_{N_T}) = {} & \sum_{w_0}\ldots\sum_{w_n}\pi_{w_0,0}\,\mathrm{P}(O_0=o_0|W_0=w_0) \\ & \times \mathrm{P}_{(w_0,0),(w_1,0)}(0,\tau_1|\mathscr{H}_0)h_{(w_1,0),(w_1,1)}(\tau_1|\mathscr{H}_{\tau_1})\,\mathrm{P}(O_1=o_1|W_1=w_1) \\ & \times [\prod_{i=1}^{N_T-1}\mathrm{P}_{(w_i,i),(w_{i+1},i)}(\tau_i,\tau_{i+1}|\mathscr{H}_{\tau_i})h_{(w_{i+1},i),(w_{i+1},i+1)}(\tau_{i+1}|\mathscr{H}_{\tau_{i+1}}) \\ & \times \mathrm{P}(O_{i+1}=o_{i+1}|W_{i+1}=w_{i+1})] \\ & \times \mathrm{P}_{(w_{N_T},N_T),(w_T,N_T)}(\tau_{N_T},T|\mathscr{H}_{\tau_{N_T}})\,\mathrm{P}(O_T=o_T|W_T=w_T). \end{aligned}$$

Existing modeling approaches are limited for discretely observed multistate diseases observed at random informative observation times. Work in this area has largely assumed that visit times are pre-designated and missing in an informative fashion (Chen et al., 2010; Chen and Zhou, 2011). Models assuming random visit times have been developed for panel count data, a simple multistate process counting the accrual instances of a repeated event (He et al., 2009; Li et al., 2013), but are lacking for more general modeling frameworks, motivating our research in this area.

2.4 Conclusion

This chapter has provided an overview of MSMs and has described the likelihoods for complete and partial observation scenarios. Fitting MSMs to partially observed disease processes can be challenging due to difficulties in computing transition probabilities, particularly when the disease model contains reversible transitions. Moreover, transition probabilities appear in the likelihoods of all of our partial observation scenarios. These challenges provide motivation for the latent CTMC disease modeling framework described in upcoming chapters, as it will be based on computationally tractable Markov models that also allow for flexible, duration-dependent sojourn times.

Chapter 3

COMPUTATIONAL TOOLS FOR LATENT CONTINUOUS TIME MARKOV CHAIN MODELS

This chapter describes the conceptual and computational tools used in our latent CTMC models for partially observed disease processes. We begin by reviewing the properties of standard CTMCs and methods of computing transition probabilities. Then we introduce phase type (PH) distributions, which utilize a latent CTMC framework to model time-to-event variables and provide a flexible means of modeling sojourn time distributions. We then define latent, or aggregated CTMCs. To provide background for our joint disease and informative sample time model (Chapter 5). we review Markov-modulated Poisson processes (MMPPs). Finally, we discuss hidden Markov models (HMMs) and the recursive algorithms used for efficient likelihood calculations. Methods for HMMs have broad applicability in the latent CTMC disease modeling framework, and we will use them extensively in our estimation methods.

3.1 Time-homogeneous continuous time Markov chains

A continuous time Markov chain $X(t)$ is a continuous time multistate process on the discrete state space $S = \{1, 2, \ldots, s\}$ obeying the Markov property: observations only depend on the past history of the chain, $\mathscr{H}_t$, through the most recently occupied state. That is,

$$\mathrm{P}_{ij}[s, t+s|\mathscr{H}_t) = \mathrm{P}_{ij}[s, t+s|X(t) = i].$$

We focus on time-homogeneous CTMCs, where $\mathrm{P}_{ij}(s, t+s|\mathscr{H}_t) = \mathrm{P}_{ij}[t|X(0) = i]$, which we shorten to $\mathrm{P}_{ij}(t)$. Time-homogeneous CTMCs are parameterized by an initial distribution, $\boldsymbol{\pi}$, and an intensity matrix, $\boldsymbol{\Lambda}$, that does not depend on t. The entries λ_{ij} provide the instantaneous rate of transitions

between states i and j, i.e.,

$$\lambda_{ij} = \lim_{\Delta t \to 0} \frac{\mathrm{P}_{ij}(t+\Delta t)}{\Delta t}.$$

Our interest is limited to stable ($0 \leq \lambda_{ij} \leq \infty$) and conservative ($-\lambda_{ii} = \sum_{j \in S} \lambda_{ij}$) CTMCs.

Due to the constant hazard assumptions, state sojourn time distributions in homogeneous CTMCs have exponential distributions. That is, if $X(t) = i$, $\lambda_i = \sum_{j \in S} \lambda_{ij}$ is the constant hazard of leaving state i from time t, and the sojourn duration in state i has the distribution $\mathrm{Exp}(\lambda_i)$. Further, after the process transitions out of state i, it moves to state j with probability λ_{ij}/λ_i. This competing risks formulation provides an intuition for the evolution of CTMCs. It also suggests one can simulate a chain starting in state i by sampling an $\mathrm{Exp}(\lambda_i)$ random variable, then randomly choosing the chain's next state, then proceeding similarly for subsequent transitions (Gillespie, 1976).

Komolgorov's forward differential equation relates the transition probabilities to the intensity function (Kolmogorov, 1931). Suppose that $\mathbf{P}(t) = \{\mathrm{P}_{ij}(t)\}$ is the matrix representing transition probabilities in the interval [t,t+s]. The forward differential equation is

$$\frac{d\mathbf{P}(t)}{dt} = \lim_{\Delta t \to 0} \frac{\mathbf{P}(t) - \mathbf{P}(t+\Delta t)}{\Delta t} = \mathbf{P}(t)\mathbf{\Lambda}.$$

Combined with the initial condition $\mathbf{P}_{[t,t+s]} = \mathbf{I}$, the solution to the forward equation is the matrix exponential

$$\mathrm{P}(t) = \exp(\mathbf{\Lambda} t).$$

CTMC transition probabilities are thus computationally tractable, provided we have means of calculating the matrix exponential.

In practice, we have two preferred methods for computing the matrix exponential (Moler and Loan, 2003). If $\mathbf{\Lambda}$ is diagonalizable, we prefer to use the eigen-decomposition method. Diagonalizability means that we can express

$$\mathbf{\Lambda} = \mathbf{T}\mathbf{D}\mathbf{T}^{-1},$$

where $\mathbf{D} = diag(d_0, ...d_q)$ is the diagonal matrix of eigenvalues and $\mathbf{T}$ is a matrix of eigenvectors.

We can then compute the transition probabilities as

$$\mathbf{P}(t) = \mathbf{T}diag(e^{d_0 t}, ..., e^{d_q t})\mathbf{T}^{-1}.$$

If $\mathbf{\Lambda}$ is not diagonalizable, we prefer to use the uniformization method. Letting $\mu = max_i\{\lambda_{ii}\}$, we can obtain the discrete time Markov chain transition matrix that characterizes the sequence of states occupied by $X(t)$, given by $\boldsymbol{R} = \mathbf{\Lambda}/\mu + \boldsymbol{I}$. One can show that the number of jumps, including self-transitions, is distributed as Poisson(μt) (Ross, 1996). The transition probability can thus be calculated as the infinite sum

$$\mathbf{P}(t) = \sum_{n=0}^{\infty} \mathbf{R}^n \frac{(\mu t)^n}{n!} \exp(-\mu t).$$

One can approximate the transition probability by the finite sum

$$\mathbf{P}(t) = \sum_{n=0}^{m} \mathbf{R}^n \frac{(\mu t)^n}{n!} \exp(-\mu t).$$

Since each entry of $\boldsymbol{R}^n$ is on $[0,1]$, the truncation error for the finite approximation is bounded by the *Poisson*(μt) tail probability.

3.1.1 Discretely observed CTMCs as discrete time inhomogeneous Markov chains

A discrete time Markov chain X_t is a multistate Markov process observed at discrete epochs $t \in \{1, \ldots, n\}$, characterized by an initial distribution $\boldsymbol{\pi}$ and transition probabilities
$\mathrm{P}(X_i = x_i | X_{i-1} = x_{i-1})$ that govern the evolution of the chain. Homogeneous discrete time Markov chains have transition probabilities that are constant with respect to t, whereas inhomogeneous chains have transition probabilities that change over the observation period. Suppose a CTMC $X(t)$ is observed at fixed, but not necessarily equally spaced, sampling times $t_1, \ldots, t_n$, and
$X(t_1) = X_1, \ldots X(t_n) = X_n$. Then we can view X_t, $t \in \{1, 2, \ldots, n\}$, as a discrete time Markov chain, with initial distribution $\boldsymbol{\pi}$ and time-dependent transition probability matrix,

$$\mathrm{P}(X_i = x_i | X_{i-1} = x_{i-1}) = \exp\left[\mathbf{\Lambda}(t_i - t_{i-1})\right]_{x_{i-1} x_i}.$$

The perspective of a discretely observed CTMC as a discrete time Markov chain will prove useful in our estimation methods for models of discretely observed disease processes.

3.2 Phase-type sojourn distributions

We now turn our focus to phase-type models, the basis for disease state sojourn time distributions in the latent CTMC framework. PH distributions, introduced by Neuts (1995), are based on underlying time-homogeneous CTMC $X(t)$ with one or more absorbing states. PH models characterize the random variable $Y_k = inf\{t : X(t) = k\}$, the time X(t) enters absorbing state k. A key aspect of PH distributions is that Y_k is not exponentially distributed. Rather, its distribution is a mixture of distributions for time to absorption in state k via each of the possible paths to that state. The distribution corresponding to a particular path is hypoexponential, as it is characterized by sums of exponentially distributed sojourn times, each with a different rate. PH distributions are quite flexible: they are dense in the set of distributions of time-to-event random variables (Asmussen, 2003) and can be used to approximate distributions of time-to-event variables arbitrarily well, with accuracy of the approximation increasing with the dimension of X(t)'s state space.

PH distributions are parametrized via an initial distribution $\boldsymbol{\pi}$ (a row vector) and intensity matrix $\boldsymbol{\Lambda}$ for $X(t)$. We assume multiple absorbing states without loss of generality. The intensity matrix is given by

$$\boldsymbol{\Lambda} = \begin{bmatrix} \mathbf{T} & \mathbf{T}^0 \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{T}$ is an $m \times m$ sub-matrix characterizing transitions between transient states and $\mathbf{T}^0_k$ is the kth column of $\mathbf{T}^0$ providing the exit rates from transient states to absorbing state k. The sub-distribution functions for time to absorption in state k are given by

$$F_k(y|a) = \mathrm{P}(Y_k \leq y|X_0 = a) = \mathrm{P}_{[t,t+y]}[a,k] = \mathbf{e_a}(\exp(\mathbf{T}y) - \mathbf{I})\mathbf{T}^{-1}\mathbf{T}^0_k, \tag{3.1}$$

with corresponding densities

$$f_k(y|a) = \frac{\partial}{\partial y}\mathrm{P}(Y_k \leq y|X_0 = a) = \mathbf{e_a}\exp(\mathbf{T}y)\mathbf{T^0_k},$$

where $\mathbf{e_a}$ is row indicator vector for initial transient state a. The hazard of leaving through state k, given initial state a, is

$$\frac{\partial}{\partial y}\mathrm{P}(Y_k < y | X_0 = a, Y < y) = \frac{f_k(y|a)}{\sum_k F_k(y|a)}, \tag{3.2}$$

where $Y = min_k\{Y_k\}$ corresponds to the time of exit through any state. In PH models, the hazard of absorption by any state is asymptotically constant with respect to t, with exit rates equal to the negative of the dominant eigenvalue of $\mathbf{\Lambda}$, i.e., the unique eigenvalue whose non-zero real part is the smallest in absolute value (O'Cinneide, 1990).

In general, PH distributions are non-unique, in that distinct $(\boldsymbol{\pi}', \mathbf{\Lambda}') \neq (\boldsymbol{\pi}, \mathbf{\Lambda})$ may yield the same distributions for Y_k (O'Cinneide, 1989). Moreover, a PH distribution with general $\mathbf{\Lambda}$ and $\boldsymbol{\pi}$ is overparameterized: with a single absorbing state and k transient states, $\mathbf{\Lambda}$ and $\boldsymbol{\pi}$ have $k^2 - 1$ free parameters, but the distribution's Laplace transform has only $2k-1$ degrees of freedom (Cumani, 1982), indicating that there are in fact only $2k-1$ free parameters. However, even when a particular PH distribution is overparameterized, the sojourn distribution of Y_k, as well as related functionals, such as density and hazard functions, is well-defined and estimable even in the absence of unique $\mathbf{\Lambda}$ and $\boldsymbol{\pi}$ (see (Bladt et al., 2003) for a discussion).

Because general PH distributions are overparameterized, structured $(\boldsymbol{\pi}, \mathbf{\Lambda})$ with $2k-1$ parameters provide more parsimonious PH representations. Coxian PH distributions represent one such structure (Cumani, 1982) and will serve as our basic disease state sojourn time model in our latent CTMC disease models. Coxian PH distributions assume that $X(t)$ starts in state 1, and at transient state i, the process can either progress to $i+1$ or absorbing state k (Figure 3.1). Coxian distributions

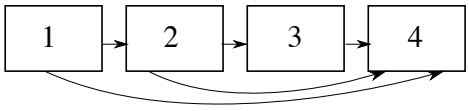

Figure 3.1: Transitions in a Coxian phase-type model with 3 transient states, $\{1,2,3\}$ and an absorbing state, 4. The model assumes the process starts in state 1.

have several compelling features. First, any PH distribution with an upper triangular $\mathbf{T}$ matrix has

a Coxian representation with the same latent state space dimension. The Coxian representation is minimal if the distribution of Y_k represented by $(\boldsymbol{\pi}, \boldsymbol{\Lambda})$ of dimension m cannot also be represented by $(\boldsymbol{\pi}', \boldsymbol{\Lambda}')$ with dimension $p < m$. Provided the Coxian representation has minimal dimension, the latent structure is uniquely parameterized. Coxian distributions are also quite flexible and, as with general PH distributions, can approximate arbitrary distributions of time-to-event random variables (Cumani, 1982).

Various methods have been developed for estimating model parameters for PH distributions. Asmussen et al. (1996) introduced an EM algorithm to obtain maximum likelihood estimates (MLEs) of PH model parameters, treating the "complete data" as the underlying unobserved trajectory of $X(t)$. Bladt et al. (2011) developed methods for obtaining estimated variances of MLEs based on the Fisher information. Estimation methods have also been developed in Bayesian contexts (Bladt et al., 2003; McGrory et al., 2009).

While PH models have their proponents in health science research (Aalen, 1995), their use is relatively limited, an exception being health care research modeling length of hospital visits (Fackrell, 2008; McGrory et al., 2009; Faddy et al., 2009). Despite the flexibility of PH models, their parametric framework and potential for overparameterization have likely limited their use in modeling survival or competing risks data in disease modeling contexts, particularly given compelling non-parametric or semi-parametric modeling alternatives.

3.3 Latent/aggregated continuous time Markov chains

Like PH models, latent, or aggregated, CTMCs are also based on an underlying CTMC, X(t), and specify that certain states cannot be differentiated by observers. In such models, rather than observing $X(t)$, we observe the process $V(t)$, formed by grouping states in $X(t)$'s state space S. Suppose the state space for $V(t)$ is R. Therefore latent CTMCS are formed by applying an onto (or surjective) function $u : S \to R$. Latent CTMCs are parameterized by $(\boldsymbol{\pi}, \boldsymbol{\Lambda}, u)$, where $\boldsymbol{\pi}$ and $\boldsymbol{\Lambda}$ are the initial distribution and intensity matrix of $X(t)$, respectively.

Our latent CTMCs assume that the states in S are ordered according to how they are grouped in R. That is, suppose $R = \{1, \ldots, r\}$; then $S = \{1_1, 1_2, \ldots, 1_{s_1}\} \cup \{2_1, 2_2, \ldots, 2_{s_2}\} \cup \cdots \cup \{r_1, r_2, \ldots, r_{s_r}\}$,

such that state $k \in R$ corresponds to states $\{k_1, k_2, \ldots, k_{s_k}\}$ in S, i.e., $u(k_i) = k$. Given the structure of S, the intensity matrix of X(t), $\mathbf{\Lambda}$, can be written as a partitioned matrix

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} & \mathbf{\Lambda}_{1r} \\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22} & \mathbf{\Lambda}_{2r} \\ \ldots & & \\ \mathbf{\Lambda}_{r1} & \mathbf{\Lambda}_{r2} & \mathbf{\Lambda}_{rr} \end{bmatrix}. \tag{3.3}$$

The structure is such that diagonal elements $\mathbf{\Lambda}_{ii}$ corresponds to transitions between $\{i_1, i_2, \ldots, i_{s_i}\}$, and that $\mathbf{\Lambda}_{ij}$ corresponds to transitions from $\{i_1, i_2, \ldots, i_{s_i}\}$ to $\{j_1, j_2, \ldots, j_{s_j}\}$. Moreover, the process $V(t)$ has phase-type sojourn time distribution, where, for all $i \in R$, $\mathbf{\Lambda}_{ii}$ corresponds to a PH distribution's $\mathbf{T}$ matrix.

As with PH distributions, latent/aggregated CTMCs may have non-unique representations, in that distinct $(\boldsymbol{\pi}, \mathbf{\Lambda}, u) \neq (\boldsymbol{\pi}', \mathbf{\Lambda}', u')$ may represent the same finite dimensional distributions for $V(t)$. Ryden (1996b) provides conditions whereby $(\boldsymbol{\pi}, \mathbf{\Lambda}) \neq (\boldsymbol{\pi}', \mathbf{\Lambda}')$ are equivalent, although in practice, these conditions are non-trivial to assess. That being said, it is possible to specify parsimonious, identifiable models, for example, by assuming that $V(t)$ has Coxian PH sojourn time distributions. This specification will result in a specific form for $\mathbf{\Lambda}$. With Coxian sojourn distributions for $V(t)$, the sub-matrix $\mathbf{\Lambda}_{ij}$, for $i \neq j$, is a $s_i \times s_j$ matrix whose first column represents the exit vector from state $i \in R$ to state $j \in R$ and has zeros everywhere else. Accordingly, $\boldsymbol{\pi}$, the initial distribution for the latent process $X(t)$, will be 0 except for entries corresponding to states $\{1_1, 2_1, ..., r_1\} \in S$.

3.3.1 Markov modulated Poisson processes

Our joint model for a discretely observed disease process and informative observation times, described in Chapter 5, assumes that the observation times follow a Markov modulated Poisson process, motivating its introduction here. Poisson processes are point processes that assume events cannot occur simultaneously, event counts in disjoint intervals are independent, and that the instantaneous rate of events is q, for a homogeneous process, and $q(t)$, for an inhomogeneous Poisson process (Ross, 1996). The counting process, $N(t)$, describing the accrual of events in the interval

$[0,t]$, is a CTMC with state space $\{0,1,\ldots,\infty\}$, instantaneous transition rates q (or $q(t)$, for the inhomogeneous case) between states i and $i+1$ and zero for all other pairs of states.

A MMPP is a doubly stochastic inhomogeneous Poisson process (Cox, 1955) with event rates that vary according to a time homogeneous CTMC (Freed and Shepp, 1982). Suppose $X(t)$ is the underlying CTMC, with initial distribution $\boldsymbol{\pi}$ and intensity matrix $\boldsymbol{\Lambda}$ and that $q(t)$ describes the Poisson rate of events over time. Under the MMPP assumptions $q(t) = q(X(t))$ is a piecewise constant function with rates that change with jumps in $X(t)$. That is, $q(t) = q_i$ when $X(t) = i$. Conditional on $X(t)$, event counts in disjoint intervals are independent; but unconditionally, this no longer holds.

Instead of above conditional construction of a MMPP, one can also consider the joint bivariate CTMC $Y(t) = [X(t), N(t)]$ (Ephraim, 2012). If $X(t)$ has state space $\{1,\ldots,s\}$, and $N(t)$ has state space $\{0,1,\ldots,\infty\}$, then the state space for $Y(t)$ is

$$S' = \{(1,0),(2,0),\ldots,(s,0),(1,1),\ldots(s,1),\ldots,(1,\infty),\ldots,(s,\infty)\}.$$

Suppose $\boldsymbol{Q} = diag(q_1,\ldots,q_s)$ is a diagonal matrix of Poisson event rates corresponding to the states in S, and that $\boldsymbol{\Lambda}$ is the infinitesimal generator for $X(t)$. We assume that jumps in $N(t)$ and $X(t)$ cannot co-occur, so $Y(t)$ has an intensity matrix

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{\Lambda}-\boldsymbol{Q} & \boldsymbol{Q} & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{0} & \boldsymbol{\Lambda}-\boldsymbol{Q} & \boldsymbol{Q} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Lambda}-\boldsymbol{Q} & \boldsymbol{Q} & \ldots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Accordingly, the first $\boldsymbol{\Lambda}-\boldsymbol{Q}$ block provides the instantaneous transition rates between state $(i,0)$ and $(j,0)$, and the first $\boldsymbol{Q}$ block provides rates of transitions between $(i,0)$ and $(i,1)$. The rest of the matrix has a similar interpretation.

In the MMPP setting, we typically observe $N(t)$, but the underlying CTMC, $X(t)$ is unobserved. Likelihood computations for $N(t)$ on the interval $[0,T]$ involve integrating the joint likelihood based on $Y(t)$ over the unobserved states of $X(t)$ at times when $N(t)$ jumps. We refer the reader to

Fearnhead and Sherlock (2006) for details of the likelihood calculations. Methods for maximum likelihood-based inference for MMPP parameters $(\boldsymbol{\pi}, \boldsymbol{\Lambda}, \boldsymbol{Q})$ use an EM algorithm, treating the full trajectory of $Y(t)$ as the complete data (Ryden, 1996a). Ryden (1996b) also demonstrates that $(\boldsymbol{\pi}, \boldsymbol{\Lambda}, \boldsymbol{Q})$ are identifiable provided that $X(t)$ is irreducible (all states are accessible from all other states) and the Poisson rates q_i are distinct, leading to consistency of MLEs.

3.4 Hidden Markov models

We now turn to a discussion of hidden Markov models. We provide a brief introduction here, and refer the reader to Cappe et al. (2005) for a more extensive exposition.

The HMM specification is as follows. Suppose that X_t is a discrete time Markov chain with state space $S = \{1, \ldots, s\}$ and initial distribution $\boldsymbol{\pi}$; we refer to X_t as the hidden layer, since it is typically unobserved. Suppose that each X_t is associated with a variable, O_t, which we refer to as the observed data. We assume that the observations $O_1, \ldots, O_n$ are independent of each other conditional on the underlying value X_t (Figure 3.2). Generally, O_t can be a continuous or discrete random variable; in our models, we assume O_t is discrete. The relationship between X_t and O_t is characterized by the emission matrix $\boldsymbol{E} = \{e(i, j)\}$ with entries $e(i, j) = \mathrm{P}(O_t = j | X_t = i)$.

Given the conditional independence structure, the likelihood for data consisting of underlying values $(x_1, \ldots, x_n)$ and observed values $(o_1, \ldots, o_n)$ is

$$\mathrm{P}(o_1, \ldots o_n, x_1, \ldots, x_n) = \pi_{x_1} \prod_{i=2}^{n} \mathrm{P}(X_i = x_i | X_{i-1} = x_{i-1}) \prod_{i=1}^{n} e(x_i, o_i).$$

The likelihood for the observed data only sums over $x_1, \ldots, x_n$:

$$\mathrm{P}(o_1, \ldots o_n, x_1, \ldots, x_n) = \sum_{x_1} \sum_{x_2} \ldots \sum_{x_n} \pi_{x_1} \prod_{i=2}^{n} \mathrm{P}(X_i = x_i | X_{i-1} = x_{i-1}) \prod_{i=1}^{n} e(x_i, o_i),$$

The summations over the hidden states make naive computations of the observed data likelihood very computationally expensive, of order $O(ns^n)$. The likelihood can be computed recursively, using forward and backward probabilities developed by Baum et al. (1970), which reduces the

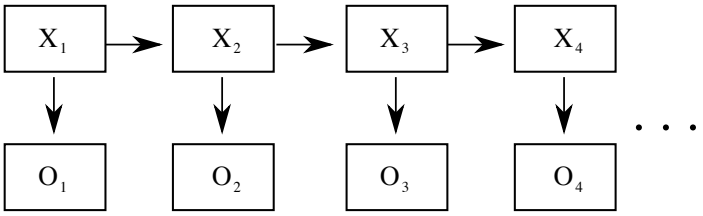

Figure 3.2: Graphical structure for a hidden Markov model, where $X_1, \ldots, X_n$ are the hidden data, and $O_1 \ldots, O_n$ are the observed data.

computation complexity of likelihood calculations to $O(ns^2)$. Using the notation $\boldsymbol{x}_{1:k} = (x_1, \ldots, x_k)$ and $\boldsymbol{o}_{1:k} = (o_1, \ldots, o_k)$ forward probabilities are defined as

$$\alpha_k(u) = \mathrm{P}(\boldsymbol{o}_{1:k}, X_k = u).$$

The forward probabilities are initialized as

$$\alpha_1(u) = \mathrm{P}(O_1 = o_1, X_1 = u) = e(u, o_1)\pi_{x_1},$$

and the recursion for $k = 2, \ldots, n-1$ is

$$\alpha_k(u) = \sum_i \alpha_{k-1}(i) e(u, o_k) \mathrm{P}(X_k = u | X_{k-1} = i).$$

Backward probabilities are defined as

$$\beta_k(u) = \mathrm{P}(\boldsymbol{o}_{k+1:n} | X_k = u).$$

They are initialized $\beta_n(u) = 1$ and have a backward recursion

$$\beta_k(u) = \sum_i \beta_{k+1}(i) e(i, o_{k+1}) \mathrm{P}(X_{k+1} = i | X_k = u),$$

for $k = 1, \ldots, n-1$. Via the forward or backward recursions, one can efficiently compute the ob-

served data likelihood. By the forward algorithm, the observed data likelihood is

$$\mathrm{P}(\boldsymbol{o}_{1:n}) = \sum_u \alpha_n(u).$$

By the backward algorithm, it is

$$\mathrm{P}(\boldsymbol{o}_{1:n}) = \sum_u \beta_1(u) e(u, o_1) \pi_{x_1}.$$

Other derived quantities for HMMs will also be of interest in our estimation methods. In particular, our information matrix calculations use recursive smoothing methods (Cappe et al., 2005), which require the filtering probability, $\mathrm{P}(X_t = j|\boldsymbol{o}_{1:t})$, and the conditional likelihood, $\mathrm{P}(O_t = o_t|\boldsymbol{o}_{1:t-1})$. Both of these quantities are obtained through a modified forward algorithm. The modified forward probabilities, described by Lystig and Hughes (2002), are given by

$$a_t(j) = \mathrm{P}(X_t = j, O_t = o_t|\boldsymbol{o}_{1:t-1}).$$

The recursion for modified forward probabilities is the following: initialize

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

and the recursion is

$$a_{k+1}(j) = \sum_{i \in S} \frac{a_k(i)}{\sum_{j \in S} a_k(j)} e(x_k, o_k) P(X_k = j|X_k = i).$$

The modified forward probabilities are related to the conditional likelihood via

$$P(o_t|\boldsymbol{o}_{1:t-1}) = \sum_{j \in S} a_t(j)$$

and the filtering probability via

$$P(X_t = j|\boldsymbol{o}_{1:t}) = \frac{a_t(j)}{\sum_{j \in S} a_t(j)}.$$

While the recursive tools for efficient computations in HMMs assume an underlying discrete time Markov chain, we will see that they apply considerably more broadly–to discretely observed CTMCs with misclassified observations, to aggregated/latent CTMCs, and even to MMPPs (Ryden, 1996a). Tools for HMMs will thus play an important role in estimation methods for our discretely observed disease process models. For example, the forward and backward probabilities will prove useful, not only in likelihood computations, but also in the expectation step of the EM algorithm we develop for fitting latent CTMCs to discretely observed data.

Chapter 4

LATENT CTMCS FOR PANEL-OBSERVED DISEASE PROCESSES WITH NON-INFORMATIVE OBSERVATION TIMES

4.1 Introduction

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. In this chapter, 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. This family of models enjoys tractable likelihoods and has established methods of obtaining 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, such as sojourn time CDFs or hazard functions, are typically 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 framework based on an underlying discretely observed CTMC, with or without misclassification error in the disease state observations. 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.

4.2 Model description

4.2.1 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 4.1A 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 4.1B). These restrictions induce sparseness in $\mathbf{\Lambda}$. Figure 4.1C 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 4.1A. 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

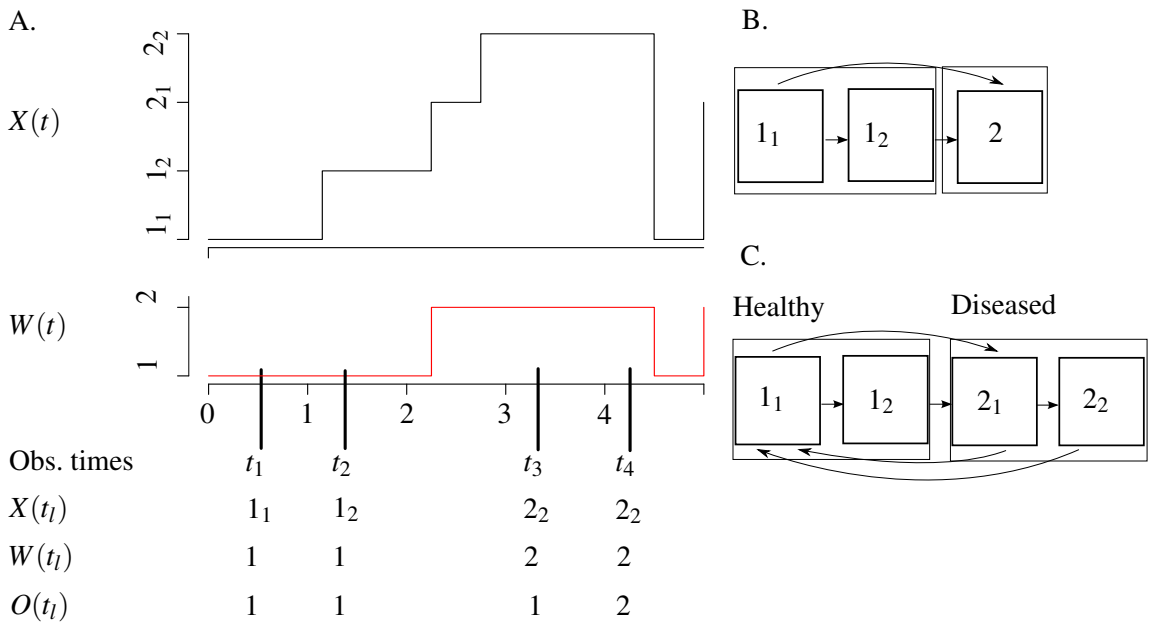

Figure 4.1: A. Example of latent trajectory $X(t)$, disease trajectory $W(t)$, and observed data $O(t_l)$ at discrete observation times for model in subfigure C, assuming possible misclassification error. B. 2-state survival model of W(t) assuming $R = \{1,2\}$ and $S = \{\{1_1,1_2\},\{2\}\}$, where disease state 2 is absorbing. The Coxian PH structure 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.

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.

4.2.2 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

$$\mathrm{P}(\mathbf{o}) = \sum_{x_1}\sum_{x_2}...\sum_{x_n}\pi_{x_1}\prod_{i=2}^{n}\mathrm{P}_{x_i x_{i+1}}(t_{i+1}-t_i)\prod_{i=1}^{n}e(x_i,o_i), \tag{4.1}$$

where $\mathrm{P}_{x_i x_{i+1}}(t_{i+1}-t_i) = \mathrm{P}\left(X(t_{i+1}) = x_{i+1} | X(t_i) = x_i\right)$ and $\pi_{x_1} = \mathrm{P}\left(X(t_1) = x_1\right)$. 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 (4.1). 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}.

4.2.3 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 $\log(\lambda_{ij}^h) = \beta_{ij}^T \mathbf{w}^h$, where h denotes the individual. In latent CTMCs, different constraints on covariate effects provide different interpretations. Adding the same covariate parameter to all latent transitions originating from disease state p, i.e., $\left\{\lambda_{ij} : i \in \{p_1, \ldots, p_{s_p}\}\right\}$, implies a multiplicative effect on the sojourn time in state p. To represent covariate effects on cause-specific hazard functions, one can add a separate covariate parameter for each transition out of disease state p to disease state r, i.e., $\left\{\lambda_{ij} : i \in \{p_1, \ldots, p_{s_p}\}, j \in \{r_1, \ldots, r_{s_r}\}\right\}$. This specification does not, however, represent a proportional hazards parameterization without additional non-linear constraints (Lindqvist, 2013).

One can also add covariates to emission and initial distribution parameterizations. Initial distributions and emission distributions are multinomial. Assuming S has s total states, the initial distribution $\boldsymbol{\pi}$ has natural parameters $\{\eta_i = \log(\pi_i/\pi_1)) : i = 2, ..., s\}$, and the emission distribution $\mathbf{e}_i$ has natural parameters $\left\{\eta_{ij} = \log(e(i,j)/e(i,1)) : j = 2, ..., g\right\}$. Subject-level covariates $\mathbf{w}^k$ are added to the multinomial models via a linear predictor, e.g., specifying $\eta_{ij}^k = \boldsymbol{\gamma}_{ij}^T \mathbf{w}^k$.

The choice to add covariates to different components of the model should be governed by scientific focus, keeping in mind model parsimony. The latter is a concern given that HMMs based on discrete time Markov models may not have identifiable covariate effects if the same covariates

are used to parameterize the emission and transition probabilities (Rosychuk and Thompson, 2004). Lack of identifiable covariate effects will be evident in a multimodal likelihood, detectable when fitting the model via our EM algorithm from multiple random starting values.

4.2.4 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 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

$$\begin{aligned} l(\boldsymbol{\theta};\mathbf{o},\mathbf{x}) &= l(\boldsymbol{\pi};x_1) + l(\boldsymbol{\Lambda};\mathbf{x}|x_1) + l(\mathbf{E};\mathbf{o}|\mathbf{x},x_1) \\ &= \sum_{i}^{s} z_i \log(\pi_i) + \sum_{i=1}^{s}\sum_{j\neq i} n_T(i,j)\log(\lambda_{ij}) - \sum_{i=1}^{s} d_T(i)\left(\sum_{j\neq i}^{s}\lambda_{ij}\right) \\ &+ \sum_{i=1}^{s}\sum_{j=1}^{r} o_T(i,j)\log\{e(i,j)\}. \end{aligned} \tag{4.2}$$

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.

4.3 EM algorithm

4.3.1 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 $\mathbf{\Lambda}$, $\boldsymbol{\pi}$ and $\mathbf{E}$ are provided in the appendix of this chapter.

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 $\boldsymbol{\pi}, \mathbf{\Lambda}$ 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).

4.3.2 E-step

The expectation step (E-step) requires computing the expectation of the complete data log-likelihood (5.2) conditional on the observed data. The log-likelihood for an individual is additive across time

intervals $T_l = [t_l, t_{l+1}]$. Hence,

$$\begin{aligned}\mathrm{E}[l(\boldsymbol{\theta};\mathbf{o},\mathbf{x})] = &\sum_{i=1}^{s}\mathrm{E}[z_i|\mathbf{o}]\log(\pi_i) + \sum_{l=1}^{n-1}\sum_{i=1}^{s}\sum_{j\neq i}\mathrm{E}[n_{T_l}(i,j)|\mathbf{o}]\log(\lambda_{ij}) \\ &- \sum_{l=1}^{n-1}\sum_{i=1}^{s}\mathrm{E}[d_{T_l}(i)|\mathbf{o}]\left(\sum_{j\neq i}\lambda_{ij}\right) + \sum_{l=1}^{n-1}\sum_{i=1}^{s}\sum_{j=1}^{r}\mathrm{E}[o_{T_l}(i,j)|\mathbf{o}]\log\left(e(i,j)\right).\end{aligned} \tag{4.3}$$

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 $\mathrm{P}(X_l = m|\mathbf{o}) = \frac{b_l(m)\alpha_l(m)}{\mathrm{P}(\mathbf{o})}$, where $\alpha_l(m)$ and $\beta_l(m)$ are HMM forward and backward probabilities, described in Chapter 3, and $\mathrm{P}(\boldsymbol{o})$ refers to equation (4.1). Hence

$$\mathrm{E}[z_i|\mathbf{o}] = \mathrm{P}(X_1 = i|\mathbf{o}) = \frac{\beta_1(i)\alpha_1(i)}{\mathrm{P}(\mathbf{o})}$$

and

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

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.

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

$$M_{ij}(t)[a,b] = \mathrm{E}[n_t(i,j)I(X_0 = a)|X_t = b]$$

and

$$H_i(t)[a,b] = \mathrm{E}[d_t(i)I(X_t = b)|X_0 = a]$$

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).

Computationally, it is efficient to calculate the joint expectations via matrix formulations. Specifically, we define $\boldsymbol{M}_{ij}(t) = \{M_{ij}(t)[a,b]\}$ and $\boldsymbol{H}_{ij}(t) = \{H_{ij}(t)[a,b]\}$. The formulae for the expectations are

$$\mathbf{M}_{ij}(t) = \int_0^t e^{\Lambda\tau}B_\alpha e^{\Lambda(t-\tau)}d\tau,$$

where $\mathbf{B}_\alpha$ is a $s \times s$ matrix that is 0 except for $\mathbf{B}_\alpha[i,j] = \lambda_{ij}$, and

$$\mathbf{H}_i(t) = \int_0^t e^{\mathbf{\Lambda}\tau}G_\alpha e^{\mathbf{\Lambda}(t-\tau)}d\tau,$$

where $\mathbf{G}_\alpha$ is a $s \times s$ matrix that is 0 except for $\mathbf{G}_\alpha[i,i] = 1$.

We consider two methods for calculating these integrals. When $\mathbf{\Lambda}$ is diagnonalizable, we use the method employed by Minin and Suchard (2008a,b), which relies on eigen-decomposition to calculate the integral. When $\mathbf{\Lambda}$ is not diagonalizable, we use uniformization to calculate the joint expectations. This method is described by Hobolth and Jensen (2011) and Bladt et al. (2011). To see how the uniformization method is derived, let $\mu = max_i\{\lambda_{ii}\}$ and $\mathbf{R} = \mathbf{\Lambda}/\mu + \mathbf{I}$, corresponding to the discrete time Markov chain transition probability matrix characterizing the state trajectory of $X(t)$. Given that the number of jumps, including self transitions, in time t is distributed as Poisson(μt),

we can compute the integral

$$\begin{aligned}\int_0^t e^{\mathbf{\Lambda} u}\mathbf{B}_\alpha e^{\mathbf{\Lambda}(t-u)}du &= \int_0^t \sum_{k=0}^{\infty} e^{-\mu u}\left(\mathbf{\Lambda}/\mu+\mathbf{I}\right))^k \frac{(\mu u)^k}{k!}\mathbf{B}_\alpha \sum_{n=0}^{\infty}\left(\mathbf{\Lambda}/\mu+\mathbf{I}\right)^n e^{-\mu(t-u)}\frac{(\mu(t-u))^n}{n!}du \\ &= e^{-\mu t}\sum_{k=0}^{\infty}\sum_{n=0}^{\infty}\frac{\mu^{k+n}}{k!n!}\left(\mathbf{\Lambda}/\mu+\mathbf{I}\right))^k\mathbf{B}_\alpha\left(\mathbf{\Lambda}/\mu+\mathbf{I}\right)^n\int_0^t u^k(t-u)^n du \\ &= e^{-\mu t}\sum_{k=0}^{\infty}\sum_{n=0}^{\infty}\frac{\mu^{k+n}}{k!n!}\left(\mathbf{\Lambda}/\mu\right)+\mathbf{I})^k\mathbf{B}_\alpha\left(\mathbf{\Lambda}/\mu+\mathbf{I}\right)^n t^{n+k+1}\frac{\Gamma(n+1)\Gamma(k+1)}{\Gamma(n+k+2)} \\ &= \sum_{s=0}^{\infty} e^{-\mu t}\frac{(\mu t)^{s+1}}{(s+1)!}\mathbf{A}(s),\end{aligned}$$

where $\mathbf{A}(s)=\sum_{n=0}^{s}\left(\mathbf{\Lambda}/\mu+\mathbf{I}\right)^n\frac{1}{\mu}\mathbf{B}_\alpha\left(\mathbf{\Lambda}/\mu+\mathbf{I}\right)^{s-n}$. This integral can be approximated by the finite sum

$$\sum_{s=0}^{m} e^{-\mu t}\frac{(\mu t)^{s+1}}{(s+1)!}\mathbf{A}(s).$$

As Bladt et al. (2011) point out, the resemblance of this sum to a shifted Poisson distribution and the fact that entries of $\mathbf{A}(s)$ are bounded on $[0,1]$ mean we can obtain an upper bound on the truncation error via Poisson tail probabilities. In practice, we set m such that

$$1-\sum_{s=0}^{m} e^{-\mu t}\frac{(\mu t)^{s+1}}{(s+1)!} < 10^{-}8,$$

bounding the truncation probability by $10^{-}8$.

Conditional joint second and cross moments of occupancy durations and transition counts

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
$U_{ijlm}(t)[a,c]=\mathrm{E}[n_t(i,j)n_t(l,m)I(X_t=c)|X_0=a]$; $W_{ij}(t)[a,c]=\mathrm{E}[d_t(i)d_t(j)I(X_t=c)|X_0=a]$; and $V_{ilm}(t)[a,c]=\mathrm{E}[d_t(i)n_t(l,m))I(X_t=c)|X_0=a]$. Details for these computations using eigen-decomposition are provided by Minin and Suchard (2008b), and using uniformization by Hobolth and Jensen (2011).

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

Joint first and second moments are also desired when the interval endpoint coincides with the time of absorption, Y, as is the case when the final observation coincides with the patient's death. 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).

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 previously defined joint moments, $M_{ij}(t)[a,b]$, $H_i(t)[a,b]$, $U_{ijlm}(t)[a,c]$, $W_{ij}(t)[a,c]$, and $V_{ilm}(t)[a,c]$.

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

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

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

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

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 4.2.2).

4.4 Recursive smoothing for complete data sufficient statistics

4.4.1 Motivation

Our E-step calculates conditional expectations of the complete data likelihood 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. 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 **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.

To further motivate the recursive smoothing method, we consider the alternative of directly computing the second moment of the complete data likelihood split up by time intervals (Equation (4.3)), conditional on the all observed data. This approach requires computing cross terms across different time intervals–e.g., $E[d_{T_k} d_{T_l}|\boldsymbol{o}]$ for intervals T_k and T_l–yielding a computational complexity of at least $O(n^2)$. Moreover, obtaining this expectation via conditioning requires

$$E[d_{T_k} d_{T_l}|\boldsymbol{o}] = E\left[E[d_{T_k} d_{T_l}|x_k, x_{k+1}, x_l, x_{l+1}|\boldsymbol{o}]\right],$$

which means we need to compute

$$\mathrm{P}(X_l = a, X_{l+1} = b, X_k = c, X_{k+1} = d|\boldsymbol{o}).$$

Unlike the bivariate smoothing probabilities used for the first moment calculations, this quantity cannot be directed calculated from the forward and backward functions and in fact requires summing over all hidden states between T_l and T_k.

4.4.2 First moments of complete data sufficient statistics

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

$$t_{k+1}(\mathbf{x}_{1:k+1}) = m_k(x_k, x_{k+1}) t_k(\mathbf{x}_{1:k}) + s_k(x_k, x_{k+1}), \tag{4.4}$$

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

$$\begin{aligned} \tau_{k+1}(x_{k+1}) = \frac{\mathrm{P}(\mathbf{o}_{1:k})}{\mathrm{P}(\mathbf{o}_{1:k+1})} \Big\{ \sum_{x_k} \Big[& \tau_k(x_k) m_k(x_k, x_{k+1}) \\ & + \mathrm{P}(X_k = x_k|\mathbf{o}_{1:k}) s_n(x_k, x_{k+1}) \Big] \times e(x_{k+1}, o_{k+1}) \mathrm{P}_{x_k x_{k+1}}(t_{k+1} - t_k) \Big\}. \end{aligned} \tag{4.5}$$

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 Chapter 3, Section 3.4.

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

$$t_{k+1}(\mathbf{x}_{1:k+1}) = \mathrm{E}[\mathbf{S}[t_1,t_{k+1}]|\mathbf{x}_{1:k+1}] = \mathrm{E}[\mathbf{S}[t_1,t_k]|x_{1:k}] + \mathrm{E}[\mathbf{S}[t_k,t_{k+1}]|x_k,x_{k+1}] = t_k(\mathbf{x}_{1:k}) + s_k(x_k,x_{k+1}). \tag{4.6}$$

Here, $m_k(x_k,x_{k+1}) = 1$. The specific values of $t_1(x_1)$ and $s_k(x_k,x_{k+1})$ are 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 $o_T(i,j)$.

4.4.3 Second moments

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. 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

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

The first component is identical to first moment recursive function (equation (4.6)); 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. Table 4.1 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: $\tau(x_k) = \left\{ \tau^{(1)}(x_k), \tau^{(2)}(x_k) \right\}$. The updates for $\tau(x_k)$ follow from a multivariate version of equation (4.5). The $\tau^{(1)}(x_k)$ component is defined as in equation (4.5). 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 4.1: Definition of recursive smoothing quantities $s_k^{(2)}(x_k, x_{k+1})$ and $t_1^{(2)}(x_1)$ for second moment calculations.

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

4.5 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's formula (Louis, 1982) as

$$-\ddot{l}(\boldsymbol{\theta}; \mathbf{o}^m) = \mathrm{E}[-\ddot{l}(\boldsymbol{\theta} \mid \mathbf{o}^m] - \mathrm{Cov}[\dot{l}(\boldsymbol{\theta} \mid \mathbf{o}^m] = \mathrm{E}[-\ddot{l}(\boldsymbol{\theta}) \mid \mathbf{o}^m] - \Big\{ \mathrm{E}[\dot{l}(\boldsymbol{\theta}) \dot{l}(\boldsymbol{\theta})^T \mid \mathbf{o}^m] - \mathrm{E}[\dot{l}(\boldsymbol{\theta}) \mid \mathbf{o}^m] \mathrm{E}[\dot{l}(\boldsymbol{\theta}) \mid \mathbf{o}^m]^T \Big\}.$$

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 $\mathrm{E}[-\ddot{l}(\boldsymbol{\theta}) \mid \mathbf{o}^m]$ readily given the factorization of the log likelihood (5.2) and the relatively simple forms for Hessian functions (Chapter 4 Appendix) for $\boldsymbol{\pi}$, $\boldsymbol{\lambda}$ and $\mathbf{E}$. At the MLE, $\mathrm{E}[\dot{l}(\boldsymbol{\theta}) \mid \mathbf{o}] = \mathbf{0}$, so we only need to calculate $\mathrm{E}[\dot{l}(\boldsymbol{\theta}) \dot{l}(\boldsymbol{\theta})^T \mid \mathbf{o}^m]$. 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 5.5.3 and use recursive smoothing to integrate over latent states.

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). 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 $\exp(\boldsymbol{\Lambda}(\psi)t)$ with respect to entries of $\boldsymbol{\psi}$. These derivatives involve similar integrals as first moments of occupancy durations and transition counts and are computed with similar methods (Najfeld and Havel, 1994). For example, consider the functional $\mathrm{P}_{ij}(t,\boldsymbol{\psi}) = \exp(\boldsymbol{\Lambda}(\boldsymbol{\psi})t)_{ij}$. Then $\frac{\partial \mathrm{P}_{ij}(t,\psi)}{\partial \psi[k]}$ is the i,j entry of the matrix given by

$$\int_0^t e^{\Lambda(\psi)\tau} \mathbf{B}_{\psi[k]} e^{\Lambda(\psi)(t-\tau)} d\tau,$$

where $\mathbf{B}_{\psi[k]} = \{B_{\psi[k]}(i,j)\}$ and $B_{\psi[k]}(i,j) = \frac{\partial \lambda_{ij}(\psi)}{\partial_{\psi[k]}}$.

4.6 Implementation

4.6.1 Software

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).

4.6.2 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 six without substantial loss of robustness.

4.6.3 Practical considerations for using the EM algorithm

EM algorithms will converge to local maxima, global maxima, or stationary points (Wu, 1983). Latent parameter models are frequently multi-modal 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(μ=0, σ=.25), but it is worth experimenting with different starting distributions for specific models and data sets.

With discretely observed data, MLEs with finite entries for $\mathbf{\Lambda}$ may not exist (Bladt and Sorensen, 2005). 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, non-existence 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.

4.6.4 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. While other latent structures are possible, we are advocating models with disease state sojourn distributions characterized by Coxian PH structure, since these models can represent distributions with increasing, decreasing, and non-monotonic hazard functions and will be uniquely parameterized except in degenerate situations (Cumani, 1982).

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 Bayesian Information Criterion (BIC), since it is easy to obtain, can be used to compare non-nested models, and has been shown to be adequate in choosing the number of mixture components (Steele and Raftery, 2010). In Chapter 5, Section 5.6.7, we will provide validation for the BIC for latent CTMC models in the context of informative sampling times, and it is reasonable to assume the method is also valid for choosing latent CTMC models in the context of data with non-informative observations.

Generally, 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.

4.7 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 (RMSE) 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

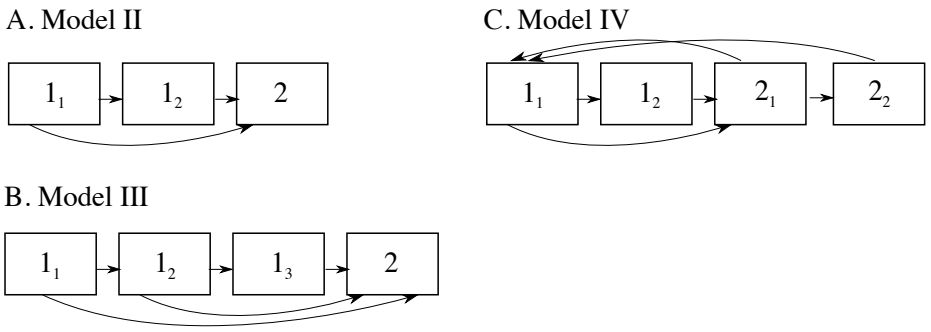

Figure 4.2: Models fit to simulated data. The numeral designation reflects the number of transient latent states in the model. A. Model II, a Coxian survival model with 2 transient latent states. B. Model III, a Coxian survival model with 3 transient latent states. C. Model IV, fit to data from the reversible disease model, with 2 transient latent states per disease state.

errors.

Data were generated for 2-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. 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; 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 using latent CTMC models with Coxian PH distributions (Figure 4.2). 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 2-state reversible model assuming sojourn distributions with two latent states, analogous to model II. The numeral designation reflects the total number of transient states in the model. 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. Hazard and CDFs of sojourn distributions were estimated for each dataset using the corresponding models.

The fitting method encountered numeric problems for 4.7% (232/5000) of the starting values across all of the 500 datasets used in the study. Eighty-five percent of the problems occurred when fitting model III to the survival data generated from the increasing hazard Weibull(1.5,1) distribution. Problems encountered in fitting appeared to be due to lack of robustness of the SQUAREM acceleration methods in certain regions of the parameter space. After the fact, spot checks revealed that replacing the accelerated EM function with the traditional version alleviated the fitting problems.

We were interested in the performance of the MLEs for the fitted latent CTMC models and did not want to analyze results if the estimates reflected a local maximum. To make sure we were confident in attaining the MLEs, we assessed final log-likelihoods attained from each of the starting values; if only one starting value converged to the maximum among all final log-likelihoods, we excluded the estimates from our results. We therefore limited our analysis to the 96% (481/500) of the simulated datasets, all of which had more than one starting value that converged to the putative maximum log-likelihood. 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%).

4.7.1 Bias in approximations of hazard and cumulative distribution functions

Results relevant to the investigation of bias are depicted in Figure 4.3A and B, depicting the mean and bias of CTMC estimates used to approximate the Weibull (1.5,1) and Weibull (.75,10) hazard and CDF functions from the different latent CTMC models. The x-axis is sojourn time, and x-limits were chosen to zoom in on the early portion of the sojourn time period. The bias in approximations reflects the closeness of the data-generating distribution to that of the latent CTMC model as well as the functional to be estimated. 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.

Interestingly, discrete sampling schemes also affected bias in estimates of both hazard functions

and CDFs. We expected that the mean of model II and IV estimates of hazard and CDFs would be similar given that they assume the same latent CTMC model for each disease state sojourn time, but that model IV estimates would be more variable due to the discrete samples providing less information. This was in fact true for estimates of Weibull(1.5,1) hazard functions and CDFs. However, estimates of Weibull(.75, 10) hazards and CDFs based on discretely sampled data were more biased, in particular in the early part of the hazard function. We suspected the bias we observed was related to the frequency of the sampling scheme relative to the rate of change of the hazard function and that the bias would be mitigated by more closely-spaced observations. To investigate this hypothesis, we generated discrete observations of the reversible process at more closely-spaced sampling schemes: 2 and 4 observations/unit-time, versus the original simulation's frequency of 1 observation/unit-time. In fact, the bias we observed did decline with more closely spaced observations (Figure 4.7.1). The most closely-spaced observation scheme yielded hazard estimates that were no more biased than those based on fitting survival data with Model II. Furthermore, the estimates from more closely-spaced observations were also less variable (Figure 4.3).

We also expected there would be a bias-variance tradeoff to adding more latent states to the latent CTMC model. In fact, model III (with 3 latent states), did have less biased estimates of hazard and CDF relative to model II. Model III estimates did have somewhat higher variance (not shown); and overall, the RMSE of the estimates (Figure 3C in the main text) from model II and III was quite similar. The one exception was for the tail end of the Weibull(1.5, 1) hazard function, when model III's estimates were considerably less biased. Overall, on the basis of the RMSE of point estimates, there is little to recommend model III over model II. We expect that adding more states to the model (e.g., 4 versus 3) would yield more variable estimates, and the RMSE would favor models with III states. This was borne out by limited investigations with such models (results not shown).

4.7.2 Performance of delta-method standard errors and confidence interval coverage

Our investigations of 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 (Figure 4.5). Generally, delta-method standard errors from model III better reflected estimate

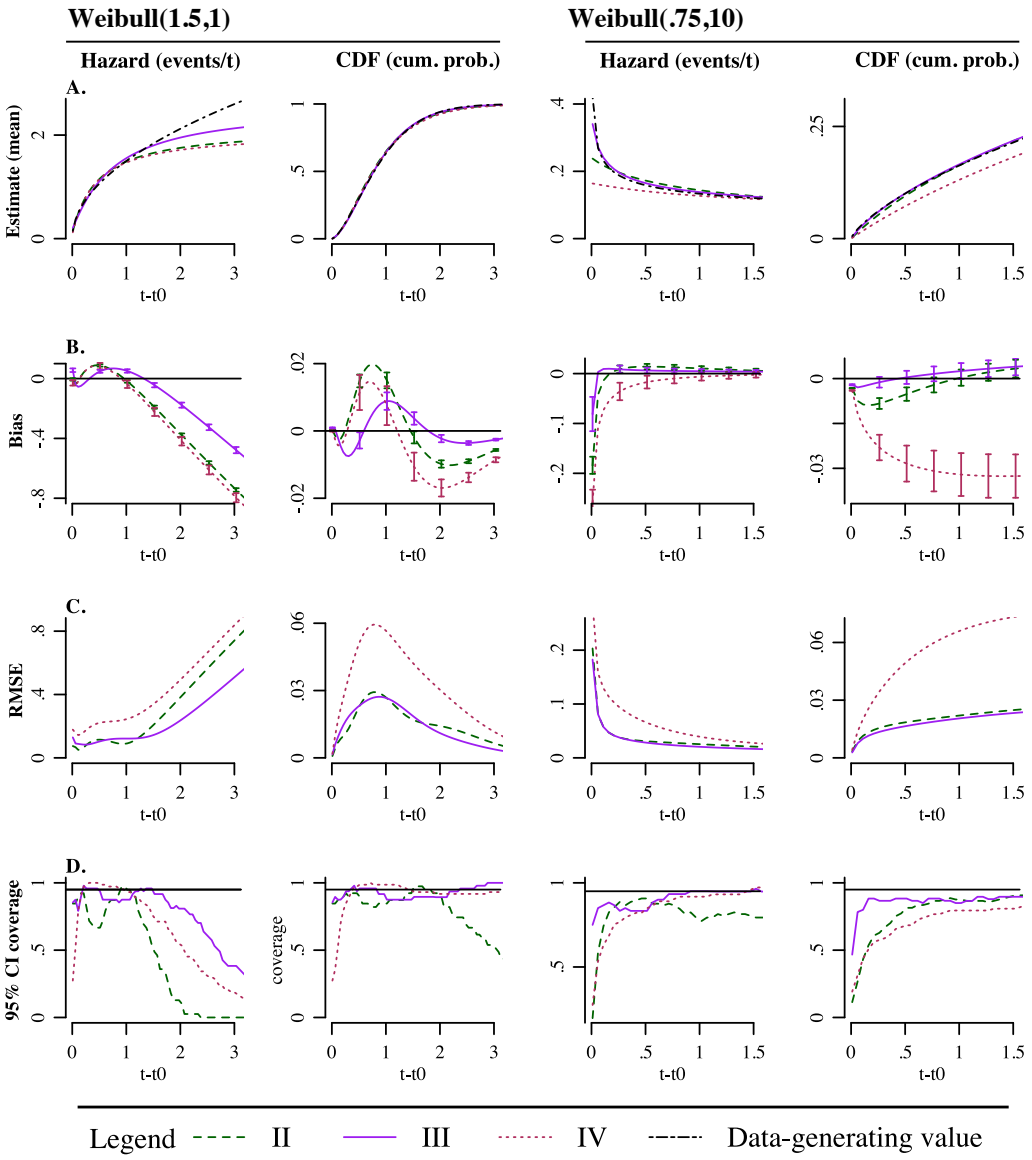

Figure 4.3: Summary of estimates of CDFs and hazard functions based on models fit to data generated from Weibull(1.5,1) and Weibull(.75,10) sojourn distributions. Models II and III fit survival data with Coxian PH models with 2 and 3 transient states, respectively; Model IV fits discretely observed data from a 2-state reversible model assuming sojourn distributions analogous to model II (Figure 4.2). 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.

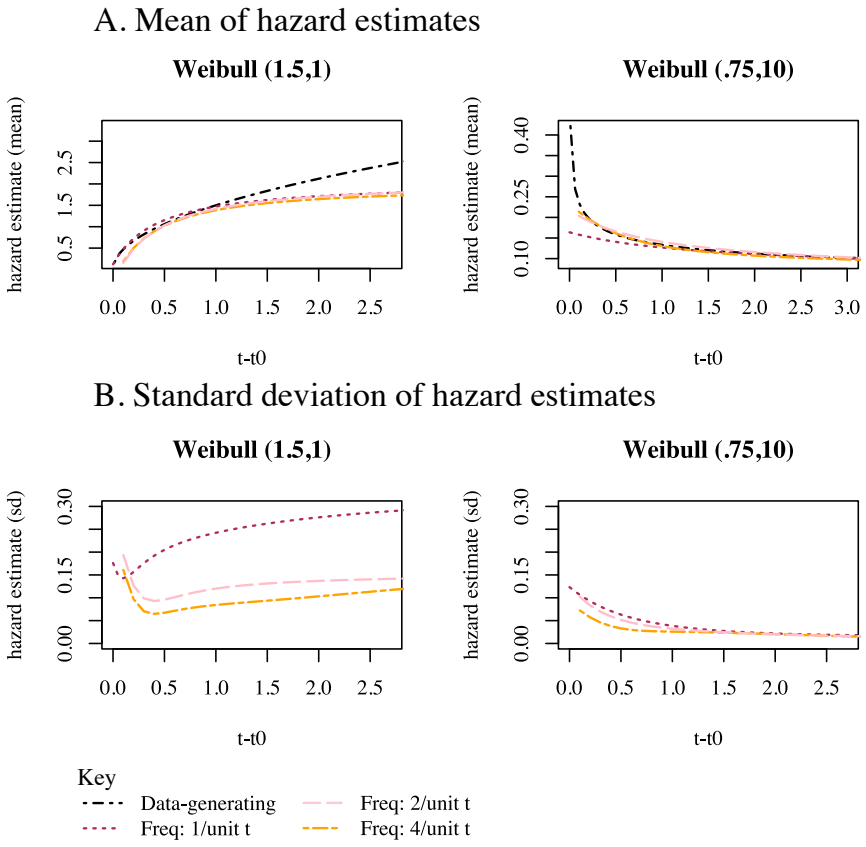

Figure 4.4: Mean and standard deviation of hazard estimates from latent CTMC Model IV (Figure 4.2 C) fit to data generated from the reversible disease model, with varying frequency of discrete observations.

variability than model II.

Coverage of 95% confidence intervals based on delta-method standard errors is shown in Figure 4.5D Again, performance was quite mixed. 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. 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 underestimated the true variability of the estimates (Figure 4.5), 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.

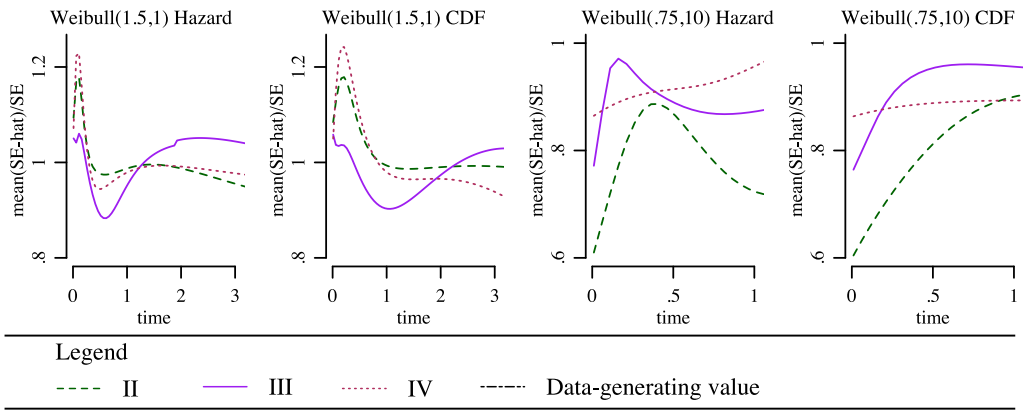

Figure 4.5: Ratio of average delta-method standard errors to the empirical standard errors of the estimates from simulated data. Models II and III fit survival data with Coxian PH models with 2 and 3 transient states, respectively; Model IV fits discretely observed data from a 2-state reversible model assuming sojourn distributions analogous to model II (Figure 4.2).

4.8 Application: bronchiolitis obliterans syndrome in lung transplant patients

4.8.1 Motivation

We now turn to the real data application presented in (Titman and Sharples, 2010). Following lung transplantation, patients are at risk of developing bronchiolitis obliterans syndrome (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 (Estenne et al., 2002). Titman and Sharples (2010) developed 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. Our objective was to use the model of Titman and Sharples (2010) on the BOS data as a means of comparing our EM estimation method to standard numeric optimization and to examine in more detail the implications of the model results for scientifically describing the disease process, which Titman and Sharples (2010) only cursorily explored.

4.8.2 Dataset

The BOS study and dataset is described in full by Jackson et al. (2002) and was obtained via communication with Andrew Titman. The dataset consisted of 122 double lung and 244 heart-lung transplant patients. For the purpose of analysis, we excluded 43 individuals with only baseline observations, since our software implementation of the EM algorithm requires at least one follow-up observation per person.

4.8.3 BOS model

The model of Titman and Sharples (2010) assumes that the BOS disease process, $W(t)$, has a state space with 3 states: $R = \{1 = \text{"}healthy\text{"}, 2 = \text{"}BOS\text{"}, 3 = \text{"}death\text{"}\}$, where death is absorbing. $W(t)$ 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 $W(t)$. The transitions depicted in the model

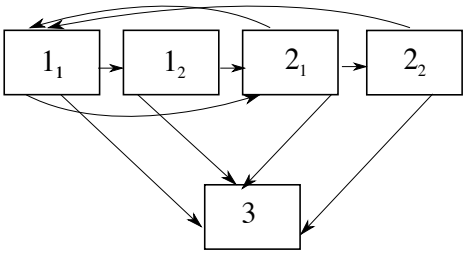

Figure 4.6: The BOS model used by Titman and Sharples (2010). States 1_1 and 1_2 map to healthy states; 2_1 and 2_2, to BOS states; and 3, to death.

are shown in Figure 4.8.3, and we refer the reader to Titman and Sharples (2010) for a discussion of their model selection. However, in Section 4.8.6, we will revisit their decision to include a transition from the BOS state back to the healthy state despite the irreversible nature of the BOS disease process.

The specifics of the parameterization of Titman and Sharples (2010) are as follows. 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. This specification can be expressed on the log-rate scale as

$$log(\lambda_{1_2 2_1}) = log(\tau_1) + log(\lambda_{1_1 2_1}),$$

which means we can specify the model according to our framework via parameter constraints and dummy covariates.

The emission model includes transplant type in the probability of misclassification of healthy patients as diseased, such that

$$logit(e(Healthy, BOS)) = \gamma_0 + \gamma_1 \times Z_{DL},$$

where Z_{DL} is an indicator of double lung transplant. Misclassification of diseased patients as healthy

does not depend on covariates, such that

$$logit(e(BOS, Healthy)) = \nu_0.$$

The initial distribution is such that individuals occupy either state 1_1 or 2_1 with a probability depending on transplant type, according to the parameterization

$$logit(\pi_{2_1}) = B_0 + B_1 \times Z_{DL}.$$

4.8.4 Comparison between our EM and other optimization methods

We now discuss experiments aimed at comparing the performance of our EM algorithm with other optimization methods using the BOS dataset. 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) 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). Based on preliminary experiments, we expected that our method, EM1, would outperform the other optimization methods that use more generic approaches, at least in terms of speeed. NM is a direct search method and is known for being robust but slow to converge to a stationary point (Lagarias et al., 1998). BFGS is a quasi-Newton method that requires numeric differentiation, which may be unstable in some regions of the parameter space. For general likelihoods, global convergence is not guaranteed for any method. We did not have an a priori hypothesis about which methods were more likely to obtain the global maximum of the log-likelihood.

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$(0, \sigma = .25)$. 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 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

Table 4.2: 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	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

the BFGS stopping criteria based on a relative convergence tolerance of 10^{-3}. We accelerated both EM algorithms by SQUAREM (Varadhan, 2011).

The performance of each of the algorithms for both BOS data models is summarized in Table 4.2. Runtime, either in time to convergence or to the maximum number of iterations, and the final value of the log-likelihood are shown in Figure 4.7. Our method, EM1, was the clear winner in terms of runtime, taking a median of 80 seconds to converge when π and $\mathbf{E}$ are unknown. Other methods ran between 5.5 to 18 minutes 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 non-differentiable 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 log-likelihood for the BOS data model was -1,248.602. 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 $(\pi, \mathbf{E})$, were unknown EM1 converged to local optima for 18/30 trials, EM2 in 16/30 and BFGS in 10/30 trials. In the scenario where $(\pi, \mathbf{E})$ were known, EM1 converged to the global maximum for all but one starting value.

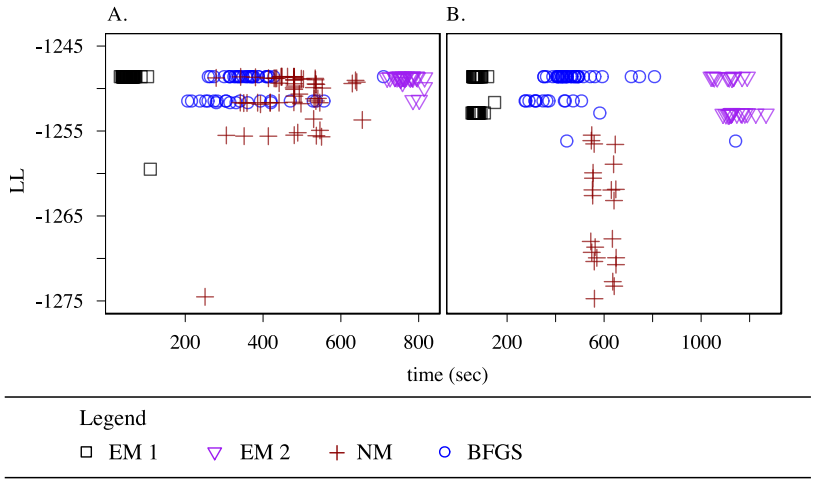

Figure 4.7: Runtime and attained log-likelihood (LL) when EM1 (our method), EM2, BFGS, and NM algorithms were used to fit the BOS data, using 30 random starting values and assuming either $(\mathbf{E}, \boldsymbol{\pi})$ was fixed (A), or was unknown (B).

4.8.5 BOS results

We now turn to our re-examination of results of fitting the model for scientific understanding of the BOS disease process. In our dataset, 64% (202/316) had at least one positive BOS test, and 50% (159/316) died. After fitting the model, we found that parameter estimates are similar, but not identical to those obtained by Titman and Sharples (2010), due to our exclusion of individuals with single observations from the dataset. As with the results reported by Titman and Sharples (2010), the MLEs we obtained were evidently unique, based on the 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 Table 4.3.

The first passage CDF for BOS development is depicted in 4.8A. The model estimates that the probability of an initially 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 5 year disease free probability ranging from 15% to 37% (Chan and Allen, 2004). The model also predicts that the rate of entry into the diseased states declines with time since transplant; disease rates

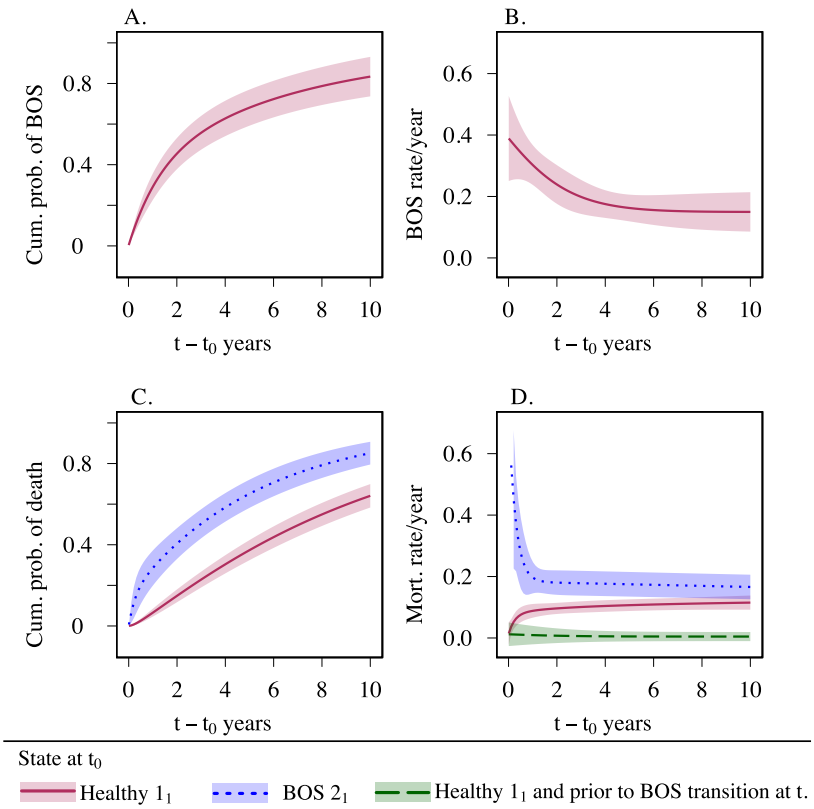

Figure 4.8: A. Cumulative probability of having transitioned to BOS state at least once, conditional on being in 1_1 at t_0. B. Disease rate conditional on being in healthy state 1_1 at t_0. C. Cumulative probability of death. C. Mortality rate per year, as a function of state at t_0. In all figures the shaded regions represent 95% point-wise confidence intervals for the estimates.

Table 4.3: 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

are initially 35-40% and drop to 15% per year after 5 years (Figure 4B). The non-constant 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 (Jackson et al., 2002).

The cumulative probabilities of death conditional on starting in healthy state 1_1 versus BOS state 2_1, is shown in Figure 4C. 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 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 (Copeland et al., 2010).

The 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 one 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 (Lama et al., 2007; Jackson et al., 2002).

4.8.6 Investigation of justification for BOS to healthy transitions in disease model

Although, biologically, BOS is an irreversible disease, Titman and Sharples (2010) decided to include a BOS $\rightarrow$ healthy transition in their latent CTMC model. Our estimates of these rates were relatively low and declined with disease duration: upon BOS onset, the rate of reversion is initially 6%/year, dropping to 1.6%/year after one year of having BOS. In our view, a more thorough investigation of the utility of the reversible transition was warranted.

The decision to include the BOS $\rightarrow$ healthy transition arose during the course of model building. Titman and Sharples (2010) describe starting their model development by assuming standard CTMC disease models and finding via a likelihood ratio test that adding a BOS $\rightarrow$ healthy transition offered a significant improvement in fit. They retained the transition when adding latent structure to the model. One question we had was whether there was still statistical evidence for including a reversible transition within a latent disease framework. To that end, we compared our model to two alternative models with no BOS $\rightarrow$ healthy transitions. Model 1 had two latent states per healthy state, and Model 2 had 3 (Figure 4.9). Other aspects of these models were identical to the original model.

As an additional test, not reported by Titman and Sharples (2010), we compared Model 1 to the original model via a likelihood ratio test. These models are nested and only differ in that the original model contains reversible healthy to BOS transitions. The null hypothesis (the transition between BOS to healthy states has a rate of zero) is on the boundary of the parameter space; thus the distribution of the likelihood ratio test statistic is not a standard chi-square distribution, but rather a 50:50 mixture of a chi-square with 1 df and a point mass at zero (Self and Liang, 1987). The p-value for the LR test statistic was .006, which supports the hypothesis that the reversible transition

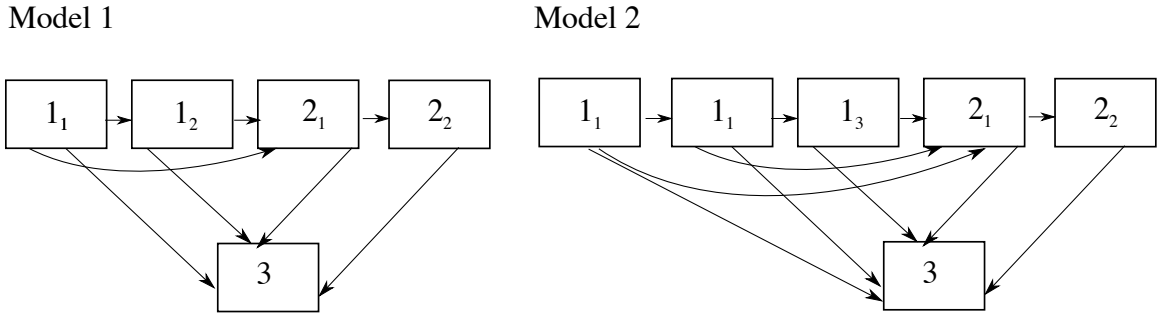

Figure 4.9: Alternative models fit to BOS data, without BOS → healthy transitions

Table 4.4: Summary of models fit to BOS data with and without BOS → healthy transitions

	BOS → Healthy transition	No. Healthy latent states	No. param.	Log likelihood	BIC
Original model	Yes	2	13	-1248.602	2572.03
Model 1	No	2	12	-1251.757	2572.58
Model 2	No	3	11	-1251.389	2583.36

improves the fit within the latent disease framework.

We were also interested whether Model 2, with three latent BOS states but no reversible transitions, might be able to fit the data as well as the original model, given Model 2's additional latent state. In fact, Model 2's fit was nearly identical to that of Model 1, in terms of the log-likelihood (Table 4.4). Moreover, there was also evidence that the intensity parameters in Model 2 were not uniquely identifiable: the information matrix was not positive definite at the maximum log-likelihood, and two separate runs of the algorithm from different starting values generated distinct sets of intensity rate estimates at log-likelihoods that differed by $< .3$.

Finally, we can compare all three models via the BIC. On the basis of lower BIC reflecting better fit, the original model is preferred, although the difference in BIC between the original model and Model 1's is $< .5$ (Table 4.4). This suggests that it is possible that the original model with reversible transitions may be over-fitting the data. Indeed, plotting the estimated disease rate and first passage distribution to BOS onset under the original and two alternative models (Figure 4.10) shows only minor differences in the estimates, suggesting that the statistical improvement in fit may have little

impact on our scientific interpretation of this aspect of the results.

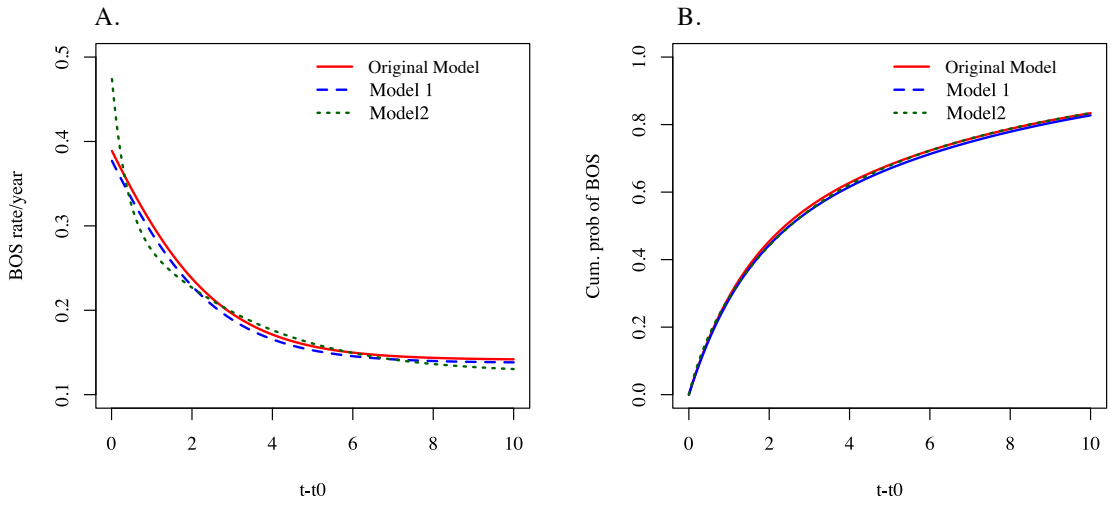

Figure 4.10: Point estimates for the BOS onset hazard rate (A) and CDF (B) from the original model and from models without BOS→ healthy transitions. Transitions in Model 1 and 2 are depicted in Figure 4.9.

The suggestion of statistical evidence in favor of a model with BOS to healthy transitions may seem counterintuitive given the biology of the disease. However, it highlights the fact that misclassification and disease model jointly explain the observed data, and, in this situation, it is plausible that evidence in favor of a reversible transition actually points to misspecification of the misclassification model. BOS is diagnosed with FEV1, a continuous measure with inherent variability. Our model assumes that misclassification probabilities are constant over the course of the disease. Misclassified disease outcomes are more likely to occur in individuals who have recently developed BOS, since their FEV1 may be near the diagnostic cutoff. The model that allows for BOS to healthy transitions may reflect non-constant misclassification probabilities with respect to disease duration.

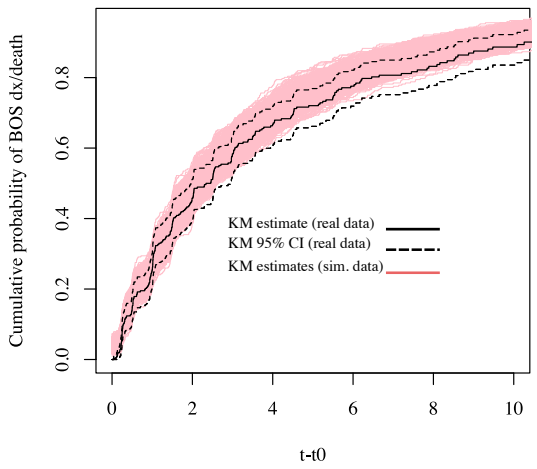

Figure 4.11: Kaplan-Meier estimates of time to first diagnosis of BOS or death from both the real data set and data simulated from the fitted model. Dashed lines are the 95% confidence intervals of the real data estimates.

4.8.7 Investigation of goodness of fit of BOS model

Titman and Sharples (2010) assessed goodness of model fit via a chi-square test comparing observed versus expected transition counts (Titman and Sharples, 2008) and did not find evidence for lack of fit. To supplement their investigation, we were interested in evaluating the plausibility of the model in describing time to BOS onset. To assess how well the distribution of the time to BOS development, induced by the latent CTMC model, fit the observed data, we compared model-based simulations of time of first observed BOS diagnosis or death to 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 to analogous K-M estimates in the artificial data (Figure 4.11). 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.

4.9 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. On the BOS dataset, the method considerably outperformed other optimization approaches, including those implemented in the R package msm (Jackson, 2011) – Nelder-Mead and BFGS – and the EM algorithm of Bureau et al (Bureau et al., 2003). We suspect that that results will be similar for other data sets 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 (Lystig and Hughes, 2002) at each iteration.

The utility of latent CTMC models lies in their ability to approximate functionals of disease processes from non-exponential sojourn time distributions. Our simulation studies investigated frequentist properties of such estimates using simple survival and 2-state disease models. We should note that while we approximated distributions with monotonic hazard functions, the latent CTMC models are not limited to these settings and can approximate hazard functions with other shapes. As discussed by Aalen (1995), Coxian distributions with three or more latent transient states can yield non-monotonic hazards, including functions with a bathtub shape. Ultimately, while somewhat limited in scope, our simulations fill a gap in the literature of frequentist properties of latent CTMC parameter MLEs under model misspecification, and should generalize to disease models consisting

of more than two states.

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, while 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. Pragmatically, investigators planning studies should consider setting the observation frequency only after examining fitted models using simulated data from disease models with a range of plausible hazard functions.

In general, the delta-method standard errors represented reasonable approximations of the true variability of hazard and CDF estimates, even though the latent CTMC models only approximated the data-generating distributions. Obtaining standard errors via a non-parametric bootstrap standard errors provides another option, but the computational intensiveness of the fitting method suggests they may not be practical for most datasets. Finally, for future methods development, it may be worth examining the feasibility of obtaining robust variance estimates, although an implementation in this setting may require substantial investment in additional computational machinery, given the partial observation context (Elashoff and Ryan, 2004).

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 (Steele and Raftery, 2010). 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 non-identifiable parameters under the null model (Chen et al., 2001). Currently, implementation in the latent CTMC context is limited to null models with exponential sojourn distributions (Titman and

Sharples, 2010), and extending this approach for more general testing is an area of future research. 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 (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.

4.10 Appendix: Complete data score and Hessian

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

4.10.1 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 β_{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

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

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

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

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.

4.10.2 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,
$\mathbf{O}_T(i) = \{O_T(i,1),\ldots,O_T(i,r)\} \sim Multinomial\{\mathbf{e}_i, N(i)\}$, 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
$\mathbf{o}_{\mathbf{i}[-\mathbf{1}]} = \{o_T(i,2),...,o_T(i,r)\}$. 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

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

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

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

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

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

With covariates, the Hessian matrix is given by

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

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

Chapter 5

LATENT CTMC MODELS FOR DISCRETELY-OBSERVED DISEASE PROCESSES WITH INFORMATIVE OBSERVATION TIMES

5.1 Introduction

In this chapter, we turn our attention to discretely-observed disease processes with observation times that are informative about an individual's underlying disease status. This scenario is particularly relevant given recent interest in mining large databases of electronic medical records (Dean et al., 2009). These data have complex features that present statistical and computational challenges. As with panel data from designed studies, observational data based on medical records provides information on patients' disease statuses only at clinic visits, and the disease may be observed with misclassification error. The difference here is that visit times, initiated by patients based on the symptoms, must also be captured by the modeling framework. Our focus in this chapter will be extending the latent CTMC framework for such data.

Most methods developed for panel observed multistate processes treat visit times as non-informative — an assumption that often does not hold in observational studies (Kalbfleisch and Lawless, 1985; Kay, 1986; Titman, 2011; Hubbard et al., 2008). Visits scheduled in advance, even those based on observations at previous time points, are ignorable; but times of patient-initiated, symptom-based visits cannot be ignored in the analysis because these times depend on the underlying disease process (Gruger et al., 1991). Non-ignorable visit times necessitate joint modeling of the disease process and visit times. However, existing joint models of this sort, capable of analyzing panel data (Chen et al., 2010; Chen and Zhou, 2011, 2013; Sweeting et al., 2010), assume pre-designated visits with informative missingness, which is appropriate for clinical trials but not for observational clinical data with random visit times.

In this chapter, we develop a joint model of a discretely observed multistate disease process and

a random observation time process. We treat the random, patient-initiated visit times as a temporal point process, which consists of a time series of binary events that occur in continuous time (Daley and Vere-Jones, 2003). Due to their tractability and flexibility, inhomogeneous Poisson processes are commonly used to model observation time point processes jointly with a longitudinal outcome, including continuous (Sun et al., 2005) and panel-count variables (Li et al., 2013). However, in these models the dependence of observation times and the disease process is specified by modeling the disease process conditional on the observation process. In contrast, we flip the conditioning, assuming that the observation process is a doubly stochastic Poisson process with rates that depend on the disease state. Our multistate-disease-driven observation (multistate-DDO) model can be viewed as an extension of the "preferential sampling" approach for spatial data to multistate disease processes (Diggle et al., 2010).

Our joint modeling framework is as follows. The disease process follows a latent CTMC trajectory. We condition on all scheduled visits and assume that patient-initiated DDO times accrue according to a Markov-modulated Poisson process (MMPP) with rates that depend on the patient's current disease status. The disease process is observed, with possible misclassification error, at informative and non-informative visit times. Our multistate-DDO model is similar to the earthquake timing model of Lu (2012), but our model also allows for observations at non-informative times. We demonstrate that the likelihood of our joint model is computationally tractable. Moreover, we develop an efficient expectation-maximization (EM) algorithm to fit our joint multistate-DDO model to panel data (Section 5.5). This EM algorithm shares a similar structure to the algorithm developed for optimizing the panel data likelihood in Chapter 4, and as such is able to utilize similar computational tools. Via simulations (Section 5.6), we demonstrate the importance of accounting for random informative sampling times in preventing bias and increasing precision of estimates of disease process parameters. We also investigate practical issues in fitting the multistate-DDO models, including the validity of our standard error estimates (Section 5.6.6) and performance of our model selection approach (Section 5.6.7).

5.2 Modeling framework

5.2.1 Joint model for disease process and disease driven observation process

The disease process, denoted $X(t)$ and modeled as a time homogeneous CTMC, has state space $S = \{1,\ldots,s\}$, infinitesimal generator matrix $\mathbf{\Lambda} = \{\lambda_{ij}\}$, and initial distribution $\boldsymbol{\pi}$. Jumps in $X(t)$ correspond to an individual's transitions between states in the disease process. The observation process, denoted $N(t)$, is a Markov-modulated Poisson process with piecewise constant rates $q(t) = q(X(t))$ that depend on the underlying disease state. $N(t)$ has state space $\{0,1,\ldots,\infty\}$, corresponding to the accrual of patient-initiated disease-driven observations (DDOs): the process jumps and the state increases by one each time a DDO occurs. Rates of DDOs corresponding to disease states $\{1,\ldots,s\}$ are denoted $\mathbf{q} = (q_1,\ldots,q_s)'$.

Jointly, the disease process and counts of DDOs evolve according to a bivariate time-homogeneous continuous time Markov chain, $Y(t) = (X(t),N(t))$ (Mark and Ephraim, 2013). The state space for $Y(t)$ is the Cartesian product of the state space of $X(t)$ and $N(t)$,

$$S' = \{(1,0),(2,0),\ldots,(s,0),(1,1),\ldots(s,1),\ldots,(1,\infty),\ldots,(s,\infty)\}.$$

Figure 5.1A shows an example of a joint three-state disease and observation process trajectory. Supposing $\mathbf{Q} = diag(q_1,\ldots,q_s)'$, the transition generator matrix for the joint process $Y(t)$ is

$$\mathbf{R} = \begin{bmatrix} \mathbf{\Lambda}-\mathbf{Q} & \mathbf{Q} & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{\Lambda}-\mathbf{Q} & \mathbf{Q} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{\Lambda}-\mathbf{Q} & \mathbf{Q} & \ldots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

The structure of $\mathbf{R}$ follows from the assumption that DDOs and changes in disease states cannot occur simultaneously. The first $\mathbf{\Lambda}-\mathbf{Q}$ block yields the transition rates between states $(i,0)$ and $(j,0)$ and the first $\mathbf{Q}$ block yields the rates between state $(i,0)$ and $(j,1)$; the rest of the generator matrix is structured similarly (Fearnhead and Sherlock, 2006).

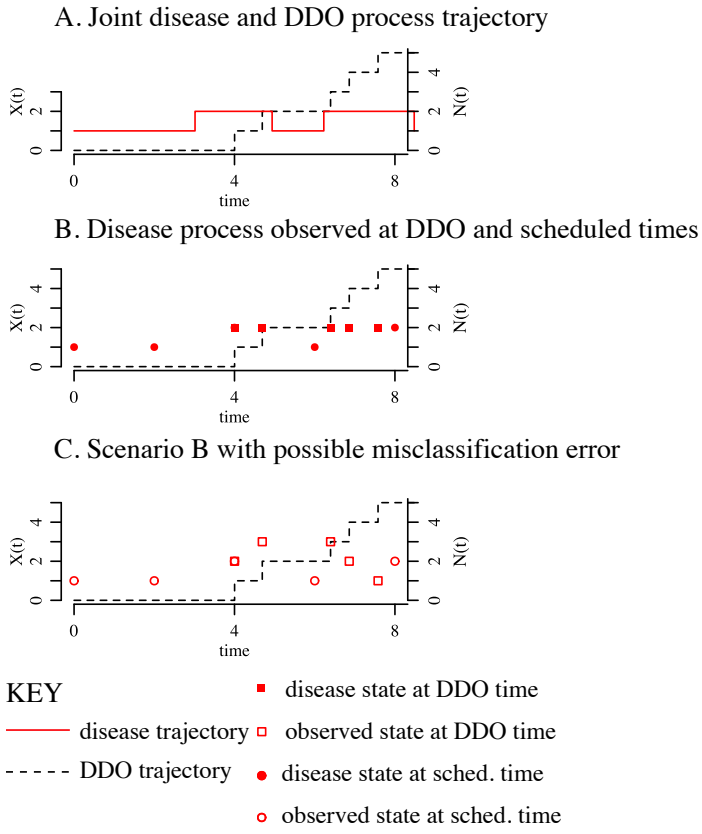

Figure 5.1: A. Example of a joint informative observation and disease process, $Y(t) = (X(t), N(t))$. B. The informative observation time process and the disease process observed at DDO and scheduled times. C. Same as B, with misclassification error.

5.2.2 Likelihood for observed data

Our observed data consist of partial observations of the joint disease and DDO process, since we only see an individual's disease status at DDO times or scheduled visit times. The observation times are $t_1, \ldots, t_n$, and DDO times are distinguished from scheduled visit times via indicator functions $\mathbf{h} = (h_1, \ldots, h_n)$. We denote the collection of DDO event times as $\boldsymbol{\tau} = \{t_i : h_i = 1, i = 1, \ldots, n\}$. Disease states at the observation times are $x_1, \ldots x_n$.

We first consider the likelihood where we observe $X(t)$ at DDO and scheduled visit times without misclassification error (Figure 5.1B). The likelihood conditions on scheduled visit times. The

random variable h_k is a censoring indicator that denotes whether a DDO observation occurred before or after the next scheduled visit time from time t_{k-1}. The Markov property and time-homogeneity of $Y(t)$ enables us to obtain the likelihood of the observed data as a product of density or survival functions for the first passage time of $Y(t)$ into state $(j,k+1)$, given $Y(t_k) = (i,k)$ across each observation interval $[t_{k-1},t_k]$. Given the time-homogeneity of $Y(t)$ and the structure of $\mathbf{R}$, it suffices to consider $U_{i0,j1}$, the first passage time into state $(j,1)$, given state $(i,0)$ at time 0. When t_k is a DDO time, the contribution to the likelihood for the interval $[t_{k-1},t_k]$ is the density of $U_{i0,j1}$, $f_{ij}(\Delta t_k)$, where $\Delta t_k = t_{k+1} - t_k$. When t_k is a scheduled visit time, we know that $U_{i0,j1} > \Delta t_k$, and the contribution to the likelihood is the survival function for $U_{i0,j1}$, $S_{ij}(\Delta t_k)$. Thus, the likelihood based on the observed data is

$$\mathrm{P}(x_1,\ldots,x_n,\boldsymbol{\tau},\mathbf{h}) = \nu_{h_1}\pi_{x_1}(h_1)\prod_{k=2}^{n}[f_{x_{k-1}x_k}(\Delta t_k)]^{h_{t_k}}[S_{x_{k-1}x_k}(\Delta t_k)]^{1-h_{t_k}}.$$

More generally, the disease process is observed with misclassification error at scheduled visits and DDO times (Figure 5.1C). Thus, we observe $\mathbf{o} = (o_1,\ldots,o_n)$ rather than $x_1,\ldots,x_n$. We assume that disease process observations are conditionally independent given $X(t)$. The relationship between observed and latent states is described by an emission matrix $\mathbf{E} = \{e(i,j)\}$ with entries $e(i,j) = \mathrm{P}[o_t = j|X(t) = i]$. The likelihood includes emission probabilities and sums $\mathrm{P}(x_1,\ldots,x_n,\mathbf{o},\boldsymbol{\tau},\mathbf{h})$ over the possible values of $\mathbf{x}$:

$$P(\mathbf{o},\boldsymbol{\tau},\mathbf{h}) = \sum_{x_1}\sum_{x_2}\ldots\sum_{x_n}\nu_{h_1}\pi_{x_1}(h_1)\prod_{k=2}^{n}[f_{x_{k-1}x_k}(\Delta t_k)]^{h_k}[S_{x_{k-1}x_k}(\Delta t_k)]^{1-h_k}\prod_{i=1}^{n}e(x_i,o_i). \tag{5.1}$$

One can derive the density and survival functions $f_{ij}(t)$ and $S_{ij}(t)$ explicitly in terms of $\boldsymbol{\Lambda}$ and $\mathbf{Q}$ using standard CTMC techniques (Freed and Shepp, 1982). First passage time $U_{i0,j1}$ has the same distribution of the absorption time of an auxiliary process $Y'(t)$, corresponding to $Y(t)$ for $\{t : N(t) \in \{0,1\}\}$, with state space $\{(1,0),\ldots(s,0),(1,1),\ldots(s,1)\}$, absorbing states $(1,1)\ldots(s,1)$, and rate matrix

$$\bar{\mathbf{R}} = \begin{bmatrix} \boldsymbol{\Lambda}-\mathbf{Q} & \mathbf{Q} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

The survival function for $U_{i0,j1}$ is

$$S_{ij}(t) = P[U_{i0,j1} > t | Y(0) = (i,0)] = P\left[Y'(t) = (j,0) | Y'(0) = (i,0)\right] = \exp[(\mathbf{\Lambda} - \mathbf{Q})t]_{ij},$$

and the density function is

$$f_{ij}(t) = \frac{d}{dt} P[U_{i0,j1} < t | Y(0) = (i,0)] = \frac{d}{dt} P\left[Y'(t) = (j,1) | Y'(0) = (i,0)\right] = \exp\left[(\mathbf{\Lambda} - \mathbf{Q})\, t\right]_{ij} q_j,$$

via the Kolmogorov forward equation.

Known times of death must be accounted for in the observed data likelihood (equation (5.1)). Let A be the set of all absorbing states in disease state space S. Assuming that absorption in other states and informative observation events are competing risks, the density of the time of absorption in state $k \in A$, designated by the random variable $U_{i0,k0}$, is given by

$$g_{ik}(t) = \frac{d}{dt} P[U_{i0,k0} < t | Y(0) = (i,0)] = \frac{d}{dt} P[Y'(t) = (k,0) | Y'(0) = (i,0)] = \sum_{j \notin A} S_{ij}(t) \lambda_{jk},$$

where i is a transient state.

When the final time t_n corresponds to absorption of $X(t)$ in state k, we modify the observed data likelihood (equation (5.1)) by replacing the terms $f_{x_{n-1}x_n}(\Delta t_n)$ or

$$[f_{x_{n-1}x_n}(\Delta t_n)]^{h_{t_n}} [S_{x_{n-1}x_n}(\Delta t_n)]^{1-h_n}$$

with $g_{x_{n-1}x_n}(\Delta t_n)$.

5.3 Forward and backward functions

Efficient calculation of the observed data likelihood (5.1) is facilitated through recursive methods developed for hidden Markov models and MMPPs (Baum et al., 1970). Forward and backward recursive functions will also be used in our EM algorithm to obtain MLEs for model parameters. We use the abbreviation $\mathbf{x}_{1:k}$ for $x_1, \ldots, x_k$, $\mathbf{o}_{1:k}$ for $o_1, \ldots o_k$, $\mathbf{h}_{1:k}$ for $h_1, \ldots, h_k$. The sequence of DDO times up to observation time t_k is denoted $\boldsymbol{\tau}(1,k) = \{t_i : h_i = 1, i = 1, \ldots, k\}$. Forward functions are

defined as

$$\alpha_{t_k}(u) = P[\mathbf{o}_{1:k}, \boldsymbol{\tau}(1,k), \mathbf{h}_{1:k}, X_k = u]$$

and backward functions as

$$\beta_{t_k}(u) = \mathrm{P}[\mathbf{o}_{k+1:n}, \boldsymbol{\tau}(k+1,n), \mathbf{h}_{k+1:n} | X_k = u].$$

The forward function is initialized with

$$\alpha_{t_1}(u) = \mathrm{P}(O_1 = o_1, X_1 = u, H_1 = h_1) = e(u, o_1)\nu_{h_1}\pi_{x_1}(h_1),$$

and the recursion for $k = 2, \ldots, n-1$ is

$$\alpha_{t_k}(u) = \sum_i \alpha_{t_{k-1}}(i) e(u, o_k) [f_{iu}(\Delta t_k)]^{h_k} [S_{iu}(\Delta t_k)]^{1-h_k}.$$

The backward function is initialized with $\beta_{t_n}(u) = 1$, and the recursion for $k = 1, \ldots, n-1$ is

$$\beta_{t_k}(u) = \sum_i \beta_{t_{k+1}}(i) e(i, o_{k+1}) [f_{ui}(\Delta t_{k+1})]^{h_{k+1}} [S_{ui}(\Delta t_{k+1})]^{1-h_{k+1}}.$$

5.3.1 Observed data likelihood

The observed data likelihood is

$$P(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}) = \sum_u \alpha_{t_n}(u),$$

via the forward algorithm; by the backward algorithm, it is

$$P(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}) = \sum_u \beta_{t_1}(u) e(u, o_1) \nu_{h_1} \pi_{x_1}(h_1).$$

The forward and backward recursions make the likelihood evaluation practical because, similarly to the standard HMM forward-backward algorithm, the algorithmic complexity of both recursions is $O(ns^2)$.

5.3.2 Latent CTMC model parameterization

Disease process models based on standard CTMCs assume that disease state sojourn times are exponentially distributed. To permit more flexibility, we assume a latent CTMC framework for the disease process. The specification is analogous to that in Chapter 4, section 4.2.1, but we repeat it to refresh the reader's memory. We denote the disease process $W(t)$, with state space $G = \{1, 2, \ldots, g\}$. Underlying $W(t)$ is a latent time-homogeneous CTMC $X(t)$, with transition intensity matrix $\mathbf{\Lambda}$ and initial distribution $\boldsymbol{\pi}$ and state space

$$S = \{1_1, 1_2, \ldots, 1_{s_1}\} \cup \{2_1, 2_2, \ldots, 2_{s_2}\} \cup \cdots \cup \{g_1, g_2, \ldots, g_{s_g}\}.$$

Each observable disease state corresponds to multiple states in the latent state space, such that $W(t) = j <=> X(t) \in \{j_1, j_2, \ldots, j_{s_j}\}$. The mapping of multiple latent states in S to a single disease state in G yields phase-type sojourn distributions of W(t), which can be used to approximate distributions with hazard functions having different shapes (Aalen, 1995). We assume a Coxian structure for $\mathbf{\Lambda}$ for its flexibility and the fact that, up to trivial permutation of states, it is uniquely parametrized when the latent space has a minimal dimension (Titman and Sharples, 2010; Cumani, 1982). Latent CTMC models can be specified in the framework of the observed data likelihood (5.1) through use of an emission matrix with observed state space G and hidden state space S that equates emission probabilities $e(j_1, k) = e(j_2, k), \ldots, e(j_{s_j}, k)$ for all $j, k \in G$, permitting the mapping of the latent disease space onto the observed disease space.

To incorporate baseline subject-level covariates $\mathbf{w}^k$ in the disease model, emission, and initial distribution models we use the approach described in Chapter 4, Section 4.2.3. One can also add covariates to DDO parameterizations, by relating

$$\log(q_i^k) = \boldsymbol{\zeta}_{ij}^T \mathbf{w}^k,$$

where k denotes the individual.

5.4 Model selection

We recommend selecting models via the Bayesian information criterion (BIC), given its good performance for selecting general mixture models (Steele and Raftery, 2010) and applicability to comparing non-nested models. The BIC can assist in choosing the dimension of latent space as well assessing parameter constraints in the DDO rates. Finally, hypothesis tests for covariate effects based on likelihood ratio or Wald tests are appropriate, provided parameter identifiability holds under the null model (Sundberg, 1973), which is achievable by constraining covariate effects rather than estimating them separately for each latent disease state (see Chapter 4, Section 4.2.3 for possible constraints).

5.5 Parameter Estimation

The parameters of interest in the multistate-DDO model, $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\Lambda}, \mathbf{E}, \mathbf{q})$, characterize the initial distribution, the disease process, the misclassification probabilities, and the DDO process rates, respectively; we will condition on h_1 rather than estimate its distribution. The standard approach for MMPPs and partially-observed bivariate CTMCs (Ryden, 1996a; Mark and Ephraim, 2013) is to use an EM algorithm to arrive at the maximum likelihood estimates (MLEs) of model parameters (Dempster et al., 1977), as it exploits the ease of maximizing a "complete data" likelihood compared to the observed data likelihood. The EM is iterative: at the $k+1$th step, one computes the expected complete data log-likelihood conditional on the observed data and the current estimate $\hat{\theta}^k$; then maximizes this expected likelihood with respect to $\boldsymbol{\theta}$ to yield $\hat{\theta}^{(k+1)}$. The expectation and maximization steps are repeated until successive iterates of $(\hat{\theta}^k, \hat{\theta}^{k+1})$ fulfill a specified convergence criterion.

In the multistate-DDO model, the complete data are $(\mathbf{x}, \boldsymbol{\tau}, \mathbf{o})$, the full disease trajectory, the DDO trajectory, and observed disease statuses at the discrete times, respectively. The complete data log-likelihood has exponential family form and is a linear function of complete data sufficient statistics. These sufficient statistics include $n_T(i,j)$, the total counts of transitions from state i to state j; $d_T(i)$, the total time spent in state i; z_i, the initial disease state indicator; $u_T(i) = \sum_{l=2}^{n} I(x_l = i) I(h_l = 1)$, the total number of DDOs that have occurred while X(t) was in state i; 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. As described by Lu (2012), the complete data log-likelihood for an individual is

$$\begin{aligned} l(\boldsymbol{\theta};\mathbf{o},\boldsymbol{\tau},\mathbf{x}|h_1) &= l(\boldsymbol{\pi};x_1|h_1) + l(\boldsymbol{\Lambda},\mathbf{q};\mathbf{x},\boldsymbol{\tau}|x_1) + l(\mathbf{E};\mathbf{o}|\mathbf{x},x_1) \\ &= \sum_{i}^{s} z_i \log[\pi_i(h_1)] + \sum_{i=1}^{s}\sum_{j\neq i} n_T(i,j)\log(\lambda_{ij}) - \sum_{i=1}^{s} d_T(i)\left(\sum_{j\neq i}^{s}\lambda_{ij}\right) \\ &+ \sum_{i=1}^{s} u_T(i)(q_i) - \sum_{i=1}^{s} q_i d_T(i) + \sum_{i=1}^{s}\sum_{j=1}^{r} o_T(i,j)\log[e(i,j)]. \end{aligned} \tag{5.2}$$

This likelihood is additive across multiple independent individuals, yielding the complete data likelihood for an entire sample.

5.5.1 E-step

The expectation step (E-step) of the EM algorithm requires computing the expectation of the complete data log-likelihood (5.2) conditional on observed data $(\mathbf{o},\boldsymbol{\tau},\mathbf{h})$.

To compute the E-step, we note that an individual's log-likelihood contribution (equation (5.2) is additive across time intervals $T_l = [t_l, t_{l+1}]$. Thus,

$$\begin{aligned} &\mathrm{E}[l(\boldsymbol{\theta};\mathbf{o},\boldsymbol{\tau},\mathbf{x})|\mathbf{o},\boldsymbol{\tau},\mathbf{h}] = \sum_{i=1}^{s}\mathrm{E}[z_i|\mathbf{o},\boldsymbol{\tau},\mathbf{h}]\log(\pi_i) \\ &+ \sum_{l=1}^{n-1}\sum_{i=1}^{s}\sum_{j\neq i}\mathrm{E}[n_{T_l}(i,j)|\mathbf{o},\boldsymbol{\tau},\mathbf{h}]\log(\lambda_{ij}) - \sum_{l=1}^{n-1}\sum_{i=1}^{s}\mathrm{E}[d_{T_l}(i)|\mathbf{o},\boldsymbol{\tau},\mathbf{h}]\left(\sum_{j\neq i}\lambda_{ij}\right) \\ &+ \sum_{l=2}^{n-1}\sum_{i=1}^{s} E[u_{T_l}(i)|\mathbf{o},\boldsymbol{\tau},\mathbf{h}]\log(q_i) - \sum_{l=1}^{n-1}\sum_{i=1}^{s} E[d_{T_l}(i)|\mathbf{o},\boldsymbol{\tau},\mathbf{h}]q_i \\ &+ \sum_{l=1}^{n-1}\sum_{i=1}^{s}\sum_{j=1}^{r}\mathrm{E}[o_{T_l}(i,j)|\mathbf{o},\boldsymbol{\tau},\mathbf{h}]\log\left[e(i,j)\right]. \end{aligned}$$

Computing the E-step therefore requires conditional expectations of the complete data sufficient statistics across T_l. Conditional expectations for z_i, $o_{T_l}(i,j)$, and $u_{T_l}(i)$ are computed using the Baum-Welch smoothing probabilities

$$\mathrm{P}(X_l = m|\mathbf{o},\boldsymbol{\tau},\boldsymbol{h}) = \frac{\beta_{t_l}(m)\alpha_{t_l}(m)}{\mathrm{P}(\mathbf{o},\boldsymbol{\tau},\mathbf{h})},$$

where $\alpha_{t_l}(m)$ and $\beta_{t_l}(m)$ are forward and backward functions. Hence,

$$\mathrm{E}[z_i|\mathbf{o},\boldsymbol{\tau},\mathbf{h}] = \mathrm{P}(X_1 = i|\mathbf{o},\boldsymbol{\tau},\mathbf{h}) = \frac{\beta_{t_1}(i)\alpha_{t_1}(i)}{\mathrm{P}(\mathbf{o},\boldsymbol{\tau},\mathbf{h})},$$

$$\mathrm{E}[o_T(j,m)|\mathbf{o},\boldsymbol{\tau},\mathbf{h}] = \sum_{l=1}^{n} I(o_l = m)\,\mathrm{P}(X_l = j|\mathbf{o},\boldsymbol{\tau},\mathbf{h}) = \sum_{l=1}^{n} I(o_l = m)\frac{\beta_{t_l}(j)\alpha_{t_l}(j)}{\mathrm{P}(\mathbf{o},\boldsymbol{\tau},\mathbf{h})},$$

and

$$\mathrm{E}[u_T(j)|\mathbf{o},\boldsymbol{\tau},\mathbf{h}] = \sum_{l=2}^{n} I(h_l = 1)\,\mathrm{P}(X_l = j|\mathbf{o},\boldsymbol{\tau},\mathbf{h}) = \sum_{l=2}^{n} I(h_l = 1)\frac{\beta_{t_l}(j)\alpha_{t_l}(j)}{\mathrm{P}(\mathbf{o},\boldsymbol{\tau},\mathbf{h})}.$$

Note that the sum in the last set of identities is over 2 to n, as the first time should not be considered an observed DDO event.

Expectations of CMTC sufficient statistics $C_{T_l} = d_{T_l}(i)$ or $C_{T_l} = n_{T_l}(i,j)$ can be obtained by first conditioning on x_l, x_{l+1}:

$$\mathrm{E}[C_{T_l}|\mathbf{o},\boldsymbol{\tau},\mathbf{h}] = \mathrm{E}\left[\mathrm{E}\left(C_{T_l}|\mathbf{o},\boldsymbol{\tau},\mathbf{h},X_l = a,X_{l+1} = b\right)\right] = \mathrm{E}\left[\mathrm{E}\left(C_{T_l}|X_l = a,X_{l+1} = b,H_{l+1} = h_{l+1}\right)|\mathbf{o},\boldsymbol{\tau},\mathbf{h}\right]. \tag{5.3}$$

This follows due to conditional independence of $X(t)$ on $[t_l,t_{l+1}]$ given knowledge of the joint disease and DDO process at the interval endpoints. The task of computing the expectation can be broken down into computing "inner" expectations $\mathrm{E}\left[C_{T_l}|X_l = a,X_{l+1} = b,H_{l+1} = h_{l+1}\right]$ and "outer" expectations. We describe the "inner" and "outer" expectations in turn.

5.5.2 Inner expectations for CTMC sufficient statistics

The formulae for the "inner expectations" are based on conditional expectations for CTMC sufficient statistics with absorbing states (Asmussen et al., 1996). We derive the desired quantities by considering conditional expectations of sufficient statistics $C = n_{ij}(t)$ or $C = d_t(i)$ for a generic homogeneous CTMC X(t) on the interval $[0,t]$, conditional on X(t) at interval endpoints and the informative observation status h_t at time t.

To obtain these expectations, recall that $U_{a0,b1}$ is the first passage time of the bivariate CTMC $Y(t) = (X(t),N(t))$ from state (a,0) to state (b,1). $U_{a0,b1}$ has the same distribution as the time to

absorption in state $(b,1)$ of the auxiliary process $Y'(t)$, given $Y'(0)=(a,0)$ and has survival function $S_{ab}(t)=\exp(\Lambda-\mathbf{Q})_{ab}$ and density function $f_{ab}(t)=\exp\left[(\Lambda-\mathbf{Q})t\right]_{ab}q_b$ (Section 5.2.2). We will use conditional expectation formulae applicable to $Y'(t)$ to derive the desired quantities.

When the endpoint t is a scheduled visit ($h_t=0$), we seek the conditional expectation

$$E[C|X(0)=a,X(t)=b,h_t=0]=\frac{E\left\{C\times I[Y'(t)=(b,0)]|Y'(0)=(a,0)\right\}}{S_{ab}(t)}. \tag{5.4}$$

Our bivariate representation of the process $Y'(t)$ enables us to use standard methods for computing expectations for CTMCs (Hobolth and Jensen, 2011). Thus, for $C=d_t(i)$, the numerator in equation (5.4) is the joint expectation

$$\begin{aligned}H_i[a,b]&=E\left\{d_t(i)\times I[Y'(t)=(b,0)]|Y'(0)=(a,0)\right\}\\&=\int_0^t\exp\left[(\Lambda-\mathbf{Q})(u)\right]_{ai}\exp([\Lambda-\mathbf{Q})(t-u)]_{ib}\,du,\end{aligned}$$

and for $C=n_t(i,j)$, the joint expectation

$$\begin{aligned}M_{ij}[a,b]&=E\left\{n_t(i,j)\times I[Y'(t)=(b,0)]|Y'(0)=(a,0)\right\}\\&=\int_0^t\lambda_{ij}\exp\left[(\Lambda-\mathbf{Q})u\right]_{ai}\exp\left[(\Lambda-\mathbf{Q})(t-u)\right]_{jb}du.\end{aligned}$$

When t corresponds to a DDO ($h_i=1$), we seek the conditional expectation

$$\begin{aligned}E[C|X(0)=a,X(t)=b,h_t=1]&=E[C|U_{a0,b1}=t,Y'(0)=(a,0)]\\&=\frac{\frac{\partial}{\partial t}\mathrm{E}[C,I(U_{a0,b1}<t)|Y'(0)=(a,0)]}{f_{ab}(t)}.\end{aligned} \tag{5.5}$$

To calculate the numerator, we employ expectation formulae derived for CTMCs with absorbing states (Asmussen et al., 1996). For $C=d_t(i)$, the numerator in (5.5) is given by the differentiated joint expectation

$$\frac{\partial}{\partial t}\mathrm{E}[d_t(i),I(U_{a0,b1}<t)|Y'(0)=(i,0)]=H_i[a,b]q_b,$$

and for $C = n_t(i,j)$, by

$$\frac{\partial}{\partial t}\mathrm{E}[n_t(i,j), I(U_{a0,b1} < t) | Y'(0) = (a,0)] = M_{ij}[a,b]q_b,$$

where $H_i[a,b]$ and $M_{ij}[a,b]$ are defined as before.

We also need to consider the special case of computing conditional expectations for $d_t(i)$ and $n_t(i,j)$ when the interval endpoint t corresponds to a known absorption time in the disease process, such as a time of death. Let A be the set of all absorbing states in S. Treating DDO events as a competing risk, suppose $U_{a0,k0}$ is the time of absorption of $Y'(t)$ in state $k \in A$, given $Y'(0) = (a,0)$, with density $g_{ak}(t) = \sum_{j \notin A} S_{ij}(t)\lambda_{jk}$. In this case, we need the conditional expectation

$$E[C | U_{a0,k0} = t, Y'(0) = (a,0)] = \frac{\frac{\partial}{\partial t}\mathrm{E}[C, I(U_{a0,k0} < t) | Y'(0) = (a,0)]}{g_{ak}(t)}. \tag{5.6}$$

When the complete-data statistic of interest is $C = d_t(i)$, the numerator in equation (5.6) is the differentiated joint expectation

$$\frac{\partial}{\partial y}\mathrm{E}[d_t(i) I(U_{a0,k1} < t) | Y'(0) = (a,0)] = I(i \notin A) \sum_{c \notin A} H_i(t)[a,c]\lambda_{ck}.$$

For $C = n_t(i,j)$, the numerator in equation (5.6) is the differentiated joint expectation

$$\frac{\partial}{\partial y}\mathrm{E}[n_t(i,j) I(U_{a0,k1} < t) | Y'(0) = (a,0)] = I(i,j \notin A) \sum_{c \notin A} M_{ij}(t)[a,c]\lambda_{ck} + I(i \notin A, j = k) S_{ai}(t)\lambda_{ik}.$$

One can use eigenvalue decomposition or the uniformization approach to computing the integrals in each of the joint expectation formulae (Hobolth and Jensen, 2011). Our implementation uses the efficient matrix-based methods from (Minin and Suchard, 2008a).

5.5.3 Outer expectations for CTMC sufficient statistics

After computing the “inner expectations," using the described formulae, one can compute “outer” expectations (5.3) for sufficient statistics $C_{T_l} = d_{T_l}(i)$ or $C_{T_l} = n_{T_l}(i,j)$ on the interval T_l using Baum-

Welch's bivariate smoothing probabilities

$$\mathrm{P}(X_l = a, X_{l+1} = b|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}) = \frac{e(b, o_{l+1})\alpha_{t_l}(a)\beta_{t_{l+1}}(b)[f_{ab}(\Delta t_{l+1})]^{h_{l+1}}[S_{ab}(\Delta t_l)]^{1-h_{l+1}}}{\mathrm{P}(\mathbf{o}, \boldsymbol{\tau}, \mathbf{h})}.$$

Thus, the expression for the conditional expectation of the complete data sufficient statistic C_T across the entire time interval $T = [t_1, t_n]$ is

$$\mathrm{E}[C_T|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}] = \sum_{l=1}^{n-1}\sum_{a=1}^{s}\sum_{b=1}^{s}\mathrm{E}[C_{T_l}|X_l = a, X_{l+1} = b, H_{l+1} = h_{l+1}]\,\mathrm{P}(X_l = a, X_{l+1} = b|\mathbf{o}, \boldsymbol{\tau}, \mathbf{h}).$$

5.5.4 Maximization step

The maximization step (M-step) maximizes the conditional expectation of the complete data likelihood, calculated in the E-step, with respect to $\boldsymbol{\theta}$. Covariate-free models admit closed-form M-steps (Lu, 2012). For covariate-parameterized models, we optimize the complete data likelihood via the Newton-Raphson method, as described in Chapter 4, Section 4.3.1; the extension to multistate-DDOs is straightforward, as complete-data score and information functions for the $\mathbf{q}$ parameters are identical to those for $\boldsymbol{\Lambda}$.

We provide an implementation of the EM algorithm in R (R Core Team, 2013), in the form of the R package `cthmm`, available at `http://r-forge.r-project.org/projects/multistate/`. As with all local optimization methods, convergence to the true maximum log-likelihood is not guaranteed, and the method is sensitive to starting values. To make it likely that the true maximum is obtained, we run the EM algorithm from multiple sets of initial values, such as random deviates around sensible values based on prior knowledge or MLEs obtained from fitting simpler, e.g., covariate-free, models. Finally, we use numerical differentiation, implemented in the R package "NumDeriv" (Gilbert and Varadhan, 2012), to obtain standard errors for parameter estimates from the observed Fisher information matrix.

5.6 Simulation Study

5.6.1 Objectives

We used simulated data to investigate several aspects of multistate-DDO models. The first objective was to study bias incurred by ignoring informative visits and treating such data as panel data (Section 5.6.4). We were specifically interested in the direction of bias in disease rate estimates and how varying sampling time rates or including scheduled visit times along with informative times affects such bias. The second objective was to assess how the joint model approach affects precision of disease process parameters, particularly when the disease process is observed with misclassification error (Section 5.6.5). Finally, we sought to assess the reasonableness of standard error estimates based on the Hessian function for the log-likelihood (Section 5.6.6) and to validate the model selection method via the BIC (Section 5.6.7).

5.6.2 Data

We considered three disease models: 1) a standard CTMC reversible disease model with two states (*healthy* and *diseased*); 2) a latent CTMC reversible disease model; and 3) a latent CTMC competing risks model (Figure 5.2). The competing risks model will be relevant to our data application discussed in Chapter 6, a study of secondary breast cancer events (SBCEs) in women with a history of primary breast cancer, which has a goal of estimating the cumulative incidence of mammographically detectable disease. In our simulated data models, events I and C correspond to ipsilateral (same side) and contralateral (opposite side) events, and the data correspond to a woman's sequence of mammograms following her primary cancer.

After simulating disease trajectories from each of these models, we used the MMPP DDO models to generate discretely-observed datasets with informative observation times, specifying comparatively higher DDO rates in the diseased states than in the healthy states. The competing risks model allowed for potentially misclassified observations, corresponding to disease surveillance tests with 70% sensitivity and 98% specificity. See Tables 5.1 and 5.2 for details. Each experiment involved 100 simulated datasets with 1000 independent individuals, unless otherwise noted.

5.6.3 Estimation and parameter identifiability

Our simulation studies involve fitting multistate-DDO models or non-informative panel data models to the generated data. Each model was fit using our EM algorithm at 3 randomly selected sets of starting values generated from $Normal(\mu = 0, sd = .20)$ distribution. In general, the models fit to data simulated from the reversible disease models without misclassification error and to the competing risks disease model appeared to have a unique set of parameters corresponding to the maximum attained likelihood. Additional spot checks on a subset of the simulated data with 10 starting values each suggested these models have fully identifiable parameters. In contrast, the models fit to reversible data models (Figure 5.2 A and B) in the presence of misclassification error corresponded to a likelihood with 2 modes. This appeared to be a case of label-switching observed more generally in HMMs (Zucchini and MacDonald, 2009). Label switching usually is not a problem for maximum likelihood estimation as one can impose parameter constraints after locating one of the likelihood modes. For example, in our analysis of reversible models with misclassification errors, we used the mode where the estimated probability of state misclassification was less than 0.5.

Table 5.1: Data descriptions for discretely-observed datasets simulated from reversible disease models (Figures 5.2A and 5.2B), including DDO rates, fixed observation times, and misclassification probabilities. These data specifications pertain to experiments summarized in Figure 5.3 and in Figure 5.5. Each experiment consisted of 100 simulated datasets with 1000 independent individuals.

Figure	Disease model	q_D	q_H	e(H,D)	e(D,H)	Obs. interval	Fixed times	DDOs
5.3A	A	2	0.25	0	0	[0,8]	0,2,4,6,8	Y
5.3B	A	2	0.25	0	0	[0,8]	0,8	Y
5.3C	A	0.3	0.25	0	0	[0,8]	0,8	Y
5.3D	B	2	0.25	0	0	[0,8]	0,8	Y
5.5A	A	2	0.25	0.15	0.15	[0,7.9]	0,7.9	Y
5.5B	A	0	0	0.15	0.15	[0,7.9]	0,7.9+10 obs.	N
5.5C	B	2	0.25	0.15	0.15	[0,.8.2]	0,8.2	Y
5.5D	B	0	0	0.15	0.15	[0,8.2]	0,8.2+8 obs	N

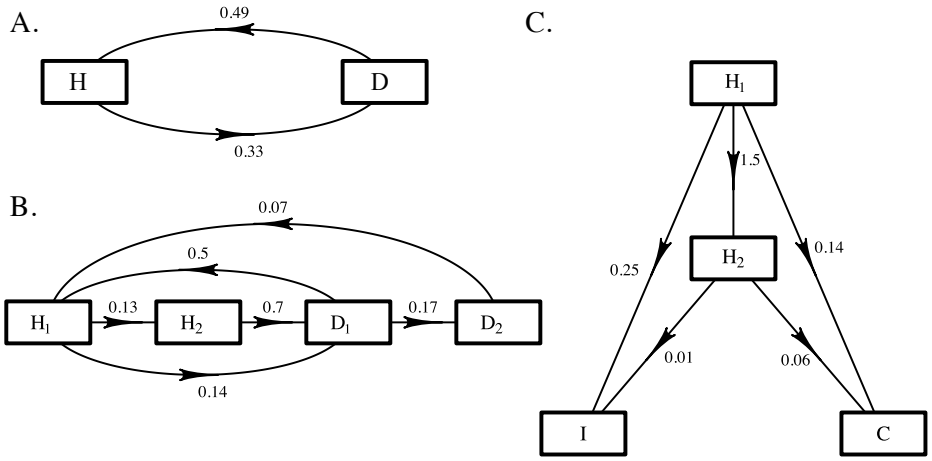

Figure 5.2: Data-generating disease models for simulation study. A. 2-state standard CTMC disease model. B. 2-state latent CTMC disease model, where latent states (H_1, H_2) and (D_1, D_2) map to *diseased* and *healthy* states, respectively. C. Competing risks disease model similar to the SBCE model. Latent states (H_1, H_2) map to the *healthy* state; I and C are two absorbing diseased states, corresponding to ipsilateral and contralateral SBCEs.

5.6.4 Bias resulting from ignoring informative visit times

To investigate bias resulting from ignoring DDO times, we fit data generated from the reversible models with correctly specified multistate-DDO models and with misspecified panel data models

Table 5.2: Data descriptions for simulated data from discretely-observed competing risks model (Figure 5.2C), including DDO rates, fixed observations, and misclassification probabilities. Notation: $q_{I/C} = q_I = q_C$ and $e(H,I/C) = e(H,I) = e(H,C)$. These data specifications pertain to experiments summarized in Figure 5.4. Each experiment consisted of 100 simulated datasets with 1000 independent individuals.

Figure	Disease mod.	$q_{I/C}$	q_H	e(H,I/C)	e(I/C,H)	Obs. interval	Fixed times	%DDOs
5.4	C	2	0.25	0.01	0.3	[0,8]	0,8	49%
5.4	C	2	0.25	0.01	0.3	[0,8]	0,2,4,6,8	35%
5.4	C	2	0.25	0.01	0.3	[0,8]	0,1,2,...,7,8	20%
5.4	C	2	0.25	0.01	0.3	[0,8]	0,.5,1,...,7.5,8	11%
5.4	C	2	0.25	0.01	0.3	[0,8]	0,.25,.5,...,7.75,8	6%

that condition on the observation times. The multistate-DDO models yielded unbiased estimates of the disease hazards. Under the misspecified panel models, bias in rate estimates from the reversible standard CTMC followed a consistent pattern: hazard rates for *healthy* $\rightarrow$ *diseased* transitions and *diseased* $\rightarrow$ *healthy* transitions were over- and under-estimated, respectively (Figure 5.3). Intuitively, informative observation times lead to more observations in the *diseased* state and fewer in the *healthy* state than would be expected under scheduled visits. Bias declined when non-informative times were included with the informative observations (Figure 5.3A vs 5.3C) and when DDO rates were less discrepant between *healthy* and *diseased* states (Figure 5.3B vs 5.3C). Ignoring informative times in the latent CTMC reversible models also led to underestimates of *diseased* $\rightarrow$ *healthy* hazard rates, but *healthy* $\rightarrow$ *diseased* hazard rates were overestimated only near the state origin time.

In the competing risks disease model similar to the SBCE application, we focused on estimates of the cumulative incidence functions of disease of events I and C. Again, to investigate bias, we either fit correctly-specified multistate-DDO models or misspecified panel data models. The correctly-specified multistate-DDO model produced unbiased cumulative incidence estimates. The bias resulting from ignoring informative visit times was consistent with results from reversible models: the hazard rates for $healthy \rightarrow I/C$ events were overestimated, yielding left-shifted cumulative incidence curves (Figure 5.4). Moreover, bias decreased with increasing numbers of scheduled visits added to supplement informative visits. Misspecification of the informative sampling times also dramatically underestimated mammography sensitivity estimates, e.g., sensitivity was estimated at 40% when 20% of visits were informative, versus the data-generating sensitivity of 70%. Finally, in addition to investigating bias given model misspecification, we also observed that cumulative incidence estimates based on the properly specified DDO model were shifted left relative to those based on a simulated time of diagnosis, i.e., the time of the first true-positive mammogram (Figure 5.4). This is consistent with diagnosis being a left censoring event for screen-detectable disease.

5.6.5 Gains in precision from using multistate-DDO model

Via simulation, we also examined the precision of estimates of disease process parameters under informative and non-informative observation schemes. Informative visit times mitigate the uncertainty

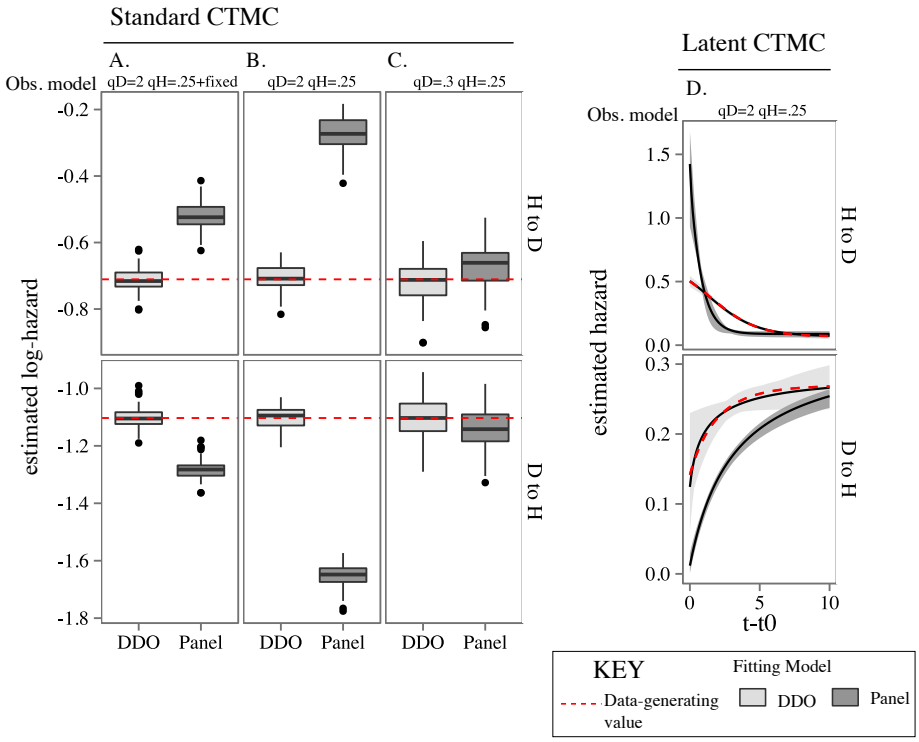

Figure 5.3: Box plots/functional box plots for hazard estimates of $H \to D$ and $D \to H$ transitions using data simulated from discretely observed 2-state standard and latent CTMC multistate-DDO models on the interval t=[0,8]. (See Table 5.1 for simulation details). Data were fit with correctly specified multistate-DDO models and incorrectly specified panel models, demonstrating bias resulting from ignoring informative visits. A. DDO rates are $q_D = 2, q_H = 0.25$; data also included fixed observation times $t = (0,2,4,6,8)$. B. DDO rates are $q_D = 2, q_H = 0.25$. C. DDO rates are $q_D = 0.35, q_H = 0.25$. D. DDO rates are $q_{H1} = q_{H2} = 0.25$ and $q_{D1} = q_{D2} = 2$.

about the underlying disease states at discrete observations with misclassification error, enabling more precise estimates. We generated data from the reversible standard and latent CTMC disease models (Figure 5.2) and simulated misclassified observations either in data sampled at DDO times or at pre-designated visit times with equivalent average observation frequencies (Table 5.1). The simulated data were fit with correctly specified multistate-DDO models or panel models, and we observed less variability in multistate-DDO estimates than in their panel model equivalents (Figure 5.5).

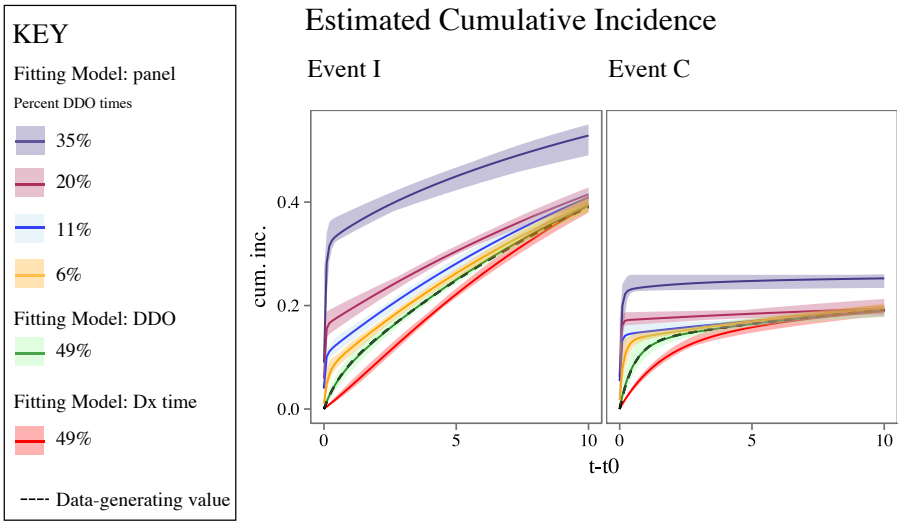

Figure 5.4: Functional box plots for simulated data estimates of cumulative incidence for disease events I and C in the latent CTMC competing risks model (Figure 5.2C.) Discretely observed data were generated from the disease trajectories according to informative observation times from a DDO model with $q_{H1} = q_{H2} = .25$ and $q_I = q_C = 2$, and varying proportions of supplemental non-informative times. Observations had 70% sensitivity and 98% specificity, corresponding to mammography data. See Table 5.2 for further dataset details. Data were fit with panel models or multistate-DDO models, demonstrating bias incurred by ignoring informative observations, and showing how increasing proportions of supplemental scheduled visits mitigates such bias. Also shown is cumulative incidence based on time of diagnosis (Dx time), the time of the first true positive mammogram.

5.6.6 Validity of standard error estimates

We also used our simulated data to investigate the validity of standard error estimates based on the observed information from the Hessian of the log-likelihood function. We fit correctly specified models to data simulated from the multistate-DDO models described in Table 5.3, and obtained the Hessian at the MLEs. The data-generating models included both standard and latent CTMC reversible models, with and without misclassification error, and the competing risks model. To evaluate the reasonableness of the variance estimates based on the Hessian function, we compared the distribution of the estimated variance (5th, 50th, and 95th quantiles) for model parameters to the empirical variance of the MLEs. The results of these experiments are shown in Figure 5.6. In

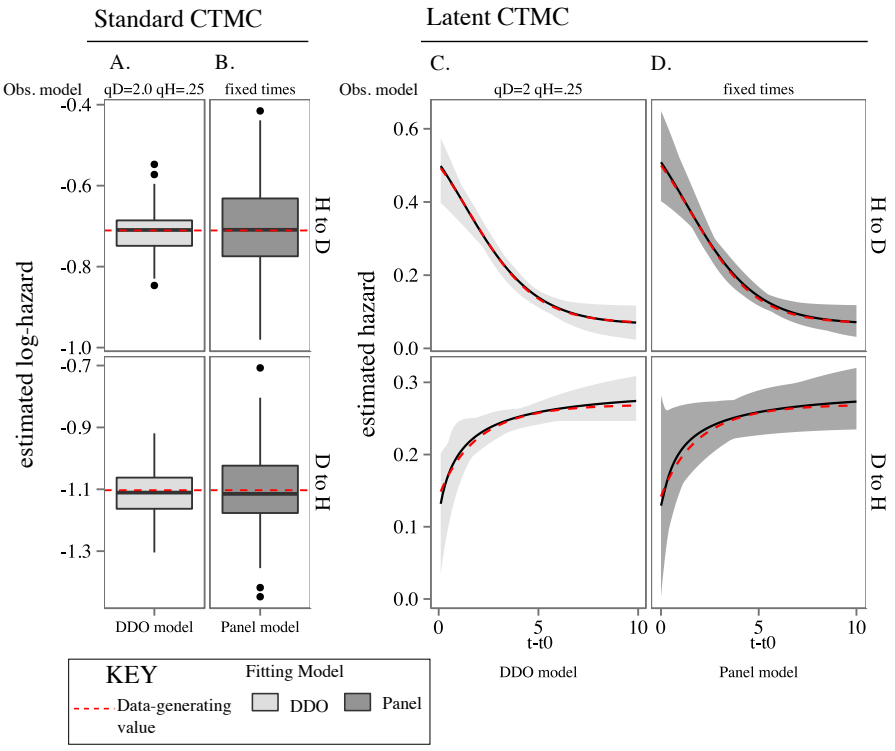

Figure 5.5: Box plots/functional box plots for hazard estimates of $H \to D$ and $D \to H$ transitions for standard and latent CTMC reversible disease models (Figure 5.2A, 5.2B), observed with 15% misclassification error at either DDO times or at fixed times with equal average frequencies. See Table 5.1 for further details. Data are fit with correctly specified multistate-DDO or panel models. These results demonstrate the gains in precision in hazard estimates via jointly modeling informative sampling times in the presence of misclassification error.

general, the agreement between empirical variance of MLEs and the Hessian-based estimates was good, and there was little indication of a pattern when estimates were discrepant, suggesting that these results may be attributable to sampling variability and the relatively small number of simulated datasets (100) for each experiment.

5.6.7 Evaluation of BIC criterion for model selection

We were also interested in evaluating the validity of the BIC criterion for model selection. To that end, we fit additional models to 50 datasets generated from the competing risks disease model

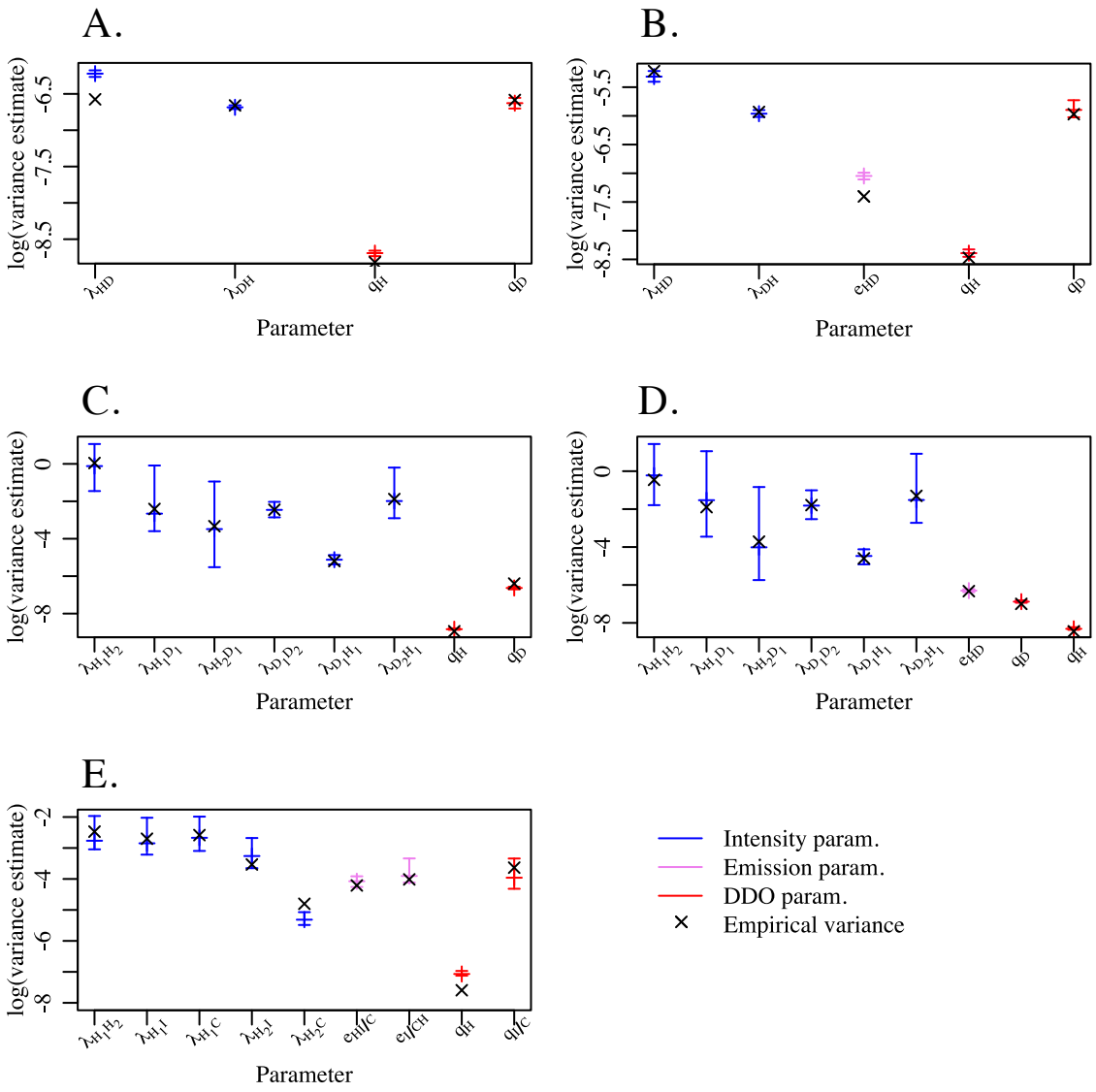

Figure 5.6: Estimated variance of model parameter estimates, based on the observed information matrix (5th, 50th, and 95th quantiles), and on the empirical variance for parameter estimates. Data generating models are fully described in Table 5.3. Data were fit with the appropriately specified multistate-DDO model. A. Reversible disease model with standard CTMC. B. Reversible disease model with latent CTMC. C. Same as A, with misclassification error. D. Same as B, with misclassification error. E. Competing risks disease model.

.

Table 5.3: Descriptions of data sets used to investigate validity of standard error estimates based on the Hessian of the log-likelihood functions. Each experiment consisted of 100 simulated datasets with 1000 independent individuals.

Figure	Disease mod.	q_D	q_H	e(H,D)	e(D,H)	Obs. interval	Fixed times	DDOs
5.6A	A	2	0.25	0	0	[0,8]	0,2,4,6,8	Y
5.6B	A	2	0.25	0.15	0.15	[0,7.9]	0,7.9	Y
5.6C	B	2	0.25	0	0	[0,8]	0,8	Y
5.6D	B	2	0.25	0.15	0.15	[0,.8.2]	0,8.2	Y

Figure	Disease mod.	$q_{I/C}$	q_H	e(H,I/C)	e(I/C,H)	Obs. interval	Fixed times	DDOs
5.6E	C	2	0.25	0.01	0.3	[0,8]	0,8	Y

Table 5.4: Multistate DDO models fit to simulated competing risks data (Table 5.2, line 1, provides the details for data-generating model). Model 3 in this table is the correctly specified multistate-DDO model.

Model label	Disease model	DDO constraints	No. params
1	Standard CTMC	$q_I = q_C$	6
2	Latent CTMC (2 state)	$q_{H_1} = q_{H_2} = q_I = q_C$	8
3	Latent CTMC (2 state)	$q_{H_1} = q_{H_2}, q_I = q_C$	9
4	Latent CTMC (2 state)	$q_{H_1} = q_{H_2}$	10
5	Latent CTMC (3 state)	$q_{H_1} = q_{H_2} = q_{H_3}, q_I = q_C$	13

(Figure 5.2C.) observed at DDO times on the interval [0,8] (DDO rates $q_{I/C} = 2$, $q_H = 0.25$), and compared the BIC from the incorrectly specified models to those from the correctly specified multistate-DDO model. The alternative, incorrect models, varied either the disease model structure, by specifying either a standard CTMC or latent CTMC with Coxian structure and 3 latent healthy states, or the DDO model, by specifying equal DDO rates across all states or allowing for different DDO rates in I and C disease states. Table 6.1 provides details of the additional models fit to the data. After fitting each model to the simulated data, we calculated and ranked the BIC for each of the models fit to the data. Across each of the 50 datasets, the ranking of the BIC was consistent: BIC was lowest for the correctly specified model (Model 3, Table 6.1), followed by Models 4, 5, 1, and 2. Thus, using the criteria of selecting a model based on the lowest BIC, the correctly specified

model was selected for 50/50 simulated datasets.

5.7 Discussion

The increasing availability of electronic medical resources presents new opportunities for modeling multistate diseases. However, as patients' disease statuses are only assessed at discrete clinic visit times – and visit times may be informative about the patients' disease histories – these data pose challenges for inference. The multistate-DDO model provides a novel and flexible approach for modeling such data: it applies to a broad class of disease models, including chronic diseases with reversible transitions and duration-dependent hazard functions; allows for covariate effects; and accommodates both patient-initiated random visit times and scheduled non-informative visits.

The model's contribution to methodology for discretely-observed disease process data is to accommodate informative patient-initiated visit times by jointly modeling the random informative observation and disease processes. Via simulations, we showed the need for such an approach to avoid bias. Ignoring the informative sampling led to overestimated rates of transitions into and underestimated rates out of preferentially sampled disease states, as well as biased estimates of misclassification probabilities. We also showed that multistate-DDO models can improve precision of estimates of disease process parameters with misclassified data, as informative visit times provide additional information about an individual's unobserved disease status. Other simulations demonstrated the reasonableness of standard errors for model parameter estimates based on the Hessian of the log likelihood, and usefulness of the BIC for selecting the disease and DDO model structure.

Our investigations of the multistate-DDO model thus far demonstrate its utility in disease modeling using observational data with informative sampling times. However, these models have multiple components, and misspecification of any component has the potential to lead to biased estimates overall. A valid and open question remains about whether the assumptions about the observation time process implied by the MMPP model adequately capture patient behavior, and if not, how much this affects inference about the disease process. The MMPP model assumes that DDO events in non-overlapping intervals are independent conditional on the latent disease process history. In reality, patient-initiated visit times likely display dependence on upcoming scheduled and prior visit

times. One can evaluate the reasonability of the MMPP assumptions using goodness of fit tests based on time-rescaling methods that transform general point processes into a standard homogeneous Poisson process with unit intensity (Brown et al., 2002; Lu, 2012), which we will expand on in Chapter 7. We also note the possibility of expanding our DDO model to accommodate prior and future visit times as time-dependent covariates, allowing for additional temporal dependence in the DDO process.

Pragmatically, with real data, it is worth fitting a variety of multistate-DDO models and assessing the sensitivity of different components to model assumptions–and verifying the reasonableness of results with findings in the literature. We will apply this approach to our secondary breast cancer application in Chapter 6.

Chapter 6

APPLICATION OF MULTISTATE-DISEASE-DRIVEN-OBSERVATION MODEL TO A STUDY OF SECONDARY BREAST CANCER EVENTS

6.1 Introduction

We illustrate the multistate-DDO model by applying it to an observational study of secondary breast cancer events (SBCEs) in women who have had a unilateral primary breast cancer (BC), using a dataset based based on electronic medical records. In contrast to the conventional studies of SBCEs based on diagnosed events (Chapman et al., 1999; Geiger et al., 2007; Buist et al., 2010), our approach will be to treat diagnosis time as a left censoring event for time of mammographically detectable disease. The target of inference is onset of mammographically-detectable ipsilateral (same side) or contralateral (opposite side) SBCEs. Because ipsilateral events include recurrences of the primary BC as well as new cancers, whereas contralateral events include only the latter, it is of interest to model these events separately (Demicheli et al., 1996). The dataset consists of the sequence of mammograms and biopsies following completion of treatment for a primary breast cancer. These data are suited for multistate-DDO models, as mammograms have misclassification error, and observation times include both scheduled screening and patient-initiated visits. Scientifically, we are interested in differences in estimates of cumulative incidence of mammographically-detectable versus diagnosed SBCEs, estimates of mammography misclassification, and estimates of covariate effects on disease process parameters. Estimates from our model are clinically meaningful, as they provide information about prevalence of undetected SBCEs in the growing population of breast cancer survivors (Siegel et al., 2012) as well as screening accuracy in this population.

6.2 Methods

6.2.1 Study population

The study population consists of women diagnosed with unilateral primary BC between 1994 and 2009 who were members of Group Health (GH), an integrated health care system in Washington State, at the time of their primary cancer diagnosis. Women were followed from 180 days after their first cancer until the earliest of the first positive biopsy for an SBCE, death, or disenrollment from the GH cohort. Women in this population were recommended to undergo annual screening mammograms in an effort to detect SBCEs before they become symptomatic. Women were also recommended to receive diagnostic evaluations for symptoms that arise in between scheduled surveillance intervals.

6.2.2 Mammography and biopsy outcomes

Mammograms were positive if the BI-RADS (Breast Imaging-Reporting and Data System) score was 0=“more imaging needed,” 4=“suspicious abnormality,” 5=“highly suggestive of malignancy,” or 6=“known malignancy” (American College of Radiology, 2003). Mammograms that were positive were followed up with further imaging workup, and, if warranted, biopsies. Biopsies with a result of invasive malignancy or ductal carcinoma in situ (DCIS) were considered positive; negative findings included benign growths and benign hyperplasias.

6.2.3 Definition of informative observation times

Mammography visit times were considered to be scheduled screening visits unless the woman and radiologist reported that the visit was for "evaluation of a breast problem," or only the radiologist coded it as such, but the woman endorsed an additional variable indicating symptoms.

6.2.4 Covariates

The specific covariates we focused on included age at diagnosis, dichotomized to age<50 versus age>50; American Joint Committee on Cancer, Version 6, stage of the primary BC (0=in-situ, 1, 2+); adjuvant endocrine therapy for the original cancer (yes or no); and race (White versus non-White).

6.2.5 Dataset exclusions

There were 4,133 women with primary unilateral breast cancers diagnosed from 1994-2009 who subsequently received mammography at Group Health. We applied sequential exclusions to obtain an analysis dataset. We excluded women with a mammographically-detectable SBCE within 180 days following the primary BC diagnosis (N=94), since events prior to that time likely reflect progression of the primary disease. We also excluded women if they had a biopsy record not preceded by a mammogram within the preceding 100 days (N=352), since this suggests missing mammography records. Finally, we excluded those with any missing laterality for mammograms or biopsy procedures (N=424), and those missing any of the covariates of interest (N=327). In total, these exclusions reduced the dataset from 4,133 to 2,936 women, removing 49% percent of ipsilateral cases, 32% of contralateral cases, 37% of those who died prior to an SBCE, and 27% of those who were alive and SBCE-free at the time they were last seen. More ipsilateral cases were dropped since they were more likely to have biopsies not preceded by mammograms within the study period, due to the preponderance events proximal to the primary cancer that occurred on the ipsilateral side.

6.2.6 SBCE Models

The disease model is a competing risks model with three absorbing states: ipsilateral SBCE, contralateral SBCE, and death before SBCE. We considered both a standard CTMC with state space $\{H = \text{healthy}, I = \text{Ipsilateral SBCE}, C = \text{contralateral SBCE}, D = \text{death before SBCE}\}$ and a latent model with state space $\{H_1, H_2, I, C, D\}$, where H_1, and H_2 are two latent states that map to the healthy disease state. The latent model is biologically plausible as it allows for SBCE hazard rates

to be higher near the time of primary BC diagnosis, reflecting recurrences of the primary BC, and to level out over time, reflecting novel cancer events (Demicheli et al., 1996). The transitions in the two models are depicted in Figure 6.1. All women are assumed to be disease free at the beginning of the study, and start in either the H or H_1 state, depending on the disease model.

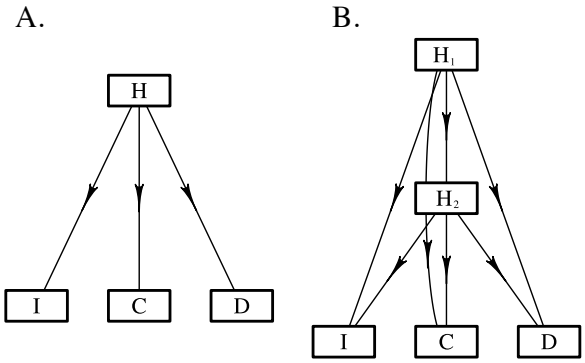

Figure 6.1: SBCE competing risks disease models. A. Standard CTMC, where H=healthy, C=contralateral SBCE, I=ipsilateral SBCE, and D=death before SBCE. B. Latent CTMC with Coxian structure. States H_1 and H_2 map to the healthy state.

Covariates were added to the disease model assuming an additive effect on the log-rates, i.e., $log(\lambda_{ij}^k) = \boldsymbol{\zeta}_{ij}^T \boldsymbol{w}^k$, where $\boldsymbol{w}^k$ are the covariates for individual k and $\boldsymbol{\zeta}_{ij}$ the coefficients for transition i, j. To ensure parameter identifiability, we constrained parameters in the latent model $\boldsymbol{\zeta}_{H_1,j} = \boldsymbol{\zeta}_{H_2,j}, j \in \{I,C,D\}$ and did not add covariates to the $H_1 \rightarrow H_2$ transition. Thus, for each covariate, there is one parameter each affecting transition rates from the healthy state to ipsilateral SBCEs, contralateral SBCEs and death prior to SBCE.

The DDO models specify rates of informative sampling times according to the individual's underlying disease state. For model comparison and sensitivity analysis we considered different restrictions on these DDO rates, i.e., assuming that the rate was the same in more than one state (for details, see Table 6.1). All models assumed that the DDO rate in the death state was zero. Models that assume DDO rates are identical across the healthy and ipsilateral and contralateral states suggest that the sampling times are not informative about the disease process: this assumption yields estimates that are quite similar to models that condition on the times, but allows for model comparison.

Table 6.1: Informative sampling time models for the SBCE data. Non-informative models assume the same DDO rate in all states.

Model label	Disease model	DDO model		
			No. DDO params	Constraints
1	Standard CTMC	non-informative	1	$q_H = q_I = q_C$
2		H/I,C	2	$q_H, q_I = q_C$
3		H/I/C	3	q_H, q_I, q_C
4	Latent CTMC	non-informative	1	$q_{H_1} = q_{H_2} = q_I, q_C$
5		H1,H2/I,C	2	$q_{H_1} = q_{H_2}, q_I = q_C$
6		H1/H2/I,C	3	$q_{H_1}, q_{H_2}, q_I = q_C$
7		H1/H2/I/C	4	$q_{H_1}, q_{H_2}, q_I, q_C$

Each mammogram and biopsy was classified as ipsilateral or contralateral. To model mammography misclassification, we assumed a zero probability of detecting an SBCE with a discordant procedure laterality; e.g., detecting an ipsilateral SBCE via a mammogram on the contralateral side. In order to promote parameter identifiability in the overall model, we estimated mammography sensitivity and specificity but fixed the biopsy false negative rate at 0.02 and false positive rate at 0, which are reasonable given modern biopsy accuracy rates (Dillon et al., 2005). To accommodate different misclassification probabilities depending on the procedure type and side, we used a time-dependent emission distribution.

We selected the disease process and DDO model on the basis of the Bayesian Information Criterion (BIC).

6.2.7 Model Fitting Details

We maximized the likelihood of the models numerically using the BFGS method as implemented in the R function optim. Our EM algorithm lacked full functionality for the model's specification of a time dependent emission distribution. Both methods are sensitive to starting values, and may converge to local rather than global maxima. For covariate free models, we fit models by randomly

selecting 5 starting values from Normal(mean=-2,sd=1) for rate and observation time model parameters and Normal(mean=0,sd=1/3) distributions for emission parameters. We based starting values for multiple covariate models on MLEs for single covariate models. Numeric calculation of the Hessian of the likelihood at the MLE provided the information matrix for the parameter estimates.

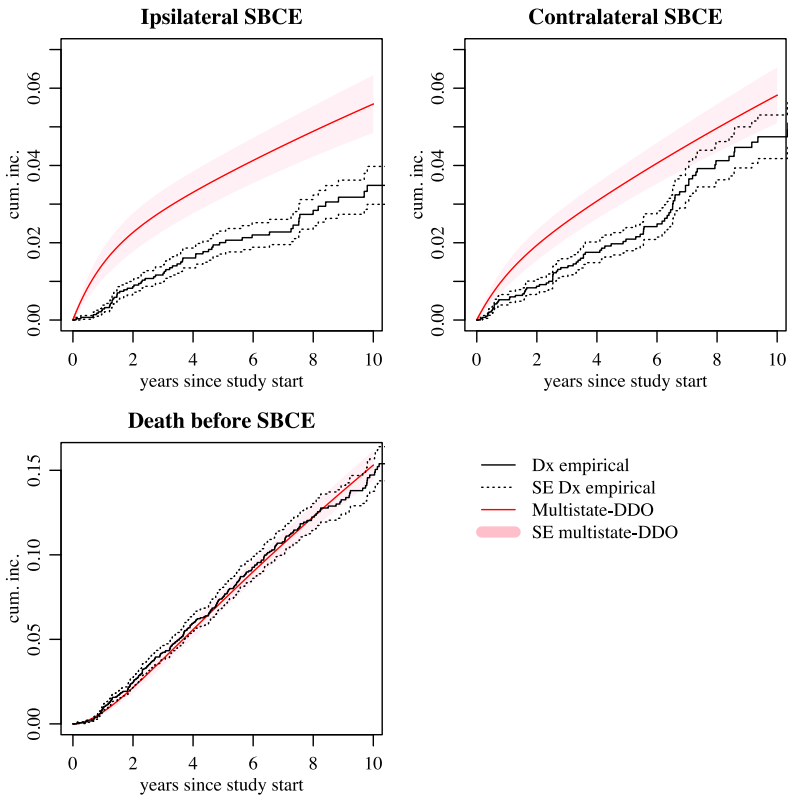

Figure 6.2: Estimated cumulative incidence for ipsilateral and contralateral SBCEs and death, via empirical estimates of the diagnosis times or using the BIC-selected multistate-DDO model (Table 6.1, model 6). The bands are point-wise standard errors. Abbreviations: Dx empirical=empirical estimate of cumulative incidence of diagnosed SBCE events; SE=standard error.

6.3 Results

6.3.1 Data description

There are 2,936 women in the analysis sample, with a median follow-up time of 5.8 years (interquartile range 2.8-9.2). The 2,936 women in the sample used for analysis, as well as the 1,197 excluded from the sample, are described in Table 6.3. The sample was predominantly white (84.7%, N=2,488), with a median age of 61 at primary BC diagnosis (IQR 52, 71). Approximately one fifth of the sample had a stage 0 (DCIS) primary BC (18.6%, N=548), whereas half had stage 1 (49.6%, N=1,456), and the rest, stage 2 or higher. The main difference between included and excluded women is that excluded individuals were more likely to have stage 2 or higher cancer. This is related to our exclusion of individuals with biopsies not preceded by mammograms within the study period being more likely to have advanced stage primary BC.

There were 14,288 contralateral and 10,468 ipsilateral mammograms and 241 contralateral and 212 ipsilateral biopsies. There are fewer ipsilateral than contralateral mammograms because some women were treated for their primary cancer with mastectomy and thus no longer require disease surveillance on the ipsilateral side. The results of the mammograms and biopsies are shown in Table 6.2. There were 84 women diagnosed with contralateral SBCEs and 64 diagnosed with contralateral SBCEs. Approximately 7% of all mammograms and 33% of biopsies were positive. Overall, there were 280 mammograms coded as patient-initiated informative visits. On average, women had 0.98 scheduled mammogram visits per person-year. In contrast, rates of informative visits were low: 0.018 per person-year.

Table 6.2: Outcomes for mammograms and biopsies by procedure laterality.

			Observed result		
Procedure type	Laterality	Total	Healthy	Ipsi.	Contra.
Mamm.	Contra.	14,288	13,305	0	983
	Ipsi.	10,468	9,800	668	0
Biopsy	Contra.	241	157	0	84
	Ipsi.	212	148	64	0

Table 6.3: Characteristics of the GH patients with a history of primary BC, either included in or excluded from the analysis sample. Percentages do not include missing data. Abbrevations: ER+=estrogen receptor positive, PR+=progesterone receptor positive.

	Included (N=2,936)		Excluded (N=1,197)	
	N	%	N	%
Age at diagnosis				
<50	557	19	264	22.1
50-59	801	27.3	330	27.6
60-69	757	25.8	281	23.5
70+	821	28	322	26.9
Missing	0		0	
Race				
White	2488	84.7	1005	86.6
Black	83	2.8	34	2.9
Asian	189	6.4	48	4.1
Other	176	6	73	6.3
Missing	0		37	
Stage of primary cancer				
0	548	18.7	138	14.1
1	1456	49.6	425	43.4
2+	932	31.7	417	42.6
Missing	0		217	
ER+ or PR+ for primary cancer				
No	386	16.3	165	17.5
Yes	1984	83.7	779	82.5
Missing	556		253	
Treatment of primary breast cancer				
Mastectomy				
None	18	0.6	24	2.3
Partial	1925	66.4	711	66.9
Complete unilateral	955	33	328	30.9
Missing	38		134	
Radiation				
No	943	33.3	323	30.9
Yes	1891	66.7	723	69.1
Missing	102		151	26.9
Chemotherapy				
No	2054	70.2	704	63.3
Yes	874	29.8	409	36.7
Missing	8		84	
Adjuvant endocrine therapy				
No	1464	49.9	500	50.8
Yes	1472	50.1	485	49.2
Missing	0		212	

Table 6.4: Model fitting results for SBCE disease and informative sampling time models.

	Disease Model						
	Standard CTMC			Latent CTMC			
	DDO model			DDO model			
	non-inf.	H/I,C	H/I/C	non-inf.	H1,H2/I,C	H1/H2/I,C	H1/H2/I/C
Model label	1	2	3	4	5	6	7
LL	-9,166	-9,155	-9,154	-9,141	-9,131	-9,103	-9,102
no. params	6	7	8	10	11	12	13
BIC	18,381	18,366	18,373	18,362	18,349	18,302	18,308

6.3.2 Model fitting results

The BIC is lowest for the latent CTMC disease model and $H_1/H_2/I,C$ DDO model, where rates of DDO times are allowed to vary in the two healthy states, but are equal in ipsilateral and contralateral SBCE states (see Table 6.4 for model comparison). The estimated DDO rate in state H_1 is 0.046/person-year (95% CI (0.036,0.058)); in H_2 it declines to 0.009/person-year (95% CI (0.007,0.012)); and in the SBCE disease states it is 0.076/person-year (95% CI (0.047,0.11)). These rate estimates are plausible given that patients may be more likely to exhibit symptoms or to initiate visits close to their primary BC diagnosis, as well as after they have developed an SBCE.

6.3.3 Estimates of Cumulative incidence of Ipsilateral and Contralateral SBCEs

Figure 6.2 plots estimates of cumulative incidence of mammographically-detectable SBCEs based on the BIC-preferred multistate-DDO model, in addition to empirical cumulative incidence of diagnosed SBCE events. The multistate-DDO model estimates that at five years after diagnosis 3.7% (95% CI [2.6,4.8]) of women will have a mammographically-detectable ipsilateral SBCE, whereas 2% (95% CI [1.14,2.6]) will have been diagnosed. Likewise, at five years, the multistate-DDO model estimates 3.6% (95% CI [2.6,4.5]) will have a contralateral SBCE, whereas 2.4% (95% CI [1.9, 2.9]) will have been diagnosed. In general, the BIC-preferred DDO model estimates that at each time between five and ten years after the primary BC 25-45% of prevalent SBCEs are undiagnosed.

6.3.4 Mammography misclassification

The multistate-DDO models allow us to estimate true and false positive rates for mammograms. Based on the BIC-selected multistate-DDO model, the estimate of the true positive rate is 69% (95% CI (55%,81%)), and the false positive rate is 5.6% (95% CI (5.3%, 5.9%)). These results are comparable with empirical estimates of mammography sensitivity of 65.4% (95% CI, (61.5%, 69.0%)) and specificity of 98.3% (95%CI (98.2%, 98.4%)) from the Breast Cancer Surveillance Consortium (BCSC), of which GH is a participating institution (Houssami et al., 2011), as well as a recent meta analysis reporting mammography sensitivity ranges of 64-67% and specificity ranges of 85-97% across studies (Robertson et al., 2011).

6.3.5 Sensitivity of covariate-free model results to assumptions

We examined how results differed if we had assumed a CTMC disease model or a non-informative observation model for the patient-initiated visit times. Unlike the BIC-selected latent disease model, the standard CTMC disease model was unable to capture higher SBCE cumulative incidence in the first five years after BC diagnosis (Figure 6.3). Further, assuming no informative observations yielded left-shifted cumulative incidence estimates relative to models allowing for DDO rates to differ across disease states. While these results are consistent with the simulation studies examining bias due to ignoring informative sampling times (Figure 5.4), the magnitude of the shift is much more subtle, probably attributable to the low incidence of DDO times. Estimates of mammography true positive rates are also sensitive to choice of disease and DDO model (Table 6.5). Indeed, higher sensitivity estimates are associated with lower estimates of the cumulative incidence of SBCEs across the observation period.

6.3.6 Covariate effects

To explore the possible associations between the covariates and the SBCE events, we examined empirical estimates of cumulative incidence of diagnosis stratified by different covariate levels (Figure 6.4). Marginally, we see that age<50 at diagnosis is associated with higher cumulative incidence of

Table 6.5: Mammography misclassification estimates for different DDO and disease models.

True positive rate				95% CI	
Model label	Disease model	DDO model	Estimate	Lower	Upper
1	Standard CTMC	Non-inf.	0.77	0.63	0.86
3	Standard CTMC	H/I/C	0.81	0.68	0.90
4	Latent CTMC	Non-inf.	0.61	0.46	0.74
6	Latent CTMC	H1/H2/I,C	0.69	0.55	0.81

False positive rate				95% CI	
Model label	Disease model	DDO model	Estimate	Lower	Upper
1	Standard CTMC	Non-inf.	0.056	0.053	0.059
3	Standard CTMC	H/I/C	0.056	0.053	0.059
4	Latent CTMC	Non-inf.	0.055	0.053	0.058
6	Latent CTMC	H1/H2/I,C	0.056	0.053	0.059

ipsilateral SBCEs, and lower incidence of death before SBCE. Stage 0 (in-situ) cancer is associated with substantially higher cumulative incidence of ipsilateral SBCEs, but lower cumulative incidence of death. Likewise, adjuvant hormone treatment associated with reduced cumulative incidence of ipsilateral SBCEs, but has little impact on contralateral SBCEs or death.

Point estimates for the covariate parameters within the BIC-selected multistate-DDO model are shown in Table 6.6. For the purpose of comparison, we also estimated covariate effects for an analogous latent CTMC disease model based on time of diagnosis, the modeled event in conventional studies of SBCEs, as well as a model specifying a non-informative observation process (Table 6.1, Model 4). Estimates for covariate effects were quite similar between the multistate-DDO and diagnosis-time models, with the exception of effect sizes for age and primary cancer stage on ipsilateral SBCEs. Interestingly, covariate effects were not only similar between diagnosis and multistate-DDO models, they also were relatively robust to specification of the informative sampling time model.

The models indicated overall significant covariate effects on rates of ipsilateral disease (Wald test $p<0.001$), but not contralateral SBCEs (Wald p-values ranged from 0.6-0.84). Further, although the chosen covariate parameterization does not imply proportional hazards, inspection of estimated

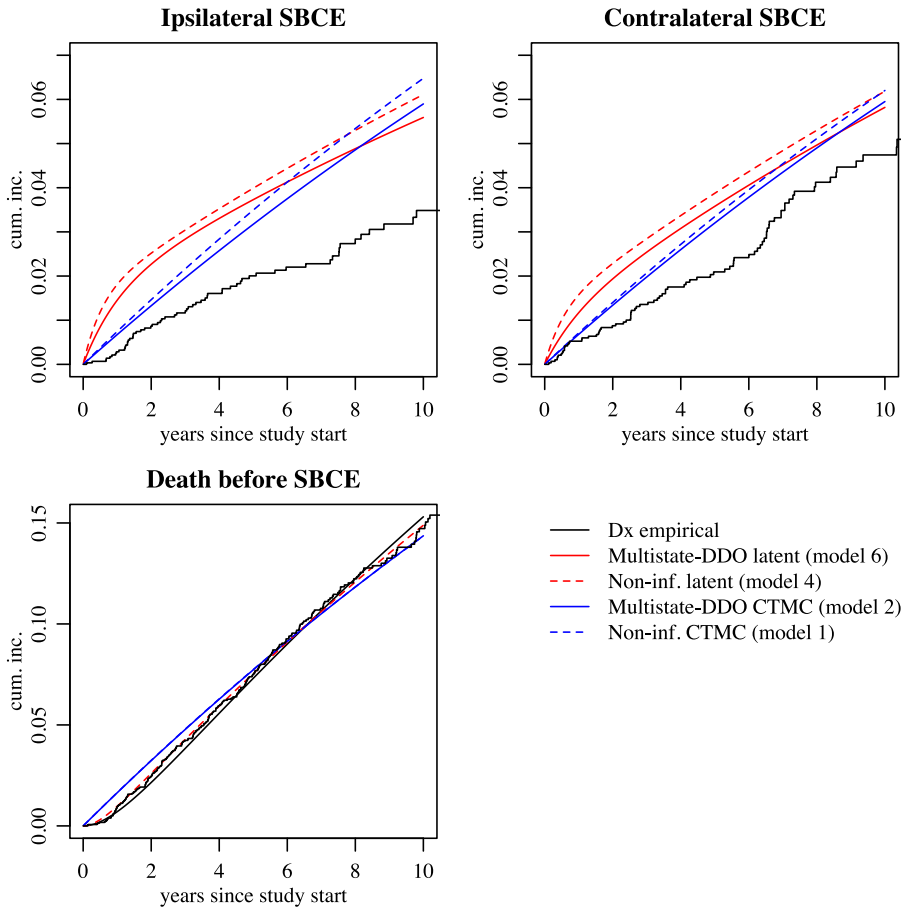

Figure 6.3: Sensitivity of SBCE cumulative incidence estimates to choice of disease and observation model. Table 6.1 shows model details. Models include informative multistate-DDO models (models 2 and 6), and misspecified non-informative observation models (models 1 and 4). Abbreviations: Dx empirical=empirical estimate of cumulative incidence of diagnosed SBCE events.

hazard ratios revealed they were very near constant over time. Thus exponentiated coefficient estimates are approximately interpretable as having multiplicative effects on hazards.

Adjusting for other factors, those with stage 0 (ductal in-situ) SBCEs had the highest rates of ipsilateral SBCEs, compared to other stages. Age <50 at diagnosis was also associated with increased ipsilateral recurrence and hormone treatment was associated with reduced rates of ipsilateral recurrence. These patterns are compatible with results from the Breast Cancer Surveillance Consortium

at large, with the exception that they found adjuvant hormone treatment also reduced the risk of contralateral SBCEs (Buist et al., 2010). Our findings were also consistent with the general literature's reports of higher rates of local recurrence in younger women (de Bock et al., 2006; Vrieling et al., 2003), and lower rates of local recurrence in women treated with hormone-based therapies (Andreetta and Smith, 2007). We found no evidence for association between minority status and any of outcomes, in contrast to literature suggesting that African American women in particular may have higher BC recurrent rates due to differences in disease biology or treatment access (Moran et al., 2008). However, our sample was predominantly white, and our power to detect any difference was likely limited.

Table 6.6: Coefficient estimates for a covariate-parameterized version of the BIC-selected SBCE multistate-DDO (M-DDO) model (Table 6.1, model 6) and an analogous latent CTMC competing risks disease model based on time of diagnosis (Dx)

		Ipsilateral			Contralateral			Death		
			95% CI			95% CI			95% CI	
	Model	Est.	Low.	Upp.	Est.	Low.	Upp.	Est.	Low.	Upp.
Endocrine.	Dx	-0.89	-1.50	-0.28	-0.06	-0.52	0.4	-0.19	-0.45	0.07
therapy	M-DDO	-0.87	-1.47	-0.27	-0.07	-0.52	0.38	-0.21	-0.47	0.05
Age < 50	Dx	0.45	-0.09	0.99	-0.36	-0.98	0.26	-0.81	-1.20	-0.42
	M-DDO	0.69	0.18	1.20	-0.28	-0.89	0.33	-0.8	-1.20	-0.40
Stage 1	Dx	-0.6	-1.18	-0.02	0.32	-0.31	0.95	0.5	0.07	0.93
(ref stage 0)	M-DDO	-0.84	-1.4	-0.28	0.33	-0.32	0.98	0.49	0.06	0.92
Stage 2+	Dx	-0.46	-1.18	0.26	0.09	-0.65	0.83	1.17	0.73	1.61
(ref stage 0)	M-DDO	-0.47	-1.15	0.21	0.22	-0.52	0.96	1.17	0.72	1.62
Non-white	Dx	-0.18	-0.92	0.56	-0.14	-0.8	0.52	-0.35	-0.76	0.06
ethnicity	M-DDO	-0.14	-0.87	0.59	-0.13	-0.79	0.53	-0.33	-0.74	0.08

6.4 Discussion

This analysis applied the multistate-DDO model to a study of the development of SBCEs in women with a personal history of breast cancer, utilizing mammography data to study the onset of mam-

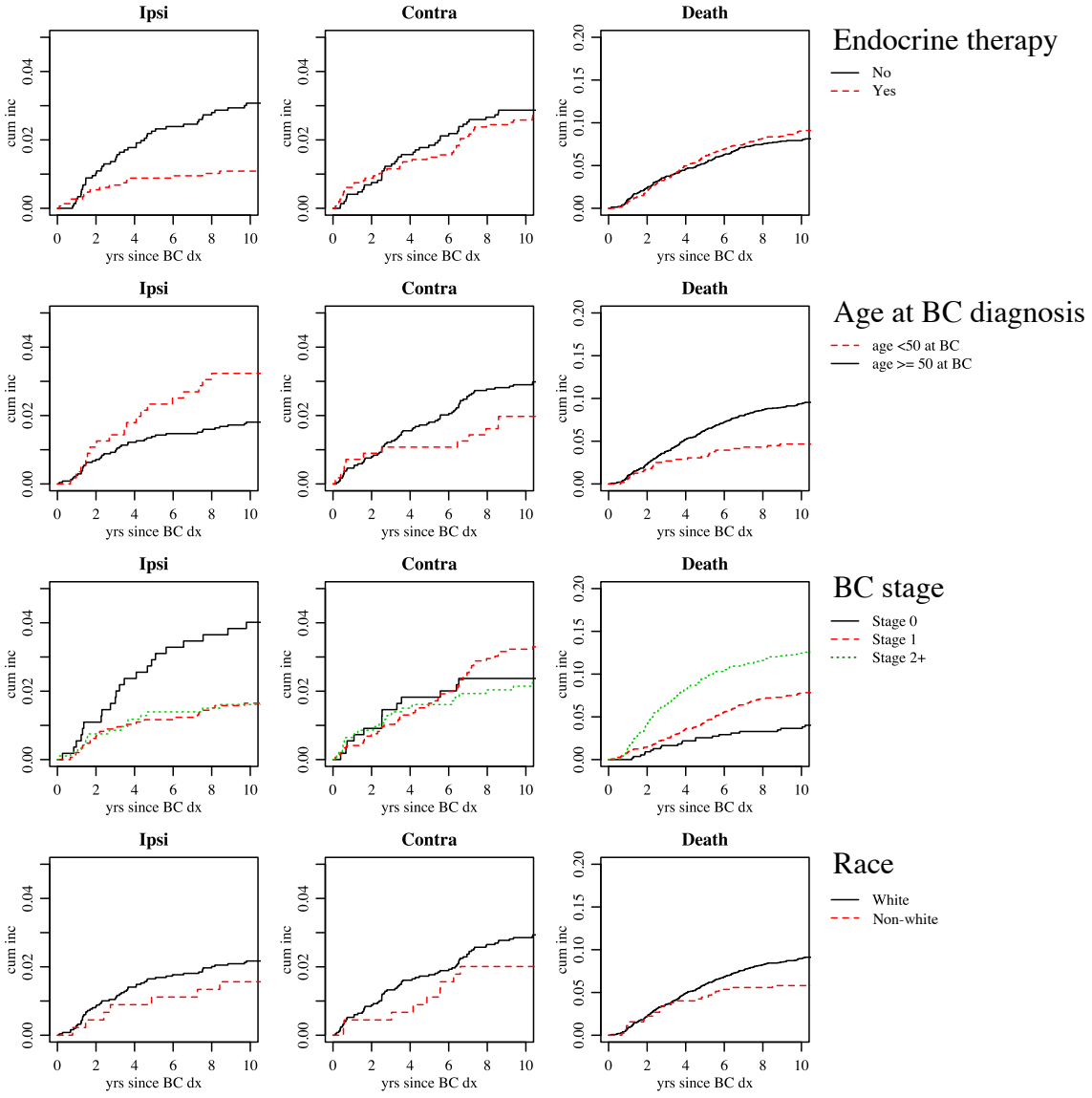

Figure 6.4: Empirical cumulative incidence estimates for diagnosis of ipsilateral and contralateral SBCEs and death prior to SBCE, stratified by covariate levels.

mographically detectable ipsilateral and contralateral secondary cancers. In general, our preferred DDO models estimates that a range of 25-45% of prevalent SBCES are undiagnosed at each time point between five and ten years after the primary BC, demonstrating the clinical significance of our results and the potential benefit of a more sensitive test for improvement of early disease detection

in this population. Our modeling also enabled us to explore covariate effects or rates of mammographically detectable disease. Overall, we found that young age, adjuvant hormone treatment, and state 0 primary BC were associated with higher rates of ipsilateral disease; though the covariates we investigated were not significantly associated with contralateral disease. We should note that there are many other potentially predictive factors for ipsilateral and contralateral SBCEs that we did not consider in this analysis, such as family history and surgical and adjuvant primary BC treatments, and any of our findings may be confounded by unaccounted for variables.

6.4.1 Sensitivity to model assumptions

The multistate-DDO model's parametric assumptions make it sensitive to model-misspecification. Our sensitivity analyses suggested that misspecification of the disease model impacts both estimates of disease cumulative incidence and mammography true positive rates. We have reason to believe that the BIC-selected latent CTMC disease model is likely reasonable. In women with unilateral primary BC, ipsilateral SBCEs reflect both recurrences and new primary cancers, which is consistent with hazard functions that are relatively high near the primary BC diagnosis and flatten out over time (Demicheli et al., 1996). Contralateral SBCEs reflect only new primary cancers, stochastic events with approximately constant rates over time. Hazard estimates from the selected latent CTMC were consistent with this basic pattern, although they did depict a slight decline in contralateral SBCE rates rather than suggesting they are merely constant. Moreover, the estimated mammography sensitivity from the model agrees quite well with empirical estimates from other studies (Houssami et al., 2011; Robertson et al., 2011), providing additional support that the disease model is not grossly misspecified.

We also examined the sensitivity of estimates of SBCE cumulative incidence and mammography sensitivity to our informative sampling time model and definition. The disease process estimates based on the "non-informative" DDO models were shifted to the left compared to those based on models that accounted for different rates in the diseased and healthy states (Figure 6.3). We expected this direction of bias given simulation results (Figure 5.3), and the intuition that "non-informative" models fail to acknowledge that sampling times related to disease status yield data with a shorter

duration between disease onset and diagnosis times. However, the magnitude of this effect was relatively small, which may be due to the low rates of informative visits in the population.

6.4.2 Comparison to other methods

Our application of the multistate-DDO model to the study of SBCEs represents a new analysis method in this setting. Existing studies of secondary BCs focus on diagnosis as the primary outcome (Chapman et al., 1999; Geiger et al., 2007; Buist et al., 2010); our method uses patient mammography data to model onset of mammographically-detectable disease, a clinically relevant outcome that indicates the fraction of a screened population at a given time with undetected disease. Further, others have studied mammography visit patterns in BC survivors (Wirtz et al., 2014), as well as the relationship between screening mammography and mortality (Buist et al., 2013), but our approach is unique in its joint modeling of disease and mammography visit processes.

The multistate-DDO approach for the SBCE data bears similarities to models developed for disease screening trials (Boer et al., 2004); both model onset of screen-detectable disease and estimate screen sensitivity. However, there are important differences between the two approaches. Disease screening models consider progression to a single disease state that is divided into symptom-free pre-clinical and symptomatic clinical sub-states. In contrast, the multistate-DDO model can handle more complicated disease frameworks, such as the SBCE model's competing risk scenario, but does not distinguish between pre-clinical and clinical sub-states. Indeed, the multistate-DDO model reflects symptom-development implicitly through the informative visit process; DDOs based on symptoms occur more frequently in diseased states but may also occur when the patient is healthy. Ultimately, the flexibility of the multistate-DDO model framework enabled us to address our scientific questions, which were descriptive rather than explicitly aimed at developing screening protocols.

6.4.3 Disadvantages of the multistate-DDO method for the SBCE application

The main challenge of method in this setting is that it demands extensive information on the participants, including complete ascertainment of mammograms and biopsies, with accurate coding of

informative visits in order to provide valid results. In order to obtain a clean dataset for analysis, we had to exclude a large proportion of individuals whose data did not meet the model's specifications. This has the potential to limit our ability to generalize to the population at large. In addition to high demands on data completeness and quality, there are challenging technical aspects of fitting these models, including sensitivity to starting values and relatively slow rates of convergence. The latent structure and misclassification component increases the possibility of problems with parameter identifiability, requiring careful consideration before increasing model complexity.

Some of the multistate-DDO model's limitations suggest alternative disease modeling approaches that might be desirable in this context. For example, in the SBCE model sojourn duration in the healthy state coincides with the external times scale of time since diagnosis. Thus, one could use an inhomogeneous standard CTMC disease model rather than a latent CTMC, which might yield models that are easier to fit and do not have latent parameters. In theory, the multistate-DDO model could be modified to accommodate this framework, but additional machinery would be required for likelihood calculations. Estimation in a Bayesian framework might also be useful, as it would allow incorporation of prior information about the disease process or misclassification probabilities and might mitigate concerns about parameter identifiability.

Chapter 7

FUTURE DIRECTIONS

The work presented here has focused on modeling multistate disease processes with latent CTMCs in partial observation settings. Using the modeling framework of Titman and Sharples (2010), we developed an efficient EM algorithm to fit latent CTMC models to disease processes observed at non-informative visit times. We then extended the framework to accommodate disease processes observed at patient-initiated, informative visit times, via the multistate-DDO model. Our applications to the BOS and SBCE studies demonstrated these models' utility in analyzing real data. This chapter discusses future directions for research, including use of a Bayesian approach for model fitting and approaches for goodness of fit evaluation for the multistate-DDO model.

7.1 Bayesian implementations

Our estimation approach has relied on a maximum likelihood framework, and our optimization methods have built on EM algorithms devised for similar contexts. Assuming a Bayesian framework for fitting latent CTMC models is a natural alternative to the maximum likelihood framework and has certain advantages, in particular with respect to mitigating concerns about parameter estimability. With Bayesian estimation, specifying priors for latent parameters will ensure that latent parameters are estimable, even if they cannot be identified from the data via maximum likelihood estimation. Moreover, the Bayesian approach allows us to focus on posterior summaries for functionals rather the latent parameters. Another advantage of the Bayesian framework is the ability to implement automated model selection procedures, e.g., using reversible jump Markov Chain Monte Carlo (MCMC) (Green, 1995).

Future implementations of latent CTMC models for partially observed disease processes will build on others' work in this area. Bladt et al. (2003) described a MCMC algorithm for obtaining

posterior distributions of functionals in phase-type models with known transition times; they extended the method for the discretely observed setting in (Bladt and Sorensen, 2005). Crespi et al. (2005) described a Bayesian procedure for estimating disease process parameters of an HMM based on discretely-observed CTMCs. McGrory et al. (2009) employed automated model selection procedures for phase-type distributions based on hospital length of stay data with known transition times.

After developing Bayesian methods for fitting latent CTMC models, it will be interesting to compare the performance of Bayesian posterior distribution summaries with MLEs. In particular it would be instructive to conduct additional simulations comparing the frequentist properties of both Bayesian and maximum likelihood based estimators of hazard and CDFs when the latent CTMC models are used to approximate generic MSMs. How do bias, mean-squared error, and confidence/credible interval coverage compare between the two approaches?

7.2 Goodness of fit evaluations for multistate-DDO models

Goodness of fit (GOF) evaluation and model diagnostics present additional opportunities for methodology development, particularly in the context of multistate-DDO models. Methods aimed at assessing GOF in discretely observed multistate models, such as Pearson-type statistics comparing observed to expected transition counts across each time interval (Aguirre-Hernández and Farewell, 2002; Titman and Sharples, 2008; Kalbfleisch and Lawless, 1985), are no longer applicable when sampling times are informative. To that end, the time rescaling theorem of Meyer (1971) for general, multivariate point processes shows promise for multistate-DDO GOF evaluation. The basic idea is that transforming DDO times for each category of observed data by their compensator (i.e., cumulative hazard) functions yields independent Poisson processes, one for each observation category, and that GOF can be evaluated via assessment of the Poisson process assumptions.

To see how this theorem applies to a multistate-DDO setting, we assume that all observations times are informative, data consist of possibly misclassified observations of the disease process, and the observed data state space is $\{1,\ldots,r\}$. We also assume that DDO times can occur when an individual is in any state in the disease model; that is, there is no death state. The

observed data at the informative observation times can be construed as realizations of a multivariate counting process $\boldsymbol{N}(t) = \{N_1(t), N_2(t), \ldots, N_r(t)\}$, where $N_i(t)$ counts the number of times state i has been observed at an informative visit time, and jumps in each component occur at non-coinciding times. Let $\{\lambda_1(t|\mathscr{H}_t), \ldots, \lambda_r(t|\mathscr{H}_t)\}$ be the vector of intensity functions for each component of $\boldsymbol{N}(t)$, conditional on the process history. Each component, accordingly, has a compensator function $A_i(t) = \int_0^t (\lambda_i(t|\mathscr{H}_t)dt$. The multivariate time transformation theorem for point processes states that if a mulitvariate point process $\{N_1(t), N_2(t), \ldots, N_r(t)\}$ is formed from times $\{\tau_{ij} : i = 1, \ldots, r; j = 1, \ldots, \infty\}$, such that $A_i(\infty) = \infty$, then the point process formed by the transformed times $A_i(\tau_{ij}), i = 1, \ldots, r, j = 1, \ldots, \infty$ are independent Poisson processes with unit rate (Brown et al., 2002).

One can develop hypothesis tests for evaluating GOF for multistate-DDO models by assessing the independent Poisson process assumptions for the transformed time sequences. A practical approach for investigating this assumption is provided by Gerhard et al. (2011). First, one assesses the Poisson process assumptions individually for each component of the transformed process. Then one assesses the Poisson process assumption for the supposition of the transformed processes. To evaluate the Poisson process assumptions, one uses a Kolmogorov-Smirnov (K-S) test to assess whether transformed inter-arrival times are exponentially distributed. Finally, one tests for independence of the processes by considering whether ordering the observed data by the transformed times is consistent with a random sequence, via a chi-square test on counts of sequential pairs of re-ordered observations. Gerhard et al. (2011) recommend a Bonferroni adjustment for multiple tests to preserve the overall alpha-level. Violations of any of these tests evaluating the independent Poisson process assumption for transformed times suggest the model is not compatible with the observed data.

The implementation of this approach to the multistate-DDO model hinges on our ability to calculate the compensators $A_k(t) = \int_0^t (\lambda_k(t|\mathscr{H}_t)dt$. Our preliminary efforts suggest these calculations are feasible. Assuming $X(t)$ is the underlying disease process with state space S and q_i is the DDO rate when $X(t)$ is in state i, the rate of observations occurring at a DDO time in category k is

$$\lambda_k(t|\mathscr{H}_t) = \sum_{i \in S} e(i,k) q_i P[X(t) = i|\mathscr{H}_t].$$

Thus, we need to be able compute

$$A_k(\tau_{kj}) = \int_0^{\tau_{kj}} \lambda_k(t|\mathscr{H}_t)dt = \sum_{i\in S} e(i,k)q_i \int_0^{\tau_{kj}} P[X(t) = i|\mathscr{H}_t]dt.$$

The computation thus reduces to integrating the forecasting probability $P[X(t) = i|\mathscr{H}_t]$ across the interval $[0, \tau_{kj}]$. To derive this forecasting probability, we first generalize the forward functions of Chapter 5. For ease of presentation, we revert to denoting *all* DDO times as $\tau_1, \ldots, \tau_n$ and their associated data as $o_1, \ldots, o_n$ rather than using the multivariate point process notation. The generalized forward function, denoted $\alpha_t(u)$, is defined in a piecewise fashion. On the interval $t \in [\tau_m, \tau_{m+1}]$,

$$\alpha_t(u) = P[o_1, \ldots, o_m, \tau_1, \ldots, \tau_m, X(t) = u] = \boldsymbol{\alpha}_{\tau_m} \exp\left[(\boldsymbol{\Lambda} - \mathbf{Q})(t - \tau_m)\right] \boldsymbol{e}'_u,$$

where $\boldsymbol{e}'_u$ is a column vector of 0's with 1 in the uth row, and $\boldsymbol{\alpha}_{\tau_m}$ is vector of forward functions at time τ_m, obtained via the formulae in Chapter 5. The forecasting probability can be expressed as

$$P[X(t) = i|\mathscr{H}_t] = \frac{\alpha_t(i)}{\sum_{j\in S} \alpha_t(j)}.$$

Given the piecewise forecasting functions we calculate the integral

$$\int_0^{\tau_m} P[X(t) = i|\mathscr{H}_t] = \sum_{l=1}^{m-1} \int_{\tau_l}^{\tau_{l+1}} P[X(t) = i|\mathscr{H}_t]dt$$

in a piecewise fashion, where

$$\begin{aligned}\int_{\tau_l}^{\tau_{l+1}} P[X(t) = i|\mathscr{H}_t]dt = \int_{\tau_l}^{\tau_{l+1}} \frac{\alpha_t(i)}{\sum_{j\in S} \alpha_t(j)} &= \int_{\tau_l}^{\tau_{l+1}} \frac{\boldsymbol{\alpha}_{\tau_l}}{\boldsymbol{\alpha}_{\tau_l} \times \mathbf{1}'} \exp\left[(\boldsymbol{\Lambda} - \mathbf{Q})(t - \tau_l)\right] \boldsymbol{e}'_i dt \\ &= \frac{\boldsymbol{\alpha}_{\tau_l}}{\boldsymbol{\alpha}_{\tau_l} \times \mathbf{1}'} \left\{exp\left[(\boldsymbol{\Lambda} - \mathbf{Q})(\tau_{l+1} - \tau_l)\right] - I\right\} (\boldsymbol{\Lambda} - \boldsymbol{Q})^{-1} \boldsymbol{e}'_i,\end{aligned} \tag{7.1}$$

where $\mathbf{1}'$ is a column vector of 1s. The matrix $(\boldsymbol{\Lambda} - \mathbf{Q})$ is invertible provided that all diagonal entries of Q are non-zero, which is the case unless one or more states in the disease model permit no informative observations.

7.2.1 Extensions and limitations

The time rescaling approach can also be used to assess the validity of the informative observation time model in the multistate-DDO framework model. To accomplish this, we can assess whether transforming all DDO times $(\tau_1, \ldots, \tau_n)$ by their compensator based on their marginal rate yields a unit rate Poisson process. The compensator can be expressed as the sum of component compensators

$$\sum_{k \in R} A_k(\tau_j) = \sum_{i \in S} q_i \int_0^{\tau_j} P[X(t) = i | \mathscr{H}_t] dt.$$

We should note that global GOF assessment via the time rescaling method only applies when all times are DDO times. The method's use is more limited in the case of a combination of scheduled and DDO times. Implicitly, the scheduled visit times also follow a point process, albeit one that is conditioned on in the multistate-DDO model. It is therefore justifiable to use the entire observed history, including observations at scheduled visits, to calculate the compensator functions for DDO times. However, we cannot use the observed data at scheduled visits to evaluate global GOF.

7.2.2 Future investigations with simulated data

To truly be practical for GOF evaluations the method needs to apply when there are death states in the model. It is an open question what modifications are required when the DDO process may be censored by time of absorption to a death state in the disease process, and demonstrating this will require both more theory work and validating simulation studies. To examine the practical applicability of this method, initial validation of the time-rescaling approach should focus on examining whether the rejection rate of the test is consistent with the alpha-level when the MLEs of data fit with the true model are used to generate the compensators. It will be interesting to consider results both with data generated with reversible disease models and those based on disease models with an absorbing death state. Then, it will be interesting to see the power of the test's performance for detecting different types of misspecified models. Implementation of these simulations will require the use of a K-S test that is compatible with censored data and an implementation of the chi-square test for assessing dependence in a sequence of disease observations reordered by their transformed

times.

7.3 Graphical evaluations based on augmented data

While classical goodness of fit tests ask if the distribution of the observed data is consistent with the posited model, more informal graphical evaluation can aid researchers in identifying where assumed models fall short and suggest how the model might be expanded. In Chapter 4 we evaluated the GOF of the BOS model by comparing the Kaplan-Meier (K-M) estimate of the distribution of time to first BOS diagnosis in the observed data to K-M curves based on times of diagnosis simulated from the fitted model. While this approach was useful for evaluating the reasonableness of the model, time of first BOS diagnosis is only a proxy for what we were truly interested in: time of BOS development. Indeed, it is often desirable to perform model evaluation based on statistics of complete data trajectories rather than just some function of the observed data. This idea is the basis for using augmented data for model evaluation.

Suppose $\boldsymbol{X}$ represents the complete data and $\boldsymbol{Y}$, the observed data. Let $T(\boldsymbol{X})$ be a statistic of the complete data of scientific interest, such as the K-M estimate of BOS onset time, not merely BOS diagnosis. We then compare the distribution of a complete data statistic $T(\boldsymbol{X})$ conditional on the observed data, $\boldsymbol{Y}$, to its unconditional distribution, were the model true, in both cases drawing samples under the MLE $\hat{\theta}$. Conditional samples can be drawn using methods reviewed in Hobolth and Stone (2009). This approach allows us to assess if a particular aspect of the observed data is consistent with the assumed model, accounting for uncertainty in our estimation of such a statistic due to missing data. One can formally test the statistical significance of the degree of discrepancy between statistics of latent trajectories conditional on observed data and their unconditional distributions under the null model using concepts of fuzzy p-values from Thompson and Geyer (2007). The Bayesian analog for this approach is posterior predictive checks, based on work by Gelman et al. (1995). The main difference with the Bayesian setting is that instead of comparing $T(\boldsymbol{X})$ to the conditional distribution under the *MLE*, we compare it to the posterior predictive distribution, which averages over the posterior distribution of $\boldsymbol{\theta}$.

The benefit of the data-augmentation approach for evaluating multistate DDO models is that it

enables us to include all of the data in the goodness of fit assessment and is not limited to observations at DDO times. It also allows us to investigate fit with scientifically meaningful statistics. Of course, its use raises several questions about the power to detect lack of model fit and the true type-I-error rate when using formal tests based on fuzzy-p-values, all of which may be examined with simulated data.

BIBLIOGRAPHY

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

Aguirre-Hernández, R. and Farewell, V. T. (2002). A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models. Statistics in medicine, 21(13):1899–911.

American College of Radiology (2003). Breast Imaging Reporting and Data System (BI-RADS). American College of Radiology, Reston, Va, 4 edition.

Andersen, P. K., Abildstrom, S. Z., and Rosthoj, S. (2002). Competing risks as a multi-state model. Statistical Methods in Medical Research, 11:203–215.

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.

Andreetta, C. and Smith, I. (2007). Adjuvant endocrine therapy for early breast cancer. Cancer letters, 251(1):17–27.

Asmussen, S. (2003). Applied Probability and Queues. Applications of mathematics (Springer).: Stochastic modelling and applied probability. Springer.

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.

Boer, R., Plevritis, S., and Clarke, L. (2004). Diversity of model approaches for breast cancer screening: a review of model assumptions by the Cancer Intervention and Surveillance Network (CISNET) Breast Cancer Groups. Statistical Methods in Medical Research, 13(6):525–38.

Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., and Frank, L. M. (2002). The time-rescaling theorem and its application to neural spike train data analysis. Neural Computation, 14(2):325–46.

Buist, D. S. M., Abraham, L. a., Barlow, W. E., Krishnaraj, A., Holdridge, R. C., Sickles, E. a., Carney, P. a., Kerlikowske, K., and Geller, B. M. (2010). Diagnosis of second breast cancer events after initial diagnosis of early stage breast cancer. Breast Cancer Research and Treatment, 124(3):863–73.

Buist, D. S. M., Bosco, J. L. F., Silliman, R. a., Gold, H. T., Field, T., Yood, M. U., Quinn, V. P., Prout, M., and Lash, T. L. (2013). Long-term surveillance mammography and mortality in older women with a history of early stage invasive breast cancer. Breast cancer research and treatment, 142(1):153–63.

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.

Chan, A. and Allen, R. (2004). Bronchiolitis obliterans: an update. Current Opinion in Pulmonary Medicine, 10(2):133–41.

Chapman, J., Fish, E., and Link, M. (1999). Competing risks analyses for recurrence from primary breast cancer. British Journal of Cancer, 79(9-10):1508–13.

Chen, B., Yi, G. Y., and Cook, R. J. (2010). Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. Statistics in Medicine, 29(11):1175–89.

Chen, B. and Zhou, X.-H. (2011). Non-homogeneous Markov process models with informative observations with an application to Alzheimer's disease. Biometrical Journal, 53(3):444–463.

Chen, B. and Zhou, X.-H. (2013). A correlated random effects model for non-homogeneous Markov processes with nonignorable missingness. Journal of Multivariate Analysis, 117:1–13.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(1):19–29.

Commenges, D. (1999). Multi-state models in epidemiology. Lifetime Data Analysis, 5(4):315–327.

Copeland, C. A. F., Snyder, L. D., Zaas, D. W., Turbyfill, W. J., Davis, W. A., and Palmer, S. M. (2010). Survival after bronchiolitis obliterans syndrome among bilateral lung transplant recipients. American Journal of Respiratory and Critical Care Medicine, 182(6):784–789.

Cox, D. R. (1955). Some Statistical Methods Connected with Series of Events. Journal of the Royal Statistical Society. Series B (Methodological), 17(2):129–164.

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.

Daley, D. J. and Vere-Jones, D. (2003). An introduction to the theory of point processes. Springer, 2nd edition.

de Bock, G. H., van der Hage, J. a., Putter, H., Bonnema, J., Bartelink, H., and van de Velde, C. J. (2006). Isolated loco-regional recurrence of breast cancer is more common in young patients and following breast conserving therapy: long-term results of European Organisation for Research and Treatment of Cancer studies. European Journal of Cancer, 42(3):351–6.

Dean, B. B., Lam, J., Natoli, J. L., Butler, Q., Aguilar, D., and Nordyke, R. J. (2009). Use of electronic medical records for health outcomes research: a literature review. Medical Care Research and Review, 66(6):611–38.

Demicheli, R., Abbattista, A., Miceli, R., Valaguss, P., and Bonadonna, G. (1996). Time distribution of the recurrence risk for breast cancer patients undergoing masectomy: further support about the concept of tumor dormancy. Breast Cancer Research and Treatment, 41(2):177–85.

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.

Diggle, P., Menezes, R., and Su, T.-l. (2010). Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(2):191–232.

Dillon, M. F., Hill, A. D. K., Quinn, C. M., O'Doherty, A., McDermott, E. W., and O'Higgins, N. (2005). The Accuracy of Ultrasound, Stereotactic, and Clinical Core Biopsies in the Diagnosis of Breast Cancer, With an Analysis of False-Negative Cases. Annals of Surgery, 242(5):701–707.

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.

Ephraim, Y. (2012). Bivariate Markov processes and their estimation. Foundations and Trends in Signal Processing, 6(1):1–95.

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.

Fackrell, M. W. (2008). Modelling healthcare systems with phase-type distributions. Health Care Management Science, 12(1):11–26.

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

Faddy, M., Graves, N., and Pettitt, A. (2009). Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. Journal of the International Society for Pharmacoeconomics and Outcomes Research, 12(2):309–14.

Fearnhead, P. and Sherlock, C. (2006). An exact Gibbs sampler for the Markov-modulated Poisson process. Journal of the Royal Statistical Society, 68(5):767–784.

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.

Foucher, Y., Giral, M., Soulillou, J. P., and Daures, J. P. (2010). A flexible semi-Markov model for interval-censored data and goodness-of-fit testing. Statistical Methods in Medical Research, 19(2):127–145.

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

Freed, D. S. and Shepp, L. A. (1982). A Poisson process whose rate is a hidden Markov process. Advances in Applied Probability, 14(1):21–36.

Frydman, H. and Szarek, M. (2009). Nonparametric estimation in a Markov "illness-death" process from interval censored observations with missing intermediate transition status. Biometrics, 65(1):143–51.

Geiger, A. M., Thwin, S. S., Lash, T. L., Buist, D. S. M., Prout, M. N., Wei, F., Field, T. S., Ulcickas Yood, M., Frost, F. J., Enger, S. M., and Silliman, R. a. (2007). Recurrences and second primary breast cancers in older women with initial early-stage disease. Cancer, 109(5):966–74.

Gelman, A., Meng, X.-l., and Stern, H. (1995). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6:733–807.

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

Gerhard, F., Haslinger, R., and Pipa, G. (2011). Apply the multivarate time-rescaling theorem to neural population models. Neural Computation, 23(6):1452–1483.

Gilbert, P. and Varadhan, R. (2012). numDeriv: Accurate Numerical Derivatives. R package version 2012.9-1.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics, 22(4):403–434.

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

Gruger, J., Kay, R., and Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. Biometrics, 47(2):595–605.

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, X., Tong, X., and Sun, J. (2009). Semiparametric analysis of panel count data with correlated observation and follow-up times. Lifetime data analysis, 15(2):177–96.

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.

Hobolth, A. and Stone, E. a. (2009). Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. The annals of applied statistics, 3(3):1204.

Houssami, N., Abraham, L. a., Miglioretti, D. L., Sickles, E. a., Kerlikowske, K., Buist, D. S. M., Geller, B. M., Muss, H. B., and Irwig, L. (2011). Accuracy and outcomes of screening mammography in women with a personal history of early-stage breast cancer. Journal of the American Medical Association, 305(8):790–9.

Hubbard, R. A., Inoue, L. Y. T., and Fann, J. R. (2008). Modeling nonhomogeneous Markov processes via time transformation. Biometrics, 64(3):843–50.

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.

Jackson, C. H., Sharples, L. D., Thompson, S. G., and Duffy, S. W. (2003). Multistate Markov models for disease progression with classification error. The Statistician, 52(2):193–209.

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.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. Biometrics, 42(4):855–65.

Kolmogorov, A. (1931). Uber die analytischen methoden in der wahrscheinlichkeitsrechnung. Mathematische Annalen, 104(1):415–458.

Lagakos, S. W., Sommer, C. J., and Zelen, M. (1978). Semi-Markov models for partially censored data. Biometrika, 65(2):311–317.

Lagarias, J. C., Reeds, J. a., Wright, M. H., and Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. SIAM Journal on Optimization, 9(1):112–147.

Lama, V. N., Murray, S., Lonigro, R. J., Toews, G. B., Chang, A., Lau, C., Flint, A., Chan, K. M., and Martinez, F. J. (2007). Course of FEV(1) after onset of bronchiolitis obliterans syndrome in lung transplant recipients. American Journal of Respiratory and Critical Care Medicine, 175(11):1192–8.

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

Li, N., Zhao, H., and Sun, J. (2013). Semiparametric transformation models for panel count data with correlated observation and follow-up times. Statistics in Medicine, 32(17):3039–54.

Lindqvist, B. H. (2013). Phase-type distributions for competing risks. In Proceedings of the 59th ISI World Statistics Congress, pages 25–30, Hong Kong, China.

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

Lu, S. (2012). Markov modulated Poisson process associated with state-dependent marks and its applications to the deep earthquakes. Annals of the Institute of Statistical Mathematics, 64(1):87–106.

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.

Mark, B. L. and Ephraim, Y. (2013). An EM algorithm for continuous-time bivariate Markov chains. Computational Statistics & Data Analysis, 57(1):504–517.

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.

Meira-Machado, L., de Una-Alvarez, J., and Cadarso-Suarez, C. (2006). Nonparametric estimation of transition probabilities in a non-markov illness-death model. Lifetime Data Analysis, 12(3):325–344.

Meira-Machodo, L., de Una-Alvarez, J., Cadarso-Suarez, C., and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. Statistical Methods in Medical Research, 18(2):195–222.

Meyer, P.-A. (1971). Démonstration simplifiée d'un théorème de Knight. Séminaire de probabilités de Strasbourg, 5(1):191–195.

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.

Moler, C. and Loan, C. V. (2003). Nineteen dubious ways to compute the exponential of a matrix , twenty-five years later. SIAM Review, 45(1):801–836.

Montoro-Cazorla, D. and Perez-Ocorn, R. (2014). Matrix stochastic analysis of the maintainability of a machine under shocks. Reliability Engineering and System Safety, 121(0):11 – 17.

Moran, M. S., Yang, Q., Harris, L. N., Jones, B., Tuck, D. P., and Haffty, B. G. (2008). Long-term outcomes and clinicopathologic differences of African-American versus white patients treated with breast conservation therapy for early-stage breast cancer. Cancer, 113(9):2565–74.

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

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

Neuts, M. F. (1995). Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Dover Publications, revised edition.

O'Cinneide, C. A. (1989). On non-uniqueness of representations of phase-type distributions. Communications in Statistics. Stochastic Models, 5(2):247–259.

O'Cinneide, C. A. (1990). Characterization of phase-Type distributions. Communications in Statistics: Simulation and Computation, 6(1):1–58.

Putter, H., van der Hage, J., de Bock, G. H., Elgalta, R., and van de Velde, C. J. (2006). Estimation and prediction in a multi-state model for breast cancer. Biometrical Journal, 48(3):366–380.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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.

Robertson, C., Ragupathy, S. K. A., Boachie, C., Fraser, C., Heys, S. D., Maclennan, G., Mowatt, G., Thomas, R. E., and Gilbert, F. J. (2011). Surveillance mammography for detecting ipsilateral breast tumour recurrence and metachronous contralateral breast cancer: a systematic review. European Radiology, 21(12):2484–91.

Ross, S. M. (1996). Stochastic Processes. Wiley and Sons, New York, 2 edition.

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.

Ryden, T. (1996a). An EM algorithm for estimation in Markov-modulated Poisson processes. Computational Statistics & Data Analysis, 21(88):431—-447.

Ryden, T. (1996b). On identifiability and order of continuous time aggregated Markov chains, Markov modulated Poisson processes, and phase-type distributions. Journal of Applied Probability, 33(3):640–653.

Scott, A. I. R., Sharples, L. D., and Stewart, S. (2005). Bronchiolitis obliterans syndrome: risk factors and therapeutic strategies. Drugs, 65(6):761–71.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association, 82(398):605–610.

Siegel, R., DeSantis, C., Virgo, K., Stein, K., Mariotto, A., Smith, T., Cooper, D., Gansler, T., Lerro, C., Fedewa, S., Lin, C., Leach, C., Cannady, R. S., Cho, H., Scoppa, S., Hachey, M., Kirch, R., Jemal, A., and Ward, E. (2012). Cancer treatment and survivorship statistics, 2012. CA: A Cancer Journal for Clinicians, 62(4):220–241.

Steele, R. and Raftery, A. (2010). Performance of Bayesian model selection criteria for Gaussian mixture models. In Chen, M.-H., Muller, P., Sun, D., Ye, K., and Dey, D., editors, Frontiers of Statistical Decision Making and Bayesian Analysis, pages 113–130. Springer.

Sun, J., Park, D.-H., Sun, L., and Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. Journal of the American Statistical Association, 100(471):882–889.

Sundberg, R. (1973). Maximum likelihood theory for incomplete data from an exponential family. Scandinavian Journal of Statistics, 1:49–58.

Sweeting, M. J., Farewell, V. T., and De Angelis, D. (2010). Multi-state Markov models for disease progression in the presence of informative examination times: an application to hepatitis C. Statistics in Medicine, 29(11):1161–74.

Thompson, E. A. and Geyer, C. J. (2007). Fuzzy p-values in latent variable problems. Biometrika, 94(1):49–60.

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. (2011). Flexible nonhomogeneous Markov models for panel observed data. Biometrics, 67(3):780–7.

Titman, A. C. and Sharples, L. D. (2010). Semi-Markov models with phase-type sojourn distributions. Biometrics, 66(3):742–752.

Touraine, C., Helmer, C., and Joly, P. (2013). Predictions in an illness-death model. Statistical Methods in Medical Research.

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.

Vrieling, C., Collette, L., Fourquet, a., Hoogenraad, W., Horiot, J.-C., Jager, J., Bing Oei, S., Peterse, H., Pierart, M., Poortmans, P., Struikmans, H., Van den Bogaert, W., and Bartelink, H. (2003). Can patient-, treatment- and pathology-related characteristics explain the high local recurrence rate following breast-conserving therapy in young patients? European Journal of Cancer, 39(7):932–944.

Wirtz, H. S., Boudreau, D. M., Gralow, J. R., Barlow, W. E., Gray, S., Bowles, E. J. a., and Buist, D. S. M. (2014). Factors associated with long-term adherence to annual surveillance mammography among breast cancer survivors. Breast cancer research and treatment.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. The Annals of Statistics, 11(1):95–103.

Zucchini, W. and MacDonald, I. L. (2009). Hidden Markov Models for Time Series: An Introduction Using R. Chapman & Hall, 1st edition.