Analysis of Multi-State Models with Mismeasured Covariates or Misclassified States

by

Feng He

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics (Biostatistics)

Waterloo, Ontario, Canada, 2015

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Multi-state models provide a useful framework for estimating the rate of transitions between defined disease states and understanding the influence of covariates on transitions in studies of the disease progression. Statistical analysis of data from studies of disease progression often involves a number of challenges. A particular challenge is that the classification of the disease state may be subject to error. Another common problem is that there are many sources of heterogeneity in the data in which situation the assumption of time-homogeneous for common Markov models is not valid. In addition, it is common for discrete covariates subject to misclassification and the panel data collected from disease progression studies is time-dependence in the covariates.

In Chapter 2, the progressive multi-state model with misclassification is developed to simultaneously estimate transition rates and account for potential misclassification. The performance of the maximum likelihood and pairwise likelihood estimators is evaluated by simulation studies. The proposed progressive model is illustrated on coronary allograft vasculopathy data, in which the diagnosis based on the coronary angiography is subject to error.

In Chapter 3, hidden mover-stayer models are proposed to provide a solution to a type of heterogeneity where the population consists of both movers and stayers in the presence of misclassification. The likelihood inference procedure based on the EM algorithm is developed for the proposed model. The performance of the likelihood method is investigated through simulation studies. The proposed method is applied to the Waterloo Smoking Prevention Project.

In Chapter 4, we propose estimation procedures for Markov models with binary covariates subject to misclassification. We show that the model is not identifiable under covariate misclassification. Consequently, we develop likelihood inference methods based on known reclassification probabilities and the main/validation study design. Simulation studies are conducted to inves-

tigate the performance of proposed methods and the consequence of the naive analysis which ignores the misclassification.

In Chapter 5, we consider two-state Markov models where time-dependent surrogate covariates are available. We exploit both functional and structural inference methods to reduce or remove bias effects induced from covariate measurement error. The performance of proposed methods is investigated based on simulation studies.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Grace Yi, for her inspirational and patient guidance. Apart from the research, Dr. Yi is kind in person, I have gained inspiration during our conversation. Her comments have benefited me not only on my study, but also my daily life.

I thank Dr. Richard Cook, Dr. Jane Law (School of Public Health and Health Systems at the University of Waterloo), Dr. Sudhir Paul (University of Windsor), and Dr. Leilei Zeng for serving as my committee members.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Multi-state models

Multi-state models provide a useful framework for estimating the rate of transitions between defined disease states and understanding the influence of covariates on transitions in studies of the disease progression (Andersen and Keiding, 2002; Commenges, 1999, 2002). Examples of medical applications include breast cancer (Duffy *et al.*, 1995; Chen *et al.*, 1996, 1997, 2000; Hsieh *et al.*, 2002; Chang *et al.*, 2007), cervical cancer (Kirby and Spiegelhalter, 1994), chronic myelogenous leukemia (Klein *et al.*, 1984), coronary occlusive disease after heart transplants (Sharples, 1993; Klotz and Sharpless, 1994), dementia (Joly *et al.*, 2002), diabetic complications (Andersen, 1988; Marshall and Jones, 1995; Kosorok and Chao, 1996), Giardia lamblia (Nagelkerke *et al.*, 1990), hairy leukoplakia (Bureau *et al.*, 2003), hepatitis C virus (Sweeting *et al.*, 2010), human immunodeficiency virus (HIV) (Longini *et al.*, 1989; Gentleman *et al.*, 1994; Satten and Longini, 1996; Aalen *et al.*, 1997; Satten, 1999; Guihenneuc-Jouyaux *et al.*, 2000; Alioum *et al.*, 2005; Binquet *et al.*, 2009), hepatocellular carcinoma (Kay, 1986), human papillomavirus (Bureau *et al.*, 2003;

Kang and Lagakos, 2007), liver cirrhosis (Andersen *et al.*, 1991), psoritic arthritis (Cook *et al.*, 2004; Sutradhar and Cook, 2008; Tolusso and Cook, 2009; Chen *et al.*, 2010; Farewell and Su, 2011; O'Keeffe *et al.*, 2011; Tom and Farewell, 2011; O'Keeffe *et al.*, 2013), screening for abdominal aortic aneurysms (Jackson *et al.*, 2003), and smoking prevention (Kalbfleisch and Lawless, 1985; Cook *et al.*, 2002; Chen *et al.*, 2011).

Continuous-time multi-state models are commonly used for characterizing disease processes due to irregularly spaced observation times in the study. The continuous-time multi-state model is a stochastic process, which is a family of random variables $\{S(t) : t \in T\}$ taking values in a discrete set of states $\{1, 2, \ldots, K\}$ with the index set $T = [0, \infty)$. The survival model in Figure 1.1 is the simplest form of multi-state model, with two states and one possible transition from "alive" to "dead".

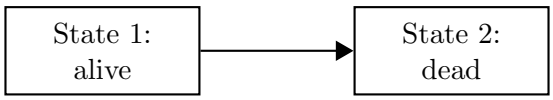

Figure 1.1: Survival model

Other multi-state models include the illness-death model in which individuals make transitions from "healthy" to "dead" possibly via "diseased" (see Figure 1.2) and the unidirectional model in which individuals move among the states sequentially and irreversibly until they reach an absorbing state (see Figure 2.1). Both the illness-death model and the unidirectional model are examples of *progressive models* in which individuals make irreversible transitions and finally reach an absorbing state. The *absorbing* state is a state that once entered, is never left. If the recovery from disease is allowed in the illness-death model, it is called the reversible illness-death model (see Figure 1.4). It is an example of a *bidirectional model*. Such models allow transitions between some transient states in both directions and contain an absorbing state. If the absorbing

state "dead" is excluded from the reversible illness-death model, the model becomes a recurrent two-state model, as in Figure 1.5, which is the simplest example of *recurrent models* that do not contain an absorbing state but include recurrent states. State i is said to be *recurrent* if the probability that, starting in state i, the process will ever return to state i is 1 and *transient* if that probability is less than 1. More details about the model structure and its influence on statistical modelling are summarized in Titman (2007, Section 1.3.1).

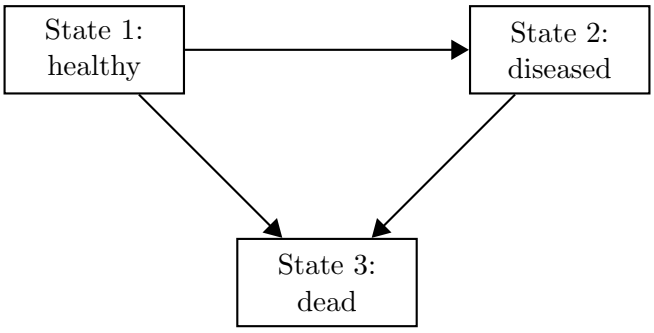

Figure 1.2: Illness-death model

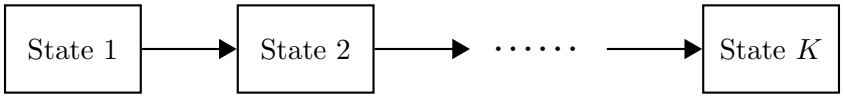

Figure 1.3: K-state unidirectional progressive model

It is of interest to determine the rate of disease onset or progression and identify prognostic variables on transitions in the medical application. Transition intensity functions $q_{ij}\left[t, H\left(t\right)\right]$ of the multi-state model are used to describe the transition rate from state i to j at time t and the history $H\left(t\right)$ of the process up to time t. These functions are defined as

$$q_{ij}\left[t, H\left(t\right)\right] = \lim_{\Delta t \downarrow 0} \frac{\Pr\left[S\left(t+\Delta t\right) = j \mid S\left(t\right) = i, H\left(t\right)\right]}{\Delta t}, \qquad i \neq j,$$

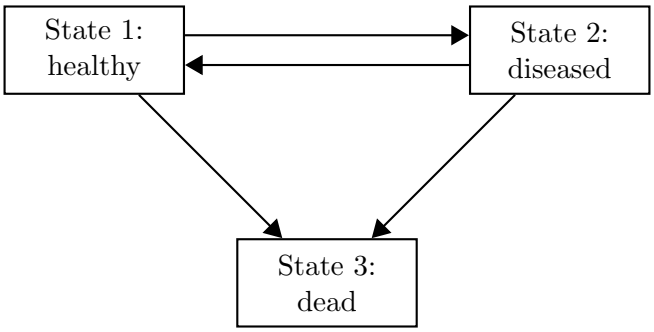

Figure 1.4: Reversible illness-death model

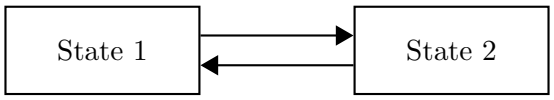

Figure 1.5: Reversible two-state model

where $S(t)$ is the state occupied at time t.

1.1.1 Continuous-time Markov models

The Markov property is usually assumed for simplifying the estimation of the intensity functions in multi-state models; i.e., the conditional distribution of the future state $S(t+s)$ given the present state $S(s)$ and the past states $\{S(u), 0 \leq u < s\}$, depends only on the present state and is independent of the past states. Mathematically, the Markov property is that for all $s, t \geq 0$, and states i, j, $\{x(u), 0 \leq u < s\}$:

$$\Pr[S(t+s) = j \mid S(s) = i, S(u) = x(u), 0 \leq u < s] = \Pr[S(t+s) = j \mid S(s) = i];$$

equivalently, for all nonnegative integers m, time points $t_1 < t_2 < \cdots < t_m$ and states $k_1, \ldots, k_{m-2}$, i, j,

$$\begin{aligned} & \Pr\left[S\left(t_m\right) = j \mid S\left(t_1\right) = k_1, \ldots, S\left(t_{m-2}\right) = k_{m-2}, S\left(t_{m-1}\right) = i\right] \\ = \; & \Pr\left[S\left(t_m\right) = j \mid S\left(t_{m-1}\right) = i\right]. \end{aligned}$$

In terms of transition intensities, the Markov property implies

$$\begin{aligned} q_{ij}\left[t, H\left(t\right)\right] &= \lim_{\Delta t \downarrow 0} \frac{\Pr\left[S\left(t+\Delta t\right) = j \mid S\left(t\right) = i, H\left(t\right)\right]}{\Delta t} \\ &= \lim_{\Delta t \downarrow 0} \frac{\Pr\left[S\left(t+\Delta t\right) = j \mid S\left(t\right) = i, S\left(u\right) = x\left(u\right), 0 \leq u < t\right]}{\Delta t} \\ &= \lim_{\Delta t \downarrow 0} \frac{\Pr\left[S\left(t+\Delta t\right) = j \mid S\left(t\right) = i\right]}{\Delta t} \\ &= q_{ij}\left(t\right), \qquad i \neq j, \end{aligned}$$

so that transition intensities are time-varying but independent of the past history of the process. For convenience, we define $q_{ii}\left(t\right) = -\sum_{j \neq i} q_{ij}\left(t\right)$ such that $\sum_j q_{ij}\left(t\right) = 0$.

Let $\mathbf{P}\left(t, u\right)$ denote the $K \times K$ transition probability matrix with (i, j) entry $\Pr\left[S\left(u\right) = j \mid S\left(t\right) = i\right]$ and $\mathbf{Q}\left(t\right)$ be the $K \times K$ transition intensity matrix with (i, j) entry $q_{ij}\left(t\right)$, where $i, j = 1, 2, \ldots, K$. As is well known (e.g. Cox and Miller, 1965, Chapter 4), the transition probability matrix for continuous-time Markov models satisfies Kolmogorov differential equations:

$$\frac{\partial \mathbf{P}\left(t, u\right)}{\partial u} = \mathbf{P}\left(t, u\right)\mathbf{Q}\left(u\right) \quad \text{and} \quad -\frac{\partial \mathbf{P}\left(t, u\right)}{\partial t} = \mathbf{Q}\left(t\right)\mathbf{P}\left(t, u\right), \tag{1.1}$$

subject to the condition $\mathbf{P}\left(t, t\right) = \mathbf{I}$, where $\mathbf{I}$ is the unit matrix. In most cases, the forward and backward systems of differential equations cannot be solved analytically. But under certain assumptions, such as time-homogeneity, which assumes transition intensities are independent of t,

transition probabilities can be expressed in a convenient way. In this case, transition probabilities depend only on the length of the time interval and the present and future states:

$$P_{ij}(t) = \Pr[S(t+s) = j \mid S(s) = i] = \Pr[S(t) = j \mid S(0) = i], \qquad i, j = 1, 2, \ldots, K.$$

That is, continuous-time homogenous Markov models are stationary. This also implies that the sojourn time within state i, which is the amount of time that the process stays in state i before making a transition into a different state, is exponentially distributed with mean $-1/q_{ii}$. When the process exists from state i, it makes a transition to state j with probability $-q_{ij}/q_{ii}$, where q_{ij} is the constant transition rate $q_{ij}(t)$ from state i to j.

In the time-homogeneous case with constant transition intensity matrix $\mathbf{Q}$, the equations

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{P}(t)\mathbf{Q} \quad \text{and} \quad -\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{Q}\mathbf{P}(t),$$

with the initial condition $\mathbf{P}(0) = \mathbf{I}$ can be solved by the matrix exponential of the element-wise scalar multiplication of the transition intensity matrix and the time interval t

$$\mathbf{P}(t) = \exp(\mathbf{Q}t) = \sum_{r=0}^{\infty} \mathbf{Q}^r \frac{t^r}{r!}$$

where the matrix exponential is defined by the power series of the matrix product and $\mathbf{Q}^0 = \mathbf{I}$.

Instead of using the algorithm for the matrix exponential based on the Taylor series, other algorithms are proposed based on the nature of transition intensity matrix $\mathbf{Q}$. If $\mathbf{Q}$ has distinct eigenvalues, the canonical decomposition for the computation of $\mathbf{P}(t)$ is available (Kalbfleisch and Lawless, 1985). In this case,

$$\mathbf{Q} = \mathbf{H}\mathbf{D}\mathbf{H}^{-1}$$

where $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_K)$ is a diagonal matrix of distinct eigenvalues of $\mathbf{Q}$ and $\mathbf{H}$ is the $K \times K$ matrix whose jth column is the eigenvector associated with d_j. Then, $\mathbf{P}(t)$ is calculated as

$$\mathbf{P}(t) = \mathbf{H} \exp\{\mathbf{D}t\} \mathbf{H}^{-1}.$$

Thus, the computation of the transition probability matrix is easy if the eigenvalue decomposition of $\mathbf{Q}$ is known. If $\mathbf{Q}$ has repeated eigenvalues, Kalbfleisch and Lawless (1985) suggested an analogous decomposition of $\mathbf{Q}$ to the Jordan canonical form (e.g. Cox and Miller, 1965, Chapter 3). More recently, Jackson (2011) recommended a method based on Padé approximation with scaling and squaring (Moler and van Loan, 2003) for cases with repeated eigenvalues. In certain simple situations, the explicit analytic expression of transition probabilities is available (Chiang, 1980), such as the three-state progressive illness-death model (Omar *et al.*, 1995), the three-state reversible illness-death model (Tuma *et al.*, 1979), the five-state progressive illness-death model (Longini *et al.*, 1989), the K-state unidirectional progressive model (Satten, 1999), and several selected 2, 3, 4 and 5-state models (Jackson, 2011).

1.1.2 Hidden Markov models

The hidden Markov model (HMM) consists of two processes $\{(S(t_k), S^*(t_k)) : k \geq 0\}$: the underlying process $\{S(t_k) : k \geq 0\}$ is unobserved and satisfies the Markov property; conditional on the state process $\{S(t_k) : k \geq 0\}$, the observed process $\{S^*(t_k) : k \geq 0\}$ is a sequence of independent random variables such that the conditional distribution of $S^*(t_k)$ depends only on the current state $S(t_k)$. The conditional independence implies that

- for any integer p and any ordered set $\{t_1 < \cdots < t_p\}$ of indices, random variables $S^*(t_1), \ldots, S^*(t_p)$ are conditionally independent given $S(t_1), \ldots, S(t_p)$;

- for any integers k and p and any ordered set $\{t_1 < \cdots < t_p\}$ of indices such that $t_k \notin \{t_1, \ldots, t_p\}$, random variables $S^*(t_k)$ and $\{S(t_1), \ldots, S(t_p)\}$ are conditionally independent given $S(t_k)$.

The discussion of the conditional independence properties of HMMs can be found in Corollary 2.2.5 of Cappé *et al.* (2005).

Figure 1.6 uses a directed graph without loops to describe the dependence structure among random variables in an HMM. The nodes in the graph represent the random variables, and the arrows represent the structure of dependence, i.e. the joint probability distribution can be factored as a product of conditional distributions of each node, given the node's direct predecessors. Figure 1.6 indicates that the conditional distribution of $S(t_m)$, given the process history $S(t_1), \ldots, S(t_{m-1})$, is determined by the value of the previous state $S(t_{m-1})$, which is the Markov property. Likewise, the conditional distribution of $S^*(t_m)$, given past observations $S^*(t_1), \ldots, S^*(t_{m-1})$, and states $S(t_1), \ldots, S(t_m)$, is determined only by the value of the current state $S(t_m)$.

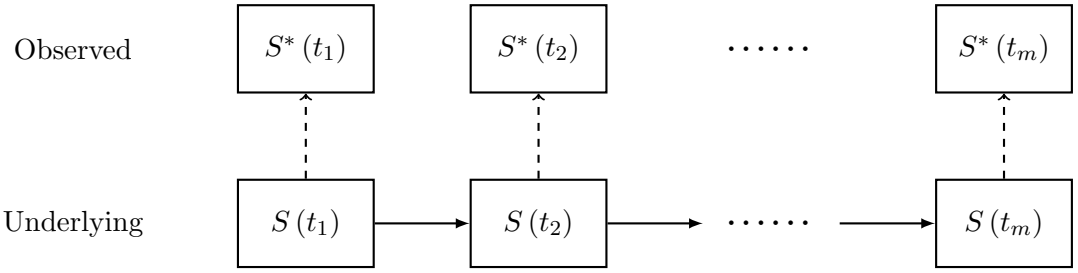

Figure 1.6: Graphical model for the dependence structure of a hidden Markov model

Given the state sequence $\{S(t_k) : k \geq 0\}$, observations $\{S^*(t_k) : k \geq 0\}$ are conditionally independent. But $\{S^*(t_k) : k \geq 0\}$ is not an independent sequence. Furthermore, $\{S^*(t_k) : k \geq 0\}$ does not satisfy the Markov property in general, even though both $\{S(t_k) : k \geq 0\}$ and

$\{(S(t_k), S^*(t_k)) : k \geq 0\}$ are Markov chains. In particular, Zucchini and MacDonald (2009, Page 39) gave a counterexample to show that a two-state Bernoulli HMM does not satisfy the Markov property. On the other hand, Spreij (2001) provided an answer to the question of conditions under which a hidden Markov model satisfies the Markov property.

Because the state sequence $\{S(t_k) : k \geq 0\}$ is not observable but satisfies the Markov property, the term hidden Markov model is used to refer to such models. They are also called *Markov-dependent mixture models* (Leroux and Puterman, 1992) in that the observation $\{S^*(t_k) : k \geq 0\}$ is conditionally independent on the states $\{S(t_k) : k \geq 0\}$, which are generated from a mixing distribution with a Markov property. Alternatively, HMMs can be considered as an extension of Markov models, in which the observation $\{S^*(t_k) : k \geq 0\}$ of the state $\{S(t_k) : k \geq 0\}$ is distorted in some way that includes some additional, independent randomness. For example, the state of cardiac allograft vasculopathy cannot be accurately assessed for patients due to the impossibility of the golden standard test of intravascular ultrasound. Instead, the state is classified based on coronary angiography (Sharples *et al.*, 2003). For another example, the state of HIV progression is not observable and therefore must be defined on the basis of CD4 cell count values (Guihenneuc-Jouyaux *et al.*, 2000).

1.1.3 Time-inhomogeneous Markov models

It is not necessarily realistic that transition intensities stay constant through time. For example, as patients are likely to accept the new therapy to improve the quality of their life, transition rates of the disease may change over time. Therefore, time-inhomogeneous Markov models in which transition intensities depend only on the states and current time are utilized to model time-dependent intensities.

A common method of fitting time-inhomogeneous Markov models is to allow the transition

intensity matrix to be a piecewise constant function (Faddy, 1976) in that transition probabilities are algebraically tractable. Change points $0 = b_0 < b_1 < \cdots < b_M < b_{M+1} = \infty$ are pre-specified and for $t \in [b_k, b_{k+1})$, $\mathbf{Q}(t) = \mathbf{Q}_k$, which implies the constant hazard in each interval. The Kolmogorov differential equations (1.1) have solutions for $s, s+t \in [b_k, b_{k+1})$,

$$\mathbf{P}(s, s+t) = \exp(\mathbf{Q}_k t),$$

which remains the closed form as the time-homogeneous case but substitutes the transition intensity matrix at time interval $[b_k, b_{k+1})$ for the common intensity matrix; transition probabilities between times containing more than one time interval can be found via the Chapman-Kolmogorov equation, such that for $s \in [b_i, b_{i+1})$ and $s+t \in [b_j, b_{j+1})$,

$$\mathbf{P}(s, s+t) = \mathbf{P}(s, b_{i+1}) \prod_{k=i+1}^{j-1} [\mathbf{P}(b_k, b_{k+1})] \mathbf{P}(b_j, s+t).$$

Markov models with piecewise constant transition intensities provide considerable flexibility in term of the time dependence. However, one limitation of these models is that the assumption of homogeneity in each time interval results in deterministic discontinuities in transition intensities, which may not be plausible for some applications (Titman, 2011).

A second approach is time transformation models in which the time-dependent intensity matrix is of the smooth parametric form

$$\mathbf{Q}(t) = \mathbf{Q}_0 \cdot g(t; \lambda),$$

where $\mathbf{Q}_0$ is a fixed intensity matrix with unknown entries, and $g(t; \lambda)$ is a known nonnegative function of time with an unspecified parameter λ. For given λ, let $s = \int_0^t g(u; \lambda)\, \mathrm{d}u$ and define

$Y(s) = S(t)$. Then, the process $\{Y(s) : s \geq 0\}$ is a time-homogeneous Markov process with intensity matrix $\mathbf{Q}_0$; transition probabilities have the solutions

$$\mathbf{P}_s(t_1, t_2) = \mathbf{P}_y(s_1, s_2) = \exp[\mathbf{Q}_0(s_2 - s_1)],$$

where $\mathbf{P}_s$ and $\mathbf{P}_y$ are transition probability matrices for $\{S(t) : t \geq 0\}$ and $\{Y(s) : s \geq 0\}$ respectively and $s_i = \int_0^{t_i} g(u; \lambda)\, \mathrm{d}u$ is an operational time defined by $g(t; \lambda)$ for given λ, $i = 1, 2$. This class of models was first suggested by Kalbfleisch and Lawless (1985). Omar *et al.* (1995) implemented the transformation $g(t; \lambda) = \lambda t^{\lambda - 1}$ for a three-state progressive illness-death model so that the sojourn time within each transient state follows the Weibull distribution with the common shape parameter λ.

Compared with piecewise constant transition intensities models, time transformation models require estimation of fewer parameters and thus observation of fewer times and provide the continuity for intensities, instead of the requirement of the homogeneity assumption in each time interval. However, these models may be quite limited because all the intensities after transformation must be monotonically increasing or decreasing depending on the choice of g and λ. Recently, Hubbard *et al.* (2008) proposed nonparametric time transformation models using a locally weighted smoother to allow more flexibility in dealing with time inhomogeneity. However, these models are still restrictive due to the requirement of a common time-varying multiplicative change for all the intensities, i.e. the ratio of transition intensities $q_{ri}(t)/q_{sj}(t)$ for $i \neq j$, $r \neq i$, $s \neq j$, remains constant through time. It may not be a realistic assumption, for instance, in illness-death models, the transition intensity to death may be expected to increase more rapidly with age than the rate of disease onset.

The third approach is general smooth intensity models (Titman, 2011) in which the intensity

$q_{ij}(t)$ takes the form of a quadratic B-spline with knots $t_{ij1}, \ldots, t_{ijM}$:

$$q_{ij}(t) = q_{ij} \sum_{m=0}^{M} \alpha_{ijm} B_{ijm}(t),$$

where $B_{ijm}(t)$ are spline basis functions, and $\alpha_{ijm} \geq 0$ are weights satisfying the identifiability constrain $\alpha_{ij0} = 1$. To make models identifiable and estimable, the number of knot points needs to be restricted so that there is sufficient information between knot points to estimate spline weights. An upper bound for the number of knot points would be the average number of times each patient is observed in the data, but in most cases fewer points than this would be necessary. The transition probabilities are obtained by numerical solutions to the Kolmogorov differential equations (1.1). Although these models are more flexible to feature the time dependence than time transformation models, it is more computationally intensive than other models, particularly in models with covariates.

1.1.4 Other extensions of Markov models

Mover-stayer models

The mover-stayer model (Blumen *et al.*, 1955) is useful to describe a particular type of population heterogeneity by assuming that the population consists of two types of individuals: the stayer stays in the initial state, whereas the mover evolves according to a Markov process. Therefore, it is a mixture of two independent Markov processes: one with degenerate transition probability matrix equal to the identity matrix at any time and the other with unknown common transition intensity matrix. Note that the proportion of stayers in each state is a time-independent parameter of the model and the transition intensity matrix for each mover does not change over time so the mover-stay model is still time homogeneous. The usual Markov model can be viewed as a special case

when the stayer proportion is zero in each state.

Semi-Markov models

Continuous-time semi-Markov models (Pyke, 1961) are generalizations of Markov models in which sojourn times between transitions have an arbitrary distribution rather than an exponential distribution. In terms of transition intensities, it implies

$$\begin{aligned} q_{ij}\left[t, H\left(t\right)\right] &= \lim_{\Delta t \downarrow 0} \frac{\Pr\left[S\left(t+\Delta t\right)=j \mid S\left(t\right)=i, H\left(t\right)\right]}{\Delta t} \\ &= \lim_{\Delta t \downarrow 0} \frac{\Pr\left[S\left(t+\Delta t\right)=j \mid S\left(t\right)=i, t_i\right]}{\Delta t} \\ &= q_{ij}\left(t_i\right), \qquad i \neq j, \end{aligned}$$

where $t_i \leq t$ is the sojourn time in current state i. Equivalently,

$$q_{ij}\left(t_i\right) = \lim_{\Delta t \downarrow 0} \frac{\Pr\left(S_{n+1}=j, \tau_{n+1} < t+\Delta t \mid S_n=i, \tau_{n+1} \geq t\right)}{\Delta t}, \qquad i \neq j,$$

where S_n denote the nth state occupied by the process and τ_{n+1} represent the sojourn time between the $(n-1)$th and nth states, $n = 1, 2, \ldots$. The semi-Markov process $\{S\left(t\right) : t \geq 0\}$ is defined by

- the time-homogenous transition probability matrix $\mathbf{P}$ of the embedded Markov chain

$$P_{ij} = \Pr\left(S_{n+1}=j \mid S_n=i\right), \qquad i, j = 1, \ldots, K, n \geq 0$$

  with the constrain $\sum_{j=1}^{K} P_{ij} = 1$ for $i = 1, \ldots, K$;

- the conditional distributions $\mathbf{F}(t) = \{F_{ij}(t) : i \neq j\}$ of sojourn times

$$F_{ij}(t) = \Pr(\tau_{n+1} \leq t \mid S_n = i, S_{n+1} = j), \qquad i, j = 1, \ldots, K, i \neq j, n \geq 0.$$

If $F_{ij}(t)$ does not depend on the state next occupied, i.e. $F_{ij}(t) = F_i(t)$ and $F_i(t)$'s are exponential distributions, then the semi-Markov model reduces to the Markov model. The statistical challenge arises from analyzing censored data in which the exact sojourn time for the state is generally unknown.

1.2 Measurement error models

Many variables of interest are difficult to measure precisely on individuals in clinical studies. For example, the presence or absence of a disease is often assessed through an imperfect diagnostic procedure, which can lead to false positives or false negatives; the covariates measured with self report, such as dietary intake and drug use, are subject to error; the actual exposure to certain pollutants is difficult to measure accurately.

In general, a measurement error problem consists of three main ingredients: a model for true values, a measurement error model specifying the relationship between the true and observed values, and possibly additional data information that may be useful to characterize the measurement error. The typical extra data and information include:

1. knowledge of parameters in the measurement error model;

2. replicate values for the error-prone measure of the true value;

3. estimated standard errors for error-prone variables;

4. internal validation data in which true values of error-prone variables are observed on a subset of the main study;

5. external validation data which contain true values of error-prone variables and are independent of the data in the main study;

6. instrumental variables which are correlated with error-prone variables but not correlated with the measurement errors or the errors in the model for true values.

These issues will be discussed in Section 1.2.3.

Two main objectives in a measurement error problem are to investigate the consequence of naive analyses which ignore the measurement error and to carry out corrections for measurement error to obtain valid inference results. These tasks can be carried out according to the inference procedures and measurement error models.

1.2.1 Classical versus Berkson models

When specifying the relationship between the true and observed values, one way is the classic measurement error model (Pearson, 1902; Cochran, 1968) which assumes the distribution of the observed values given the true values, and the other way is the Berkson model which specifies the distribution of the true values given the observed values. The Berkson model was first introduced by Berkson (1950) in cases where the observations are fixed target values but the true values of interest are not observed and random, such as protein levels in a diet in balance/intake studies to determine nutritional requirements.

For the continuous variable, the most widely-used model is the additive measurement error model $X^* = X + U$, where U is a random variable with $E(U \mid X) = 0$. The Berkson version

of the additive model is $X = X^* + U^*$, where U^* is a random variable with $E(U^* \mid X^*) = 0$. Therefore, there is no bias between the observed variable X^* and the true variable X but the true variable has more variability than the observed variable, $\text{var}(X) > \text{var}(X^*)$ in the Berkson model, in contrast with the classical model with $\text{var}(X^*) > \text{var}(X)$. Note that the error structure of U and U^* in both models could be homoscedastic (constant variance) or heteriscedastic.

1.2.2 Misclassification

If both observed and true variables are discrete, the classical measurement error model is given by $\Pr(X^* \mid X)$ where X is the true variable and X^* is the observed variable. This probability is called misclassification probability or misclassification rate. On the other hand, the probability model $\Pr(X \mid X^*)$ is called reclassification probability or reclassification rate. The reclassification and misclassification probabilities are related via

$$\begin{aligned} \Pr(X = x_i \mid X^* = x_j) &= \frac{\Pr(X = x_i, X^* = x_j)}{\Pr(X^* = x_j)} \\ &= \frac{\Pr(X^* = x_j \mid X = x_i)\Pr(X = x_i)}{\sum_k \left[\Pr(X^* = x_j \mid X = x_k)\Pr(X = x_k)\right]}, \end{aligned}$$

where x_i, x_j and x_k are the possible discrete values of the variables.

1.2.3 Measurement error mechanism

The measurement error mechanism is an important assumption imposed when linking the model for true values and the measurement error model (Carroll *et al.*, 2006; Buonaccorsi, 2010). Let Y denote the response variable.

- The measurement error model is *non-differential* (with respect to Y) if the distribution of

the observed variable X^* given the true variable X and the response variable Y does not depend on the value of Y. Otherwise, it is *differential*.

- The observed variable X^* is a *surrogate* for the true variable X (with respect to Y) if the distribution of the response Y given the true variable X and observed variable X^* does not depend on X^*. In other words, given X, X^* contains no information about Y other than what is available in X.

- The response Y and the observed variable X^* are *conditionally independent* given the true variable X if $f(y, x^* \mid x) = g(y \mid x)\, h(x^* \mid x)$, where f is the joint probability function of Y and X^* given $X = x$, g is the conditional probability function of Y given $X = x$, and h is the conditional probability function of X^* given $X = x$.

In fact, the three concepts, non-differential measurement error, surrogacy, and conditional independence, are equivalent.

1.3 Existing methods for continuous-time Markov models

1.3.1 Sampling scheme

Longitudinal data collected from disease progression studies are often under panel/intermittent observation in which the exact times of disease onset or progression are interval censored so that the transition time is only known to lie in a certain interval. It may arise from the intermittent follow-up visits, at which the disease information is collected, but the information between visits is commonly unavailable. A regular balance observation schedule at times t, $2t$, $\ldots$, mt may be specified in advance, but in practice times of visits may vary due to missing or changing scheduled

times. An important exception is the observation time for death which is commonly recorded at the exact time or within one day.

Figure 1.7 illustrates a possible sampling situation. The individual is observed at five visits through four months. The available information is the occupation of states 1, 1, 2, 3, and death at respective times 1.0, 2.0, 3.0, 3.5, 3.9. Except the death date, the entry time of each occupied state and the state occupancy between observation times are unknown.

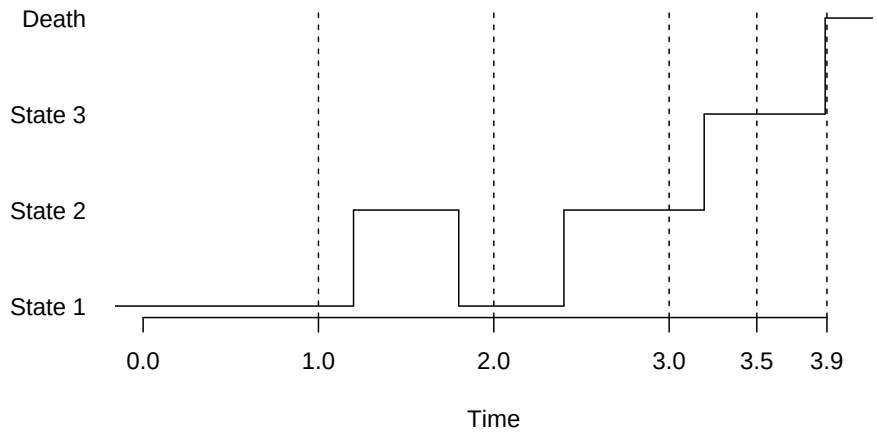

Figure 1.7: Example of panel observation of a multi-state process

Grüger *et al.* (1991) derived two conditions of the interrelationship between the disease process and the sampling scheme under which a valid inference method is possible. Suppose the disease process $S(t)$ for a particular individual is observe at a finite number of times. Let M be the number of observation times and $T_1, \ldots, T_M$ be the observation times. In fact, both observation times $T_1, \ldots, T_M$ and their number M are random variables in applications. Therefore, the observed disease states should be modelled along with the sampling scheme such that the joint

likelihood of the states $s_0, s_1, \ldots, s_m$ and the times $t_0 < t_1 < \ldots < t_m$ is

$$\mathcal{L} = \Pr\left[S\left(t_0\right) = s_0, \ldots S\left(t_m\right) = s_m, T_0 = t_0, \ldots, T_m = t_m, M = m\right].$$

Then, Grüger *et al.* (1991) extended the non-informative censoring in survival analysis (e.g. Kalbfleisch and Prentice, 2002) to the non-informative sampling scheme in multi-state models. That is, the sampling scheme is stochastically independent of the disease process under observation. In terms of the factorization of the likelihood,

$$\begin{aligned}\mathcal{L} &= \Pr\left[S\left(t_0\right) = s_0, \ldots, S\left(t_m\right) = s_m \mid T_0 = t_0, \ldots, T_m = t_m, M = m\right] \\ &\quad \times \Pr\left[T_0 = t_0, \ldots, T_m = t_m, M = m\right],\end{aligned}$$

the condition in the first factor is ignored, and the second factor is assumed to be free of any parameters of

$$\mathcal{L}_0 = \Pr\left[S\left(t_0\right) = s_0, \ldots, S\left(t_m\right) = s_m\right].$$

However, this treatment is not satisfactory, because the independence assumption may not be valid in practice. On the other hand, the likelihood can be factored in a dynamic fashion, conditional on the history H_j of disease states and observation times up to and including the jth observation,

$$\mathcal{L} = \Pr\left(H_0\right)\left\{\prod_{j=1}^{m} \Pr\left[S\left(t_j\right) = s_j \mid T_j = t_j, H_{j-1}\right]\right\}\left\{\prod_{j=1}^{m} \Pr\left(T_j = t_j \mid H_{j-1}\right)\right\}$$

where the history is defined as follows

$$H_j = \left\{T_0 = t_0, S\left(t_0\right) = s_0, \ldots, T_j = t_j, S\left(t_j\right) = s_j\right\}, \qquad j = 0, \ldots, m.$$

Hence, Grüger *et al.* (1991) derived the following conditions for the non-informative sampling scheme:

1. The probability of staying in state s_j at time t_j, given the history $H_{j-1} = \{T_0 = t_0, S(t_0) = s_0, \ldots, T_{j-1} = t_{j-1}, S(t_{j-1}) = s_{j-1}\}$, is independent of whether an observation is carried out at this time and the past observation times, i.e.,

$$\Pr[S(t_j) = s_j \mid T_j = t_j, H_{j-1}] = \Pr[S(t_j) = s_j \mid S(t_0) = s_0, \ldots, S(t_{j-1}) = s_{j-1}];$$

2. The conditional distribution of the jth observation time T_j, $\Pr(T_j = t_j \mid H_{j-1})$, is functionally independent of parameters governing transition intensities of the disease process $\{S(t), t \geq 0\}$.

The non-informative sampling schemes for monitoring the chronic disease progression include, for example,

1. Observation at regular intervals: each patient is observed at regular intervals fixed in advance. Even if observation times are irregular, the sampling scheme is still non-informative as long as it is specified in advance.

2. Random sampling: observation times are independent of the disease histories of individuals under study.

On the other hand, if a patient self-selects the visit to the doctor based on the unwell feeling or the unexpected symptom, this behaviour may violate the first condition and make the observation time informative. Moreover, the valid statistical inference method may not be possible for the informative sampling scheme in general due to the non-identifiability issue, similar to the

competing risk model (Tsiatis, 1975). Meanwhile, estimated transition intensities are subject to potential biases when the informative sampling scheme is ignored. For illustration, Grüger *et al.* (1991) demonstrated the biased estimation of transition intensities under the patient self-selection sampling scheme by analyzing the simulated data.

1.3.2 Likelihood approach

Kalbfleisch and Lawless (1985) and Kay (1986) proposed maximum likelihood methods for the analysis of panel data under continuous-time Markov models. Suppose n independent individuals are under study. The data for individual i consist of observed states $\{s_{i0}, s_{i1}, \ldots, s_{im_i}\}$ at the times $t_{i0} < t_{i1} < \cdots < t_{im_i}$. The observation times are assumed to be non-informative. The parameters related to transition intensities are denoted by $\boldsymbol{\beta}$. Then, the likelihood for individual i is

$$\mathcal{L}_i(\boldsymbol{\beta}) = \prod_{j=1}^{m_i} \Pr\left[S_i(t_{ij}) = s_{ij} \mid S_i(t_{i,j-1}) = s_{i,j-1}; \boldsymbol{\beta}\right],$$

where the conditional probability

$$\Pr\left[S_i(t_{ij}) = s_{ij} \mid S_i(t_{i,j-1}) = s_{i,j-1}; \boldsymbol{\beta}\right]$$

is the entry of the transition probability matrix $\mathbf{P}(t)$ at the $s_{i,j-1}$th row and s_{ij}th column, evaluated at $t = t_{ij} - t_{i,j-1}$.

However, if the death state is included in the multi-state model and recorded at the exact time, then the transition probability from the previous state to the death state is different from that calculated from the transition probability matrix. Suppose the last state of individual i, s_{im_i}, is the death state D which is recorded at the exact time. Then, the transition probability

from state s_{i,m_i-1} to D is of the form

$$\begin{aligned} & \Pr\left[S_i\left(t_{im_i}\right) = D \mid S_i\left(t_{i,m_i-1}\right) = s_{i,m_i-1}; \boldsymbol{\beta}\right] \\ = & \sum_{j \neq D} \left\{ \Pr\left[S_i\left(t_{im_i}\right) = j \mid S_i\left(t_{i,m_i-1}\right) = s_{i,m_i-1}; \boldsymbol{\beta}\right] q_{jD} \right\}, \end{aligned}$$

where the conditional probability $\Pr\left[S_i\left(t_{im_i}\right) = j \mid S_i\left(t_{i,m_i-1}\right) = s_{i,m_i-1}; \boldsymbol{\beta}\right]$ is calculated from the transition probability matrix and q_{jD} is the transition intensity from state j to D.

Another important exception is that the last observation of individual is subject to administrative censoring in the case that the mortality of individuals is followed-up until the end of the study. Typically, there is a gap between the last state observation time t_{m_i}, at which the disease state is observed, and the end time t_{iE}, at which no disease state is known except for the death. In this situation, if individual i is alive at the end of the study, the censored observation has the likelihood contribution

$$\sum_{j \neq D} \Pr\left[S_i\left(t_{iE}\right) = j \mid S_i\left(t_{im_i}\right) = s_{im_i}; \boldsymbol{\beta}\right].$$

The likelihood $\mathcal{L}$ is the product of all the contributions from indidividuals

$$\mathcal{L}\left(\boldsymbol{\beta}\right) = \prod_{i=1}^{n} \mathcal{L}_i\left(\boldsymbol{\beta}\right).$$

For maximum likelihood estimation, Kalbfleisch and Lawless (1985) gave a computationally efficient expression for the first derivatives of transition probabilities and presented a Fisher-scoring algorithm, in which the second derivatives of the log-likelihood are replaced by estimated expectations and thus only the first derivatives are required. However, this procedure can not be used, if the death time is not censored or the final observation does not provide the information of the disease state except the death, due to that expectations of the second derivatives of the log-likelihood

require the second derivatives of transition probabilities. While similar expressions for the second derivates are available in Kosorok and Chao (1995, 1996), they are so complex that it is not worth the effort to supply for the optimization. Instead, the BFGS quasi-Newton method (e.g. Dennis and Schnabel, 1996), in which the second derivatives are approximated by evaluations of the first derivatives, is available to obtain maximum likelihood estimates. To protect against the possible violation of the Markov or time-homogeneity assumption, one may use the sandwich-type robust variance formula (Royall, 1986) to calculate standard errors of parameter estimates.

1.3.3 Extension to mover-stayer models

The likelihood method can be applied for the analysis of panel data under a mover-stayer model. Let Z denote a mover-stayer indicator where $Z = 0$ if the individual is a stayer and $Z = 1$ otherwise. Let $\boldsymbol{\gamma}$ represent the parameters related to the mover-stayer probability, $\boldsymbol{\beta}$ represent the parameters related to transition intensities in the mover process, and $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^{\mathsf{T}}, \boldsymbol{\gamma}^{\mathsf{T}}\right)^{\mathsf{T}}$. If all the observed states of individual i are the same, then individual i is a susceptible mover or a possible stayer. In this case, the likelihood contributed from individual i takes the form

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= \Pr[Z_i = 0 \mid S_i(t_{i0}) = s_{i0}; \boldsymbol{\gamma}] \\ &\quad + \Pr[Z_i = 1 \mid S_i(t_{i0}) = s_{i0}; \boldsymbol{\gamma}] \prod_{j=1}^{m_i} \Pr[S_i(t_{ij}) = s_{i0} \mid S_i(t_{i,j-1}) = s_{i0}, Z_i = 1; \boldsymbol{\beta}].\end{aligned}$$

Otherwise, individual i is a known mover and the likelihood from individual i is given by

$$\mathcal{L}_i(\boldsymbol{\theta}) = \Pr[Z_i = 1 \mid S_i(t_{i0}) = s_{i0}; \boldsymbol{\gamma}] \prod_{j=1}^{m_i} \Pr[S_i(t_{ij}) = s_{ij} \mid S_i(t_{i,j-1}) = s_{i,j-1}, Z_i = 1; \boldsymbol{\beta}].$$

The likelihood can be written as the product over all the contributions from individuals.

The maximum likelihood estimates can be obtained from the direct maximization of the likelihood function based on the Newton-Raphson method. Alternatively, the mover-stayer indicator Z can be treated as a latent variable so the EM algorithm (Dempster *et al.*, 1977) can be used to maximize the likelihood function. The complete data likelihood for individual i is

$$\begin{aligned}\mathcal{L}_i^{\mathrm{c}}(\boldsymbol{\theta}) &= \left\{ \Pr\left[Z_i \mid S_i(t_{i0}) = s_{i0}; \boldsymbol{\gamma}\right] \right\}^{1-Z_i} \\ &\quad \times \left\{ \Pr\left[Z_i \mid S_i(t_{i0}) = s_{i0}; \boldsymbol{\gamma}\right] \prod_{i=1}^{m_i} \Pr\left[S_i(t_{ij}) = s_{ij} \mid S_i(t_{i,j-1}) = s_{i,j-1}, Z_i = 1; \boldsymbol{\beta}\right] \right\}^{Z_i},\end{aligned}$$

and then the complete data log-likelihood is given by

$$\begin{aligned}\ell_i^{\mathrm{c}}(\boldsymbol{\theta}) &= (1 - Z_i) \log \left\{ \Pr\left[Z_i \mid S_i(t_{i0}) = s_{i0}; \boldsymbol{\gamma}\right] \right\} \\ &\quad + Z_i \left\{ \log \left\{ \Pr\left[Z_i \mid S_i(t_{i0}) = s_{i0}; \boldsymbol{\gamma}\right] \right\} \right. \\ &\quad \left. + \sum_{j=1}^{m_i} \log \left\{ \Pr\left[S_i(t_{ij}) = s_{ij} \mid S_i(t_{i,j-1}) = s_{i,j-1}; \boldsymbol{\beta}\right] \right\} \right\}.\end{aligned}$$

It can be divided into two parts: the first part

$$\ell_{i1}^{\mathrm{c}}(\boldsymbol{\gamma}) = \log \left\{ \Pr\left[Z_i \mid S_i(t_{i0}) = s_{i0}; \boldsymbol{\gamma}\right] \right\}$$

contains only the parameters related to the mover-stayer distribution; the second part

$$\ell_{i2}^{\mathrm{c}}(\boldsymbol{\beta}) = Z_i \sum_{j=1}^{m_i} \log \left\{ \Pr\left[S_i(t_{ij}) = s_{ij} \mid S_i(t_{i,j-1}) = s_{i,j-1}; \boldsymbol{\beta}\right] \right\}$$

contains only the parameters related to transition intensities in the mover process. In the expec-

tation step (E-step), the expected complete data log-likelihood at the $(k+1)$th iteration is

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \sum_{i=1}^{n} Q_{i1}(\boldsymbol{\gamma}, \boldsymbol{\theta}^{(k)}) + \sum_{i=1}^{n} Q_{i2}(\boldsymbol{\beta}, \boldsymbol{\theta}^{(k)})$$

where

$$\begin{aligned} Q_{i1}(\boldsymbol{\gamma}, \boldsymbol{\theta}^{(k)}) &= E\left[\ell_{i1}^{\mathrm{c}}(\boldsymbol{\gamma}) \,\middle|\, s_{i0}, \ldots, s_{im_i}, t_{i0}, \ldots, t_{im_i}; \boldsymbol{\theta}^{(k)}\right]; \\ Q_{i2}(\boldsymbol{\beta}, \boldsymbol{\theta}^{(k)}) &= E\left[\ell_{i2}^{\mathrm{c}}(\boldsymbol{\beta}) \,\middle|\, s_{i0}, \ldots, s_{im_i}, t_{i0}, \ldots, t_{im_i}; \boldsymbol{\theta}^{(k)}\right]. \end{aligned}$$

Note that the parameters for the mover process are distinct from those for the mover-stayer distribution. The maximization step can be carried out with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ separately in the M-step. Maximum likelihood estimation can be obtained through iterations between E and M steps until convergence of $\boldsymbol{\theta}^{(k)}$.

The variance estimation in the EM algorithm can be obtained from Louis' Formula (Louis, 1982)

$$\begin{aligned} E\left[-\frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{S})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}\right] &= E_{\boldsymbol{\theta}}\left[-\frac{\partial^2 \ell^{\mathrm{c}}(\boldsymbol{\theta}; \mathbf{S}, Z)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \,\middle|\, \mathbf{S}\right] - E_{\boldsymbol{\theta}}\left\{\left[\frac{\partial \ell^{\mathrm{c}}(\boldsymbol{\theta}; \mathbf{S}, Z)}{\partial \boldsymbol{\theta}}\right]^{\otimes 2} \,\middle|\, \mathbf{S}\right\} \\ &\quad + \left\{E_{\boldsymbol{\theta}}\left[\frac{\partial \ell^{\mathrm{c}}(\boldsymbol{\theta}; \mathbf{S}, Z)}{\partial \boldsymbol{\theta}} \,\middle|\, \mathbf{S}\right]\right\}^{\otimes 2}, \end{aligned}$$

where $\ell(\boldsymbol{\theta}; \mathbf{S})$ is the log-likelihood based on the observed states $\mathbf{S}$ of one individual, $\ell^{\mathrm{c}}(\boldsymbol{\theta}; \mathbf{S}, Z)$ is the complete data log-likelihood based on the mover-stayer indicator Z and the observed state $\mathbf{S}$ of one individual, and $\mathbf{x}^{\otimes 2} = \mathbf{x}\mathbf{x}^{\mathsf{T}}$.

1.3.4 Regression models

It is of interest to investigate the relationship between covariates of each individual and transition intensities of the disease progression. Let $\mathbf{x} = (x_1, x_2, \ldots, x_p)^\mathsf{T}$ denote the $p \times 1$ vector of covariates for an individual. The proportional intensity model,

$$q_{ij}(t) = q_{ij0}(t) \exp\left(\mathbf{x}^\mathsf{T} \boldsymbol{\beta}_{ijx}\right), \qquad i \neq j,$$

is of parametric form to describe the multiplicative effect of covariates on the transition intensity, where $q_{ij0}(t)$ is the baseline intensity from state i to j at time t and $\boldsymbol{\beta}_{ijx} = (\beta_{ij1}, \beta_{ij2}, \ldots, \beta_{ijp})^\mathsf{T}$ is a $p \times 1$ vector of regression coefficients.

For time-homogenous Markov models, baseline intensities can be specified as time-independent constants, i.e. $q_{ij0}(t) = \exp(\beta_{ij0})$, and then the proportional intensity model becomes the log-linear model. One advantage of this model is to ensure that transition intensities are nonnegative for all $\mathbf{x}$ and $\boldsymbol{\beta}_{ijx}$'s; other parameterizations may be more appropriate in particular situations, such as the local equilibrium distribution model (Kosorok and Chao, 1996). For Markov models with piecewise constant transition intensities, baseline intensities can be specified to be piecewise constant of the form

$$q_{ij0}(t) = \exp(\beta_{ijk0}), \qquad b_k \leq t < b_{k+1}, \quad k = 0, 1, \ldots, M,$$

where $0 = b_0 < b_1 < \cdots < b_M < b_{M+1} = \infty$ is a pre-specified sequence of times.

It is important to note that covariates can be time-varying. In this situation, time-dependent covariates are observed at the same time points as the process but the values of covariates between two observations are not known, except deterministic time-dependent covariates, such as age. Therefore, further assumptions and approximations are often made to allow the calculation of

the transition probability matrix and the likelihood. A widely-used assumption is that transition intensities are piecewise constant. It is usually achieved by approximating the time-dependent covariate as a step function which remains constant between its observation times, such that

$$z(t) = z_i, \qquad t_i \leq t < t_{i+1},$$

where $z(t)$ represents the value of the time-dependent covariate at time t and z_i is the value of $z(t)$ observed at time t_i. Then, transition intensities can be written as

$$q_{ij}(t) = q_{ij0}(t) \exp\left[\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}_{ijx} + z(t)\,\beta_{ijz}\right], \qquad i \neq j,$$

where $q_{ij0}(t)$ is assumed to be either constant or piecewise constant.

1.4 Composite likelihood method

The composite likelihood, termed by Lindsay (1988), is a type of pseudo likelihoods constructed by multiplying a collection of marginal or conditional distributions for subsets of response components. The idea dates back to the pseudo likelihood proposed by Besag (1974, 1975) for making statistical inference in spatial random fields and the partial likelihood of Cox (1975) for dimension reduction in the nuisance parameter space. It has drawn increasing attentions in estimation and inference for data with complex structures. Application areas include statistical genetics, spatial statistics, time series, longitudinal studies, and panel surveys. More comprehensive reviews can be found in Varin (2008), Varin *et al.* (2011), Larribe and Fearnhead (2011), and Lindsay *et al.* (2011).

1.4.1 Formulation

Consider an m-dimensional random vector $\mathbf{Y}$, with probability density function $f(\mathbf{y};\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a p-dimensional parameter vector of interest. Let $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ denote a set of marginal or conditional events with associate sub-likelihoods $\mathcal{L}_k(\boldsymbol{\theta};\mathbf{y}) \propto f(\mathbf{y} \in \mathcal{A}_k;\boldsymbol{\theta})$. The composite likelihood is the weighted product

$$\mathcal{L}_{\mathrm{c}}(\boldsymbol{\theta};\mathbf{y}) = \prod_{k=1}^{K} \left[\mathcal{L}_k(\boldsymbol{\theta};\mathbf{y})\right]^{w_k},$$

where w_k are weights assigned to each sub-likelihood. If all the weights are equal, then they can be ignored; the unequal weights are chosen to improve efficiency of estimation.

Although the combination of marginal and conditional densities are allowed in the formulation, composite likelihoods are typically distinguished to be conditional or marginal versions.

Composite conditional likelihoods

The composite conditional likelihoods date back to the pseudo likelihood proposed by Besag (1974, 1975) for inference in spatial models. This pseudo likelihood is constructed from the product of conditional densities of a single observation given its neighbours,

$$\mathcal{L}_{\mathrm{c}} = \prod_{r=1}^{m} f(y_r \mid y_s \in \partial y_r;\boldsymbol{\theta}),$$

where ∂y_r represent the set of neighbours of y_r. It is further generalized by using blocks of observations for both conditional and conditioned events in spatial data analysis (Vecchia, 1988; Stein *et al.*, 2004).

For stratified case-control studies, Liang (1987) suggested composite conditional likelihoods

based on the product of conditional densities of a single observation in the case group given the pairwise sum of that observation and an observation in the control group within the same stratum. The further extensions include Hanfelt (2004) and Wang and Williamson (2005) for sparse clustered binary data and Fujii and Yanagimoto (2005) for the exponential dispersion model with multiple strata.

For repeated multivariate binary data, Molenberghs and Verbeke (2005) explored two types of composite conditional likelihoods: one constructed from the product of univariate conditional densities within one unit given all the other outcomes in the same unit, and the other constructed from the product of conditional densities of all the outcomes for one occasion, given the outcomes for the other occasions. Mardia *et al.* (2008, Chapter 12) considered composite conditional likelihoods based on the product of univariate conditional densities given all the other observations in the same dimension for the trivariate von Mises distribution with application to protein fold data.

Composite marginal likelihoods

The simplest composite marginal likelihood is the independence likelihood suggested by Chandler and Bate (2007)

$$\mathcal{L}_{\text{ind}}\left(\boldsymbol{\theta};\mathbf{y}\right)=\prod_{r=1}^{m}f\left(y_r;\boldsymbol{\theta}\right),$$

based on the working independence assumption for the analysis of the clustered data. However, the within-cluster dependence is ignored and thus the inference is limited to marginal parameters in the independence likelihood.

The most popular type of composite marginal likelihoods is the pairwise likelihood

$$\mathcal{L}_{\text{pair}}(\boldsymbol{\theta}; \mathbf{y}) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r, y_s; \boldsymbol{\theta}),$$

which models the second-order dependence explicitly without specifying the full joint distribution and takes the correlation among observations into account. There has been increasing influence of pairwise likelihood methods on spatial statistics since 1990, e.g., Hjort and Omre (1994), Heagerty and Lele (1998), Varin *et al.* (2005), Guan (2006), Li and Lin (2006), and Bai *et al.* (2012). Pairwise likelihood methods have also been applied to correlated data, including random set models in image analysis (Nott and Rydén, 1999), correlated binary data (Kuk and Nott, 2000), additive and multiplicative frailty models for multivariate survival data analysis (Parner, 2001), correlated gamma frailty models for longitudinal count data (Henderson and Shimakura, 2003), and multilevel models with binary responses and probit link (Renard *et al.*, 2004). Their extensions include the construction from larger subsets of observations (Varin and Vidoni, 2005; Caragea and Smith, 2007) and the combination of the independence likelihood and the pairwise likelihood in some optimal way (Cox and Reid, 2004).

In addition to lower-dimensional marginal densities, composite marginal likelihoods can be constructed based on the function of the lower-dimensional marginal densities, such as pairwise difference,

$$\mathcal{L}_{\text{diff}}(\boldsymbol{\theta}; \mathbf{y}) = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} f(y_r - y_s; \boldsymbol{\theta}).$$

This form was proposed by Curriero and Lele (1999) to semivariogram estimation in geostatistics and further extended to estimation of covariance components by Lele and Taper (2002).

1.4.2 Asymptotic theory

The composite likelihood can be viewed as a type of inference functions obtained from multiplying a collection of marginal or conditional densities. Therefore, the derivative of the composite log-likelihood is an unbiased estimating function.

Suppose n independent and identically distributed observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are obtained from model $f(\mathbf{y}; \boldsymbol{\theta})$. Large-sample properties of the composite likelihood can be derived under two scenarios: one is the increment in the number of independent observations with fixed observation times, i.e., $n \to \infty$ with m fixed; the other is the increment in the observation times of few realizations of the process, i.e., $m \to \infty$ with a small integer n. In disease progression studies, we focus on the first asymptotic scenario in which the number of individuals is increasing while the number of observations for each individual is fixed.

The consistency and asymptotic normality of the composite maximum likelihood estimator hold under regularity conditions for the first scenario (e.g., Lindsay, 1988; Molenberghs and Verbeke, 2005, Chapter 9): as $n \to \infty$,

1. The composite maximum likelihood estimator $\tilde{\boldsymbol{\theta}}$, which is the maximizer of the composite log-likelihood, converges in probability to $\boldsymbol{\theta}$;

2. $\sqrt{n}\,(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to $\mathbf{N}_p\left[\mathbf{0}, \mathbf{G}^{-1}(\boldsymbol{\theta})\right]$, where $\mathbf{N}_p(\boldsymbol{\mu}, \Sigma)$ is the p-dimensional normal distribution with mean $\boldsymbol{\mu}$ and variance Σ, and $\mathbf{G}(\boldsymbol{\theta})$ is the Godambe information matrix for a single observation (Godambe, 1960), given by

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta})\,\mathbf{J}^{-1}(\boldsymbol{\theta})\,\mathbf{H}(\boldsymbol{\theta}),$$

with the sensitivity matrix

$$\mathbf{H}(\boldsymbol{\theta}) = E\left[\frac{\partial^2 \log \mathcal{L}_{\mathrm{c}}(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}\right],$$

and the variability matrix

$$\mathbf{J}(\boldsymbol{\theta}) = \mathrm{Var}\left[\frac{\partial \log \mathcal{L}_{\mathrm{c}}(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}}\right].$$

The validity of the composite likelihood method relies on the model assumption of lower dimensional marginal or conditional densities, but not on the correct specification of the full joint distribution. Therefore, the composite likelihood method is robust to the possible misspecification of higher dimensional distributions. The investigation on robustness was carried out for several scenarios, including the misspecification of the random effects distribution in a one-way random effects model (Lele and Taper, 2002), the misspecification of the correlation structure in sparse clustered binary data (Wang and Williamson, 2005), and the misspecification of the missing data mechanism (Parzen *et al.*, 2007; Yi *et al.*, 2011b).

Another feature of the composite likelihood method pertains to computational efficiency. It reduces computational burden by avoiding the high dimensional integration in many situations, such as random effects models (Varin *et al.*, 2005), analysis of clustered data (Li and Yi, 2013a,b; Varin and Vidoni, 2008), and especially the cases with high dimensional covariance matrices. In addition, the composite likelihood surface can be smoother than the full likelihood surface (Blackwell, 1985; Liang and Yu, 2003). It makes composite likelihood more robust to converge and therefore is called "computational robustness" by Renard *et al.* (2004).

1.4.3 Composite likelihoods and Markov chain models

Hjort and Varin (2008) discussed and compared the full likelihood, composite marginal likelihood, and composite conditional likelihood methods of estimation for discrete-time Markov chain models in the context of spatial statistics.

Let $S_0, \ldots, S_m$ be an irreducible discrete-time Markov chain on a finite stat space with stationary transition probabilities. The traditional full likelihood is a product of all the transition probabilities because of the Markov assumption with the distribution of the initial state ignored:

$$\mathcal{L}(\boldsymbol{\theta}) \propto \prod_{i=1}^{m} \Pr(S_i = s_i \mid S_{i-1} = s_{i-1}; \boldsymbol{\theta}),$$

where s_i's are observations, and $\boldsymbol{\theta}$ is the $p \times 1$ parameter vector for transition probabilities.

The composite conditional likelihood is formulated as a product over conditional densities of a single observation given the rest of data:

$$\mathcal{L}_{\text{cc}} = \prod_{i=1}^{m-1} \Pr(S_i = s_i \mid S_0 = s_0, \ldots, S_{i-1} = s_{i-1}, S_{i+1} = s_{i+1}, \ldots, S_m = s_m; \boldsymbol{\theta}). \tag{1.2}$$

By the Markov property, the conditional distribution given all the rest observations is the same as the conditional distribution given the nearest neighbours. Therefore, (1.2) can be simplified into

$$\begin{aligned} \mathcal{L}_{\text{cc}} &= \prod_{i=1}^{m-1} \Pr(S_i = s_i \mid S_{i-1} = s_{i-1}, S_{i+1} = s_{i+1}; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{m-1} \frac{\Pr(S_{i+1} = s_{i+1} \mid S_i = s_i; \boldsymbol{\theta}) \Pr(S_i = s_i \mid S_{i-1} = s_{i-1}; \boldsymbol{\theta})}{\Pr(S_{i+1} = s_{i+1} \mid S_{i-1} = s_{i-1}; \boldsymbol{\theta})}. \end{aligned}$$

The composite marginal likelihood is a product of all the bivariate distributions of adjacent

observations:

$$\begin{aligned}\mathcal{L}_{\text{cm}} &= \prod_{i=1}^{m} \Pr\left(S_{i-1} = s_{i-1}, S_i = s_i; \boldsymbol{\theta}\right) \\ &= \prod_{i=1}^{m} \Pr\left(S_i = s_i \mid S_{i-1} = s_{i-1}; \boldsymbol{\theta}\right) \Pr\left(S_{i-1} = s_{i-1}; \boldsymbol{\theta}\right),\end{aligned}$$

where the term $\Pr\left(S_{i-1} = s_{i-1}\right)$ is the equilibrium distribution with the assumption that the chain starts out in its equilibrium distribution.

Hjort and Varin (2008) showed that the composite marginal and conditional likelihoods can be interpreted as penalized likelihoods in Markov chain models. Both theoretical and numerical analysis provided strong evidence that the composite marginal likelihood method is preferable to the composite conditional likelihood method in terms of efficiency and robustness and is a robust alternative to the full likelihood method.

1.5 Outline of the thesis

The structure of the remaining thesis is as follows. In Chapter 2, the progressive multi-state model with misclassification is developed to simultaneously estimate transition rates and account for potential misclassification. The performance of the maximum likelihood and pairwise likelihood estimators is evaluated by simulation studies. The proposed progressive model is illustrated on coronary allograft vasculopathy data, in which the diagnosis based on the coronary angiography is subject to error.

In Chapter 3, hidden mover-stayer models are proposed to provide a solution to a type of heterogeneity where the population consists of both movers and stayers in the presence of misclassification. The likelihood inference procedure based on the EM algorithm is developed for the

proposed model. The performance of the likelihood method is investigated through simulation studies. The proposed method is applied to the Waterloo Smoking Prevention Project.

In Chapter 4, we propose estimation procedures for Markov models with binary covariates subject to misclassification. We show that the model is not identifiable under covariate misclassification. Consequently, we develop likelihood inference methods based on known reclassification probabilities and the main/validation study design. Simulation studies are conducted to investigate the performance of proposed methods and the consequence of the naive analysis which ignores the misclassification.

In Chapter 5, we consider two-state Markov models where time-dependent surrogate covariates are available. We exploit both functional and structural inference methods to reduce or remove bias effects induced from covariate measurement error. The performance of proposed methods is investigated based on simulation studies.

Chapter 2

Analysis of Progressive Multi-State Models with Misclassified States

2.1 Introduction

Unidirectional progressive Markov models can be very powerful to model the processes of successive events to reflect an accumulation of damage or deterioration. Satten (1999) considered a conditionally time-homogeneous progressive Markov model for panel data, given random effects which act on each conditional intensity multiplicatively. Cook *et al.* (2004) described a conditionally time-homogeneous progressive Markov model with discrete multivariate random effects for clustered multi-state processes. Sutradhar and Cook (2008) generalized the model of Cook *et al.* (2004) to allow continuous multivariate random effects and time non-homogeneity for clustered progressive processes. Chen *et al.* (2010) proposed a progressive Markov model with piecewise constant transition intensities to address non-homogeneity under informative examination times.

When analyzing disease progression data, one serious challenge is that the disease state may

be misclassified. Misclassification of the disease state is frequently caused by sampling error, due to the poor quality of a diagnostic test or the impossibility of the accurate assessment as well as from reading error. To account for potential misclassification, hidden Markov models (HMMs) are commonly employed. For instance, Nagelkerke *et al.* (1990) considered a two-state Markov model with only one type of misclassification and constant transition rates. Bureau *et al.* (2003) presented a continuous-time hidden Markov model with two types of misclassification and covariate dependent transition rates. Rosychuk and Thompson (2003, 2004) investigated identifiably issues and bias correction of parameter estimates for the two-state hidden Markov process. Rosychuk *et al.* (2006) compared three variance estimation approaches for the two-state hidden Markov process, and Rosychuk and Islam (2009) developed a Bayesian approach for inference. Jackson *et al.* (2003) presented a general HMM in continuous time which allows covariate-dependent transitions and misclassification rates. Recently, Jackson (2011) developed the *msm* package in `R` for fitting continuous-time Markov and HMMs to panel data.

These methods are essentially likelihood based, and thereby are vulnerable to model misspecification. Furthermore, computation based on direct maximization of the likelihood can be intensive when the number of states or the number of observations is large. In this chapter, we propose an inference method to handle progressive models with misclassified states using the pairwise likelihood formulation (Lindsay, 1988; Cox and Reid, 2004; Lindsay *et al.*, 2011). This method enjoys the robustness property where the dependence assumption of transition among states can be relaxed compared to usual HMMs. In addition, we develop an EM algorithm, in which the derivatives of the expected complete data log-likelihood are in the closed form, to obtain maximum likelihood estimates under the progressive Markov model with misclassification.

The pioneering work of the EM algorithm in HMMs, known as Baum-Welch algorithm, includes Baum and Petrie (1966), Baum and Eagon (1967), and Baum *et al.* (1970). The pairwise

EM algorithm was first proposed by Liang and Yu (2003) for network tomography, and Castro *et al.* (2004) presented the pairwise EM algorithm in their review paper for recent developments of network tomography. Gao and Song (2011) established properties of the composite likelihood EM algorithm. The validity of those methods lies on the assumption that data are accurately measured. This assumption, however, is commonly violated in application. Our development here complements available work in that progressive models may contain error-prone states, and broadens the application scope.

The rest of this chapter is organized as follows. In Section 2.2, we describe the progressive Markov model with misclassification and the conditions for the non-informative observation process. The inference procedures based on the likelihood and pairwise likelihood approaches are developed in Sections 2.3 and 2.4, respectively. The performance of the proposed methods is investigated and compared through simulation studies in Section 2.5. The proposed methods are applied to the coronary allograft vasculopathy data in Section 2.6. Discussion is given in Section 2.7. Technical details are presented in Section 2.8.

2.2 Progressive model with misclassification

2.2.1 K-state progressive Markov model

Suppose an individual moves among K states, denoted by integers $1, 2, \ldots, K$. Let $S(t)$ denote the true state at time t occupied by an individual. Assume that $\{S(t), t \geq 0\}$ follows a continuous-time progressive Markov process. Let $\mathbf{P}(s, s+t)$ be the $K \times K$ transition probability matrix with (i, j) entry $P_{ij}(s, s+t) = \Pr\{S(s+t) = j \mid S(s) = i\}$ for $s \geq 0$, $t > 0$, $i, j = 1, 2, \ldots, K$,

where $P_{ij}(s, s+t) = 0$ if $i > j$. The transition intensity from state i to j at time t is

$$q_{ij}(t) = \lim_{\Delta t \downarrow 0} \frac{P_{ij}(t, t+\Delta t)}{\Delta t}, \qquad i \neq j,$$

and as a convention, define $q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t) = -\sum_{j>i} q_{ij}(t)$. Let $\mathbf{Q}(t)$ be the $K \times K$ transition intensity matrix with (i, j) entry $q_{ij}(t)$, $i, j = 1, 2, \ldots, K$.

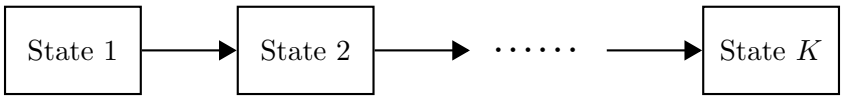

Figure 2.1: K-state unidirectional progressive model

With an unidirectional progressive model, shown in Figure 2.1, each individual passes through the states consecutively and does not escape from state K. In addition, the process is assumed to be irreversible and the transition only happens from one state to its consecutive state. That is, the individual can only go to state $i+1$ after passing state i. In this case, for a given $i = 1, \ldots, K-1$, $q_{i,i+1}(t) > 0$, $q_{ii}(t) = -q_{i,i+1}(t)$ and $q_{ij}(t) = 0$, $j \neq i, i+1$, and state K is an absorbing state with $q_{Kj}(t) = 0$, $j = 1, \ldots, K$. For ease of notation, let $q_i(t)$ denote $q_{i,i+1}(t)$, $i = 1, \ldots, K$, then the transition intensity matrix takes the form

$$\mathbf{Q}(t) = \begin{pmatrix} -q_1(t) & q_1(t) & 0 & \cdots & 0 & 0 \\ 0 & -q_2(t) & q_2(t) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -q_{K-1}(t) & q_{K-1}(t) \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

This chapter is primarily concerned with time-homogeneous progressive Markov models in which transition intensities are independent of t. We therefore let $q_i(t) = q_i$, $i = 1, \ldots, K$ and

write $\mathbf{Q}(t) = \mathbf{Q}$. It follows that $\mathbf{P}(s, s+t) = \mathbf{P}(0, t)$, which is then written as $\mathbf{P}(t)$. Transition probability from state i to j can be analytically expressed in terms of transition intensities (Satten, 1999):

$$P_{ij}(t) = \begin{cases} \sum_{k=i}^{j} C_{ijk} \exp(-q_k t) & i \leq j, \\ 0 & i > j, \end{cases}$$

where

$$C_{ijk} = \frac{\prod_{l=i}^{j-1} q_l}{\prod_{l=i,\, l\neq k}^{j} (q_l - q_k)}, \qquad i \leq k \leq j,$$

$C_{kkk} = 1$, $i, j, k = 1, \ldots, K$, and $q_K = 0$.

In application, transitions between states are associated with certain covariates, and interest lies in understanding the relationship between these covariates and transition intensities. Let $\mathbf{x}$ be a $p \times 1$ vector of prognostic variables. Consider regression models

$$q_i(\mathbf{x}) = q_{i0} \exp\left(\mathbf{x}^{\mathsf{T}} \boldsymbol{\beta}_{ix}\right), \qquad i = 1, \ldots, K-1, \tag{2.1}$$

where q_{i0} is the baseline transition intensity out of state i, and $\boldsymbol{\beta}_{ix} = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ip})^{\mathsf{T}}$ are vectors of regression coefficients which are of primary interest. Often, q_{i0} are reparameterized as $q_{i0} = \exp(\beta_{i0})$ (e.g. Kalbfleisch and Lawless, 1985; Jackson *et al.*, 2003).

2.2.2 Misclassification model

It is common that the disease state is subject to misclassification. Let $S^*(t)$ represent the observed state occupied at time t for $t \geq 0$. Suppose some supplementary information, such as clinical symptoms or measurements is associated with state misclassification. Let $\mathbf{c}(t)$ denote the predictor vector associated with the misclassification process. For $i \neq j$, $i, j = 1, \ldots, K$,

let $\pi_{ij}(t) = \Pr\{S^*(t) = j \mid S(t) = i, \mathbf{c}(t)\}$ denote misclassification probabilities at time t. By the constraint $\sum_{j=1}^{K} \Pr\{S^*(t) = j \mid S(t) = i, \mathbf{c}(t)\} = 1$, we define $\pi_{ii}(t) = 1 - \sum_{j \neq i}^{K} \pi_{ij}(t)$, $i = 1, \ldots, K$.

We employ the multinomial logistic regression model to portray the relationship between misclassification probabilities and $\mathbf{c}(t)$ (e.g. Agresti, 2002, Chapter 7):

$$\log\left\{\frac{\pi_{ij}(t)}{\pi_{ii}(t)}\right\} = \alpha_{ij0} + \boldsymbol{\alpha}_{ijc}^{\mathsf{T}}\mathbf{c}(t), \qquad i \neq j, \tag{2.2}$$

where α_{ij0} and $\boldsymbol{\alpha}_{ijc}$ are state-dependent regression coefficients.

In application, suitable constraints may be imposed on misclassification probabilities to reflect a prior knowledge of the diagnosis process. For progressive multi-state models, probabilities of misclassification may be negligibly small for those states that are far apart. For example, Jackson *et al.* (2003) considered a scenario where misclassification probabilities are non-zero constants only for some adjacent states.

2.2.3 Non-informative observation process

Let M denote the number of observations of an individual, which is a random variable, and m be the realization of M. Suppose the disease process for an individual $\{S(t), t \geq 0\}$ is assessed at a finite number of times $0 \leq t_1 < t_2 < \cdots < t_m$, which may be subject to misclassification. Let $\{S^*(t), t \geq 0\}$ denote the observed process and s_i^* denote the observed state at time t_i, $i = 1, \ldots, m$. In addition to the observed states $\{S_1^*, \ldots, S_m^*\}$, the number of observations, M, and the observation times $T_1, T_2, \ldots, T_m$ are also random variables. Inferences are, in principle, carried out based on the joint distribution of all the observed random variables, $\Pr(\mathbf{S}_m^* = \mathbf{s}_m^*; \mathbf{T}_m = \mathbf{t}_m; M = m)$, where $\mathbf{S}_m^* = \{S^*(t_1), \ldots, S^*(t_m)\}$, $\mathbf{T}_m = \{T_1, T_2, \ldots, T_m\}$,

$\mathbf{s}_m^* = \{s_1^*, \ldots, s_m^*\}$, and $\mathbf{t}_m = \{t_1, \ldots, t_m\}$. Our aim is to make inference on the parameters associated with the true disease process in the presence of possible state misclassification. To this end, a convenient factorization is invoked where the underlying true states are explicitly spelled out:

$$\begin{aligned}
& \Pr\left(\mathbf{S}_m^* = \mathbf{s}_m^*; \mathbf{T}_m = \mathbf{t}_m; M = m\right) \\
= \; & \Pr\left(M = m \mid \mathbf{S}_m^* = \mathbf{s}_m^*; \mathbf{T}_m = \mathbf{t}_m\right) \Pr\left(\mathbf{S}_m^* = \mathbf{s}_m^*; \mathbf{T}_m = \mathbf{t}_m\right) \\
= \; & \Pr\left(\mathbf{S}_m^* = \mathbf{s}_m^*; \mathbf{T}_m = \mathbf{t}_m\right) \\
= \; & \sum_{\mathbf{s}_m} \Pr\left(\mathbf{S}_m^* = \mathbf{s}_m^*; \mathbf{S}_m = \mathbf{s}_m; \mathbf{T}_m = \mathbf{t}_m\right) \\
= \; & \sum_{\mathbf{s}_m} \Big\{ \Pr\left(\mathbf{S}_m^* = \mathbf{s}_m^* \mid \mathbf{S}_m = \mathbf{s}_m; \mathbf{T}_m = \mathbf{t}_m\right) \Pr\left(\mathbf{S}_m = \mathbf{s}_m; \mathbf{T}_m = \mathbf{t}_m\right) \Big\}, \qquad (2.3)
\end{aligned}$$

where $\mathbf{S}_m = \{S(t_1), \ldots, S(t_m)\}$, and $\mathbf{s}_m = \{s_1, \ldots, s_m\}$. The second term inside the summation of (2.3) is of primary interest, and this quantity is examined by Grüger *et al.* (1991) using the factorization:

$$\begin{aligned}
& \Pr\left\{S(t_1) = s_1, \ldots, S(t_m) = s_m; T_1 = t_1, \ldots, T_m = t_m\right\} \\
= \; & \Pr\left(H_1\right) \prod_{j=2}^{m} \Big[\Pr\left\{S(t_j) = s_j \mid T_j = t_j, H_{j-1}\right\} \Big] \prod_{j=2}^{m} \Big\{ \Pr\left(T_j = t_j \mid H_{j-1}\right) \Big\},
\end{aligned}$$

where H_j is the history of true disease states and observation times up to and including the jth time point, defined as

$$H_j = \left\{T_1 = t_1, S(t_1) = s_1, \ldots, T_j = t_j, S(t_j) = s_j\right\}, \qquad j = 1, \ldots, m.$$

In the context of no misclassification, Grüger *et al.* (1991) introduced the following conditions for conducting inferences in order to avoid modelling the observation scheme:

1. The probability of staying in state s_j at time t_j, given the history $H_{j-1} = \{T_1 = t_1, S(t_1) = s_1, \ldots, T_{j-1} = t_{j-1}, S(t_{j-1}) = s_{j-1}\}$, is independent of whether an observation is carried out at this time and the past observation times, i.e.,

$$\Pr\{S(t_j) = s_j \mid T_j = t_j, H_{j-1}\} = \Pr\{S(t_j) = s_j \mid S(t_1) = s_1, \ldots S(t_{j-1}) = s_{j-1}\}; \quad (2.4)$$

2. The conditional distribution of the jth observation time T_j, $\Pr(T_j = t_j \mid H_{j-1})$, is functionally independent of parameters governing transition intensities of the disease process $\{S(t), t \geq 0\}$.

These conditions are useful to confine attention to studying the true state process. However, they are not sufficient when states are subject to misclassification. Additional care should be taken of the misclassification process. Note that the first term inside the summation of (2.3) can be factored as follows

$$\begin{aligned} & \Pr(\mathbf{S}_m^* = \mathbf{s}_m^* \mid \mathbf{S}_m = \mathbf{s}_m; \mathbf{T}_m = \mathbf{t}_m) \\ = \; & \Pr\{S^*(t_1) = s_1^* \mid H_m\} \prod_{j=2}^{m} \Pr\left\{S^*(t_j) = s_j^* \;\middle|\; H_{j-1}^{s^*}, H_m\right\}, \end{aligned}$$

where $H_j^{s^*}$ is the history of the observed states up to the jth observation, defined as $H_j^{s^*} = \left\{S^*(t_1) = s_1^*, \ldots, S^*(t_j) = s_j^*\right\}$, $j = 1, \ldots, m$. The additional condition for the non-informative observation process in the presence of state misclassification is that the conditional probability of the jth observed state, given the history of observed states, true states, and observation times, is independent of all the observation times, i.e.

$$\Pr\left\{S^*(t_j) = s_j^* \;\middle|\; H_{j-1}^{s^*}, H_m\right\} = \Pr\left\{S^*(t_j) = s_j^* \;\middle|\; H_{j-1}^{s^*}, S(t_1) = s_1, \ldots, S(t_m) = s_m\right\}. \quad (2.5)$$

This condition is as important as (2.4), since both conditions ensure that what we can estimate from the data, $\Pr\{S(t_j) = s_j \mid T_j = t_j, H_{j-1}\} \Pr\left\{S^*(t_j) = s_j^* \;\middle|\; H_{j-1}^{s^*}, H_m\right\}$, is identical to what we are interested in, $\Pr\{S(t_j) = s_j \mid \mathbf{S}_{j-1} = \mathbf{s}_{j-1}\} \Pr\left\{S^*(t_j) = s_j^* \;\middle|\; H_{j-1}^{s^*}, \mathbf{S}_m = \mathbf{s}_m\right\}$. Furthermore, in the HMM, the Markov property assumes that

$$\Pr\{S(t_j) = s_j \mid \mathbf{S}_{j-1} = \mathbf{s}_{j-1}\} = \Pr\{S(t_j) = s_j \mid S(t_{j-1}) = s_{j-1}\}, \qquad j = 2, \ldots, m,$$

and the output independence assumption in HMM suggests that

$$\Pr\left\{S^*(t_j) = s_j^* \;\middle|\; H_{j-1}^{s^*}, \mathbf{S}_m = \mathbf{s}_m\right\} = \Pr\left\{S^*(t_j) = s_j^* \mid S(t_j) = s_j\right\}, \qquad j = 1, \ldots, m. \quad (2.6)$$

2.3 Maximum likelihood estimation via the EM algorithm

Suppose N individuals are under study and each individual independently follows an unidirectional progression time-homogeneous Markov process. Let $\{S_\ell(t), t \geq 0\}$ and $\{S_\ell^*(t), t \geq 0\}$ denote the true and observed process for individual ℓ, respectively, $\ell = 1, 2, \ldots, N$. Let $\mathbf{x}_\ell$ be time-independent prognostic variables, and $\mathbf{c}_\ell(t)$ represent misclassification predictors for individual ℓ at time $t \geq 0$, $\ell = 1, 2, \ldots, N$. Let $t_{\ell 1}, \ldots, t_{\ell m_\ell}$ denote m_ℓ times at which individual ℓ is observed. Let $H_{\ell r}^{s^*} = \{S_\ell^*(t_{\ell k}) = s_\ell^*(t_{\ell k}) : 1 \leq k < r\}$ and $H_{\ell r}^{c} = \{\mathbf{c}_\ell(t_{\ell k}) : 1 \leq k < r\}$ denote the history of the observed states and the predictor history at time $t_{\ell r}$, respectively. For convenience, we write $S_\ell^*(t_{\ell r})$, $S_\ell(t_{\ell r})$, and $\mathbf{c}_\ell(t_{\ell r})$ as $S_{\ell r}^*$, $S_{\ell r}$, and $\mathbf{c}_{\ell r}$, respectively, where $r = 1, \ldots, m_\ell$. We employ models (2.1) and (2.2) to postulate the response and misclassification processes, respectively. Let $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}_1^\mathsf{T}, \ldots, \boldsymbol{\alpha}_K^\mathsf{T}, \boldsymbol{\beta}^\mathsf{T}\right)^\mathsf{T}$, where $\boldsymbol{\alpha}_i = \left(\alpha_{ij0}, \boldsymbol{\alpha}_{ijc}^\mathsf{T} : j = 1, \ldots, K, j \neq i\right)^\mathsf{T}$, and $\boldsymbol{\beta} = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp} : k = 1, \ldots, K-1)^\mathsf{T}$.

To conduct estimation of the model parameters, we employ the EM algorithm in which the

underlying true states are treated as missing data. We now elaborate on the EM algorithm for the progressive model with misclassified states and defer technical details to Section 2.8.

The log-likelihood for the complete data contributed from individual ℓ is

$$\log \mathcal{L}_\ell^{\mathrm{c}}(\boldsymbol{\theta}) = \sum_{r=1}^{m_\ell} \left[\log \left\{ \Pr\left(S_{\ell r}^* \mid S_{\ell r}, \mathbf{c}_{\ell r}; \boldsymbol{\alpha}_{s_{\ell r}}\right) \right\} \right] + \sum_{s=1}^{m_\ell - 1} \left[\log \left\{ \Pr\left(S_{\ell,s+1} \mid S_{\ell s}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \right\} \right].$$

In the expectation step (E-step), the expected complete data log-likelihood at the $(k+1)$th iteration is $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \sum_{\ell=1}^{N} Q_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$, where $\boldsymbol{\theta}^{(k)}$ is the estimate of $\boldsymbol{\theta}$ at the kth iteration, and $Q_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ is given by

$$\begin{aligned} Q_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= E\left\{ \log \mathcal{L}_\ell^{\mathrm{c}}(\boldsymbol{\theta}) \;\middle|\; H_{\ell,m_\ell+1}^{s^*}, H_{\ell,m_\ell+1}^{c}, \mathbf{x}_\ell; \boldsymbol{\theta}^{(k)} \right\} \\ &= \sum_{i=1}^{K} \sum_{r=1}^{m_\ell} \gamma_{\ell r}(i) \log \left\{ \Pr\left(S_{\ell r}^* \mid S_{\ell r} = i, \mathbf{c}_{\ell r}; \boldsymbol{\alpha}_i\right) \right\} \\ &\quad + \sum_{s=1}^{m_\ell - 1} \sum_{i=1}^{K} \sum_{j=i}^{K} \xi_{\ell s}(i,j) \log \left\{ \Pr\left(S_{\ell,s+1} = j \mid S_{\ell s} = i, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \right\} \end{aligned}$$

with conditional probabilities

$$\begin{aligned} \gamma_{\ell r}(i) &= \Pr\left\{ S_{\ell r} = i \;\middle|\; H_{\ell,m_\ell+1}^{s^*}, H_{\ell,m_\ell+1}^{c}, \mathbf{x}_\ell; \boldsymbol{\theta}^{(k)} \right\}, \\ \text{and} \qquad \xi_{\ell s}(i,j) &= \Pr\left\{ S_{\ell s} = i, S_{\ell,s+1} = j \;\middle|\; H_{\ell,m_\ell+1}^{s^*}, H_{\ell,m_\ell+1}^{c}, \mathbf{x}_\ell; \boldsymbol{\theta}^{(k)} \right\}, \end{aligned}$$

which can be computed by the forward-backward algorithm (Baum *et al.*, 1970; Rabiner, 1989).

As parameters for the progression model are distinct from those for the misclassification model, we can carry out maximization with respect to $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K$ and $\boldsymbol{\beta}$ separately in the M-step. The maximizer of the expected complete data log-likelihood does not exist in a closed form, and therefore, the Newton-Raphson procedure may be used to iteratively compute $\boldsymbol{\theta}^{(k)}$ in the

M-step. The estimator, say $\hat{\boldsymbol{\theta}}$, of the parameters $\boldsymbol{\theta}$, is obtained through iterations between E and M steps until convergence of $\boldsymbol{\theta}^{(k)}$.

Let $\mathcal{L}^{\mathrm{c}}(\boldsymbol{\theta}; \mathbf{S}, \mathbf{S}^*)$ be the complete likelihood based on the underlying true state $\mathbf{S}$ and the observed state $\mathbf{S}^*$ of one individual. Note that

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{S}^*)}{\partial \boldsymbol{\theta}} = E\left\{ \frac{\partial \log \mathcal{L}^{\mathrm{c}}(\boldsymbol{\theta}; \mathbf{S}, \mathbf{S}^*)}{\partial \boldsymbol{\theta}} \,\middle|\, \mathbf{S}^* \right\}.$$

That is, the gradient of the log-likelihood can be approximated by the expectation of the gradient of the complete data log-likelihood, and thus the Hessian of the log-likelihood is obtained by the numerically differentiation (Jamshidian and Jennrich, 2000).

To protect against possibly invalid assumptions of the independence structures among the states, the sandwich-type robust variance estimation (White, 1982) is used. Specifically, $\mathcal{L}(\boldsymbol{\theta})$ is constructed from a misspecified model, then $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$ almost surely, where $\boldsymbol{\theta}^*$ is the root of the expectation of $\partial \log \mathcal{L}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ taken with respect to the true distribution. The asymptotic normality of $\hat{\boldsymbol{\theta}}$ from a misspecified model is $\sqrt{N}\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathbf{N}\{\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}^*)\}$, where $\mathbf{C}(\boldsymbol{\theta}) = \mathbf{A}^{-1}(\boldsymbol{\theta})\, \mathbf{B}(\boldsymbol{\theta})\, \mathbf{A}^{-1}(\boldsymbol{\theta})$ with

$$\mathbf{A}(\boldsymbol{\theta}) = E\left\{ \partial^2 \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{S}^*) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}} \right\} \quad \text{and} \quad \mathbf{B}(\boldsymbol{\theta}) = E\left[\{ \partial \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{S}^*) / \partial \boldsymbol{\theta} \}^{\otimes 2} \right].$$

The matrix $\mathbf{C}(\boldsymbol{\theta}^*)$ can be estimated by $\widehat{\mathbf{A}}^{-1}(\boldsymbol{\theta})\, \widehat{\mathbf{B}}(\boldsymbol{\theta})\, \widehat{\mathbf{A}}^{-1}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, where

$$\begin{aligned} \widehat{\mathbf{B}}(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{\ell=1}^{N} \left\{ \frac{\partial \log \mathcal{L}_\ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{\otimes 2} = \frac{1}{N} \sum_{\ell=1}^{N} \left\{ \frac{\partial Q_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}')}{\partial \boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}'=\boldsymbol{\theta}} \right\}^{\otimes 2}, \\ \text{and} \qquad \widehat{\mathbf{A}}(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{\ell=1}^{N} \frac{\partial^2 \log \mathcal{L}_\ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}. \end{aligned}$$

Let $\mathbf{s}(\boldsymbol{\theta}) = \partial Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right)/\partial \boldsymbol{\theta}\big|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}$. The Hessian of the observed data log-likelihood, i.e. $\partial^2 \log \mathcal{L}_\ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\mathsf{T}$ can be obtained from the first-order Richardson extrapolation (Press *et al.*, 2007, Section 17.3) of the central difference for the gradient of the expected complete data log-likelihood. That is, the jth column of the Hessian matrix is given by

$$\frac{\mathbf{s}(\boldsymbol{\theta} - 2h\mathbf{u}_j) - 8\,\mathbf{s}(\boldsymbol{\theta} - h\mathbf{u}_j) + 8\,\mathbf{s}(\boldsymbol{\theta} + h\mathbf{u}_j) - \mathbf{s}(\boldsymbol{\theta} + 2h\mathbf{u}_j)}{12h},$$

where $\mathbf{u}_j$ is the jth co-ordinate vector with the jth element equal to 1 and others equal to 0 and h is a small positive value.

2.4 Pairwise likelihood formulation

Although the likelihood method provides a convenient way of obtaining the asymptotically efficient parameter estimators, the likelihood method relies on the validity of model assumptions. For instance, the output independence assumption in (2.6) may not hold for some applications, then the likelihood method would break down. To gain robustness to certain model assumptions, we now propose the pairwise likelihood method based on the composition of bivariate margins, which enjoys easier implementation and more robustness to misspecification of higher order association structures.

2.4.1 Non-informative observation process in the pairwise likelihood formulation

The pairwise likelihood is the product of all the bivariate densities of each distinct pair of observations. The pairwise likelihood for an individual takes the form of

$$\mathcal{L}_p = \prod_{r=1}^{m-1} \prod_{s=r+1}^{m} \Pr\left\{S^*(t_r) = s_r^*, S^*(t_s) = s_s^*; T_r = t_r, T_s = t_s; M = m\right\},$$

where

$$\begin{aligned}
& \Pr\left\{S^*(t_r) = s_r^*, S^*(t_s) = s_s^*; T_r = t_r, T_s = t_s; M = m\right\} \\
= & \sum_{s_r,\, s_s} \Pr\left\{S^*(t_r) = s_r^*, S^*(t_s) = s_s^*; S(t_r) = s_r, S(t_s) = s_s; T_r = t_r, T_s = t_s; M = m\right\} \\
= & \sum_{s_r,\, s_s} \Big[\Pr\left\{S^*(t_r) = s_r^*, S^*(t_s) = s_s^* \mid S(t_r) = s_r, S(t_s) = s_s; T_r = t_r, T_s = t_s; M = m\right\} \\
& \times \Pr\left\{S(t_r) = s_r, S(t_s) = s_s; T_r = t_r, T_s = t_s; M = m\right\} \Big].
\end{aligned}$$

Therefore, the conditions for the non-informative sampling scheme under the pairwise likelihood formulation is as follows:

1. The conditional probability of the sth observed state, given the rth observed state, as well as the rth and sth true states and observation times, is independent of observation times and the number of observations, i.e..

$$\begin{aligned}
& \Pr\left\{S^*(t_s) = s_s^* \mid S^*(t_r) = s_r^*, S(t_r) = s_r, S(t_s) = s_s; T_r = t_r, T_s = t_s; M = m\right\} \\
= & \Pr\left\{S^*(t_s) = s_s^* \mid S^*(t_r) = s_r^*, S(t_r) = s_r, S(t_s) = s_s\right\}, \qquad 1 \le r < s \le m; \quad (2.7)
\end{aligned}$$

2. The probability of staying in state s_s at time t_s, given the rth true state, is independent of whether an observation is carried out at this time and the rth observation time, as well as the number of the observation times, i.e.,

$$\begin{aligned}&\Pr\left\{S\left(t_s\right)=s_s \mid S\left(t_r\right)=s_r, T_r=t_r, T_s=t_s; M=m\right] \\ &\qquad\qquad = \Pr\left[S\left(t_s\right)=s_s \mid S\left(t_r\right)=s_r\right\}, \quad 1 \leq r<s \leq m;\end{aligned}$$

3. The conditional distribution of the sth observation time T_s,
   $\Pr\left(T_s=t_s \mid T_r=t_r, S_r=s_r, M=m\right)$, is functionally independent of parameters governing transition intensities of the disease process $\left\{S\left(t\right), t \geq 0\right\}$ and those related to the misclassification of the true states.

Equation (2.7) can be further simplified into

$$\Pr\left\{S^*\left(t_s\right)=s_s^* \mid S^*\left(t_r\right)=s_r^*, S\left(t_r\right)=s_r, S\left(t_s\right)=s_s\right\}=\Pr\left\{S^*\left(t_s\right)=s_s^* \mid S\left(t_s\right)=s_s\right\}.$$

2.4.2 Pairwise EM algorithm

The EM algorithm can be straightforwardly extended to maximization of the pairwise likelihood in HMMs. We now introduce the pairwise EM algorithm for the progressive model with misclassification. Technical details are described in Section 2.8.2.

The complete data pairwise log-likelihood for individual ℓ is

$$\begin{aligned}\log \mathcal{L}_{p\ell}^{\mathrm{c}}\left(\boldsymbol{\theta}\right) \;=\; & \sum_{s=2}^{m_\ell}\Bigg[\log\left\{\Pr\left(S_{\ell 1}^* \mid S_{\ell 1}, \mathbf{c}_{\ell 1}; \boldsymbol{\alpha}_{s_{\ell 1}}\right)\right\}+\log\left\{\Pr\left(S_{\ell s}^* \mid S_{\ell s}, \mathbf{c}_{\ell s}; \boldsymbol{\alpha}_{s_{\ell s}}\right)\right\} \\ & +\log\left\{\Pr\left(S_{\ell s} \mid S_{\ell 1}, \mathbf{x}_\ell; \boldsymbol{\beta}\right)\right\}\Bigg]\end{aligned}$$

$$+\sum_{r=2}^{m_\ell-1}\sum_{s=r+1}^{m_\ell}\left[\log\left\{\Pr\left(S^*_{\ell r}\mid S_{\ell r},\mathbf{c}_{\ell r};\boldsymbol{\alpha}_{s_{\ell r}}\right)\right\}+\log\left\{\Pr\left(S^*_{\ell s}\mid S_{\ell s},\mathbf{c}_{\ell s};\boldsymbol{\alpha}_{s_{\ell s}}\right)\right\}\right.$$
$$\left.+\log\left\{\Pr\left(S_{\ell s}\mid S_{\ell r},\mathbf{x}_\ell;\boldsymbol{\beta}\right)\right\}+\log\left[\sum_{s_{\ell 1}}\left\{\Pr\left(S_{\ell r}\mid S_{\ell 1},\mathbf{x}_\ell;\boldsymbol{\beta}\right)\Pr\left(S_{\ell 1}\right)\right\}\right]\right].$$

In the E step, the expected complete data pairwise log-likelihood at the $(k+1)$th iteration is given by $Q_p\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}\right)=\sum_{\ell=1}^N Q_{p\ell}(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)})$, where

$$\begin{aligned}
Q_{p\ell}\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}\right) &= \sum_{s=2}^{m_\ell}E\left[\log\left\{\Pr\left(S^*_{\ell 1}\mid S_{\ell 1},\mathbf{c}_{\ell 1};\boldsymbol{\alpha}_{s_{\ell 1}}\right)\right\}+\log\left\{\Pr\left(S^*_{\ell s}\mid S_{\ell s},\mathbf{c}_{\ell s};\boldsymbol{\alpha}_{s_{\ell s}}\right)\right\}\right.\\
&\quad\left.+\log\left\{\Pr\left(S_{\ell s}\mid S_{\ell 1},\mathbf{x}_\ell;\boldsymbol{\beta}\right)\right\}\;\middle|\; S^*_{\ell 1},S^*_{\ell s},\mathbf{c}_{\ell 1},\mathbf{c}_{\ell s};\boldsymbol{\theta}^{(k)}\right]\\
&\quad+\sum_{r=2}^{m_\ell-1}\sum_{s=r+1}^{m_\ell}E\left[\log\left\{\Pr\left(S^*_{\ell r}\mid S_{\ell r},\mathbf{c}_{\ell r};\boldsymbol{\alpha}_{s_{\ell r}}\right)\right\}\right.\\
&\quad\left.+\log\left\{\Pr\left(S^*_{\ell s}\mid S_{\ell s},\mathbf{c}_{\ell s};\boldsymbol{\alpha}_{s_{\ell s}}\right)\right\}+\log\left\{\Pr\left(S_{\ell s}\mid S_{\ell r},\mathbf{x}_\ell;\boldsymbol{\beta}\right)\right\}\right.\\
&\quad\left.+\log\left[\sum_{s_{\ell 1}}\left\{\Pr\left(S_{\ell r}\mid S_{\ell 1},\mathbf{x}_\ell;\boldsymbol{\beta}\right)\Pr\left(S_{\ell 1}\right)\right\}\right]\;\middle|\; S^*_{\ell r},S^*_{\ell s},\mathbf{c}_{\ell r},\mathbf{c}_{\ell s};\boldsymbol{\theta}^{(k)}\right].
\end{aligned}$$

Similarly to the EM algorithm in Section 2.3, maximization can be carried out with respect to the parameters for the observation process and the underlying process separately in the M-step, where $Q_p\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}\right)=\sum_{i=1}^K Q_{pi}\left(\boldsymbol{\alpha}_i,\boldsymbol{\theta}^{(k)}\right)+Q_{p,K+1}\left(\boldsymbol{\beta},\boldsymbol{\theta}^{(k)}\right)$. The function

$$\begin{aligned}
Q_{pi}\left(\boldsymbol{\alpha}_i,\boldsymbol{\theta}^{(k)}\right) &= \sum_{\ell=1}^N\left[\sum_{s=2}^{m_\ell}\left[E\left[\log\left\{\Pr\left(S^*_{\ell 1}\mid S_{\ell 1}=i,\mathbf{c}_{\ell 1};\boldsymbol{\alpha}_i\right)\right\}\;\middle|\; S^*_{\ell 1},S^*_{\ell s},\mathbf{c}_{\ell 1},\mathbf{c}_{\ell s},\mathbf{x}_\ell;\boldsymbol{\theta}^{(k)}\right]\right.\right.\\
&\quad\left.+E\left[\log\left\{\Pr\left(S^*_{\ell s}\mid S_{\ell s}=i,\mathbf{c}_{\ell s};\boldsymbol{\alpha}_i\right)\right\}\;\middle|\; S^*_{\ell 1},S^*_{\ell s},\mathbf{c}_{\ell 1},\mathbf{c}_{\ell s},\mathbf{x}_\ell;\boldsymbol{\theta}^{(k)}\right]\right]\\
&\quad+\sum_{r=2}^{m_\ell-1}\sum_{s=r+1}^{m_\ell}\left[E\left[\log\left\{\Pr\left(S^*_{\ell r}\mid S_{\ell r}=i,\mathbf{c}_{\ell r};\boldsymbol{\alpha}_i\right)\right\}\;\middle|\; S^*_{\ell r},S^*_{\ell s},\mathbf{c}_{\ell r},\mathbf{c}_{\ell s},\mathbf{x}_\ell;\boldsymbol{\theta}^{(k)}\right]\right.
\end{aligned}$$

$$+ E\left[\log \left\{ \Pr\left(S^*_{\ell s} \mid S_{\ell s} = i, \mathbf{c}_{\ell s}; \boldsymbol{\alpha}_i\right) \right\} \,\middle|\, S^*_{\ell r}, S^*_{\ell s}, \mathbf{c}_{\ell r}, \mathbf{c}_{\ell s}, \mathbf{x}_\ell; \boldsymbol{\theta}^{(k)} \right] \Bigg] \Bigg]$$

is maximized with respect to $\boldsymbol{\alpha}_i$ where $i = 1, \ldots, K$ and

$$\begin{aligned} Q_{p,K+1}\left(\boldsymbol{\beta}, \boldsymbol{\theta}^{(k)}\right) &= \sum_{s=2}^{m_\ell} E\left[\log \left\{ \Pr\left(S_{\ell s} \mid S_{\ell 1}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \right\} \,\middle|\, S^*_{\ell 1}, S^*_{\ell s}, \mathbf{c}_{\ell 1}, \mathbf{c}_{\ell s}, \mathbf{x}_\ell; \boldsymbol{\theta}^{(k)} \right] \\ &\quad + \sum_{r=2}^{m_\ell - 1} \sum_{s=r+1}^{m_\ell} E\Bigg[\log \left\{ \Pr\left(S_{\ell s} \mid S_{\ell r}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \right\} \\ &\quad + \log \left[\sum_{s_{\ell 1}} \left\{ \Pr\left(S_{\ell r} \mid S_{\ell 1}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \Pr\left(S_{\ell 1}\right) \right\} \right] \,\Bigg|\, S^*_{\ell r}, S^*_{\ell s}, \mathbf{c}_{\ell r}, \mathbf{c}_{\ell s}, \mathbf{x}_\ell; \boldsymbol{\theta}^{(k)} \Bigg] \end{aligned}$$

is maximized with respect to $\boldsymbol{\beta}$ by the Newton-Raphson algorithm. The estimator of parameters can be obtained through iterations between the E and M steps until convergence of $\boldsymbol{\theta}^{(k)}$.

2.4.3 Variance estimation in the pairwise likelihood formulation

Under some regularity conditions, the maximum pairwise likelihood estimators, $\tilde{\boldsymbol{\theta}}$, is asymptotically normally distributed: $\sqrt{N}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathbf{N}\left\{\mathbf{0}, \mathbf{G}^{-1}\left(\boldsymbol{\theta}\right)\right\}$, where $\mathbf{G}\left(\boldsymbol{\theta}\right)$ is the Godambe information matrix (Godambe, 1960), given by $\mathbf{G}\left(\boldsymbol{\theta}\right) = \mathbf{H}\left(\boldsymbol{\theta}\right)\mathbf{J}\left(\boldsymbol{\theta}\right)^{-1}\mathbf{H}\left(\boldsymbol{\theta}\right)$ with the sensitivity matrix $\mathbf{H}\left(\boldsymbol{\theta}\right) = E\left\{\partial^2 \log \mathcal{L}_p\left(\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}\right\}$ and the variability matrix $\mathbf{J}\left(\boldsymbol{\theta}\right) = \mathrm{Var}\left\{\partial \log \mathcal{L}_p\left(\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}\right\}$.

In the pairwise EM algorithm, the sensitivity and variability matrices can be estimated based on the expected complete data pairwise log-likelihood. The sensitivity matrix can be estimated by

$$\begin{aligned} \widehat{\mathbf{H}}\left(\boldsymbol{\theta}\right) &= \frac{1}{N}\sum_{\ell=1}^{N} \left. \frac{\partial^2 Q_{p\ell}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}} \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{(k)}=\tilde{\boldsymbol{\theta}}} \\ &\quad + \frac{1}{N}\sum_{\ell=1}^{N}\sum_{r=1}^{m_\ell - 1}\sum_{s=r+1}^{m_\ell} E\left[\left\{ \frac{\partial \log \Pr\left(S_{\ell r}, S_{\ell s}, S^*_{\ell r}, S^*_{\ell s}; \boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}} \right\}^{\otimes 2} \,\middle|\, S^*_{\ell r}, S^*_{\ell s}; \boldsymbol{\theta} \right] \Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \end{aligned}$$

$$-\frac{1}{N}\sum_{\ell=1}^{N}\sum_{r=1}^{m_\ell-1}\sum_{s=r+1}^{m_\ell}\left[E\left\{\frac{\partial \log \Pr\left(S_{\ell r},S_{\ell s},S^*_{\ell r},S^*_{\ell s};\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\,\middle|\,S^*_{\ell r},S^*_{\ell s};\boldsymbol{\theta}\right\}\right]^{\otimes 2}\Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}},$$

and the variability matrix can be estimated by

$$\widehat{\mathbf{J}}\left(\boldsymbol{\theta}\right)=\frac{1}{N}\sum_{\ell=1}^{N}\left\{\frac{\partial Q_{p\ell}\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}\right)}{\partial\boldsymbol{\theta}}\,\Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}},\boldsymbol{\theta}^{(k)}=\tilde{\boldsymbol{\theta}}}\right\}^{\otimes 2}.$$

Technical details of the information for the pairwise likelihood are placed in Section 2.8.2.

2.5 Simulation studies

Simulation studies are conducted to evaluate the performance of the MLEs and MPLEs for the progressive model with misclassification, as opposed to that of the naive MLEs and MPLEs obtained using the carry-backward method to adjust for the classification error (Couto *et al.*, 2002). For the transitions from a higher state to a lower one, the carry-backward method replaces the higher state by the lower one, where the classification error is ignored in the analysis of the progressive multi-state model.

2.5.1 Simulation setting

Simulation studies assess the progressive models with $K = 3$ states under the Markov assumption. The number of individuals is $N = 500$ and a total of 5000 replications are used in the progressive model with misclassification.

Each individual is assumed to start from state 1 at the initial time $t_{\ell 0} = 0$ and be observed at six equally spaced examination times, $t_{\ell 1} = 1$, $t_{\ell 2} = 2$, $t_{\ell 3} = 3$, $t_{\ell 4} = 4$, $t_{\ell 5} = 5$, and $t_{\ell 6} = 6$. One unit uniformly distributed time-dependent misclassification predictor $c\left(t\right)$, and a fixed prognostic

covariate x following the standard normal distribution are introduced. In the transition intensity model (1), we set $\beta_{10} = -1.0$, $\beta_{11} = 0.6$, $\beta_{20} = -0.7$, and $\beta_{21} = 0.4$, such that the mean sojourn times from state 1 to 2 and from state 2 to 3 are 3.25 and 2.18, respectively (see Section 2.8.3).

For simplicity, we assume no misclassification for the initial state of each individual and non-adjacent states and consider the misclassification models with the same regression coefficients. Specifically, we set $S^*_{\ell 0} = 1$ and in the misclassification model (2), $\alpha_{120} = \alpha_{210} = \alpha_{230} = \alpha_{320} = \alpha_0$, $\alpha_{121} = \alpha_{121} = \alpha_{231} = \alpha_{321} = \alpha_1$, and $\alpha_{130} = \alpha_{310} = \alpha_{131} = \alpha_{311} = 0$. We consider three degrees of the misclassification to investigate the effects of misclassification: (a) 5% misclassification rate $(\alpha_0, \alpha_1) = (-2.50, -1.50)$; (b) 15% misclassification rate $(\alpha_0, \alpha_1) = (-1.50, -0.90)$; (c) 30% misclassification rate $(\alpha_0, \alpha_1) = (-0.75, -0.55)$. More details on the setting of the misclassification model can be found in Section 2.8.3.

2.5.2 Simulation results

Table 2.1 summarizes the results for the progressive models with misclassification obtained from the EM and pairwise EM algorithms. For all three scenarios, the proposed MLEs and MPLEs have little biases, while the naive MLEs and MPLEs are biased. The asymptotic standard errors (ASEs) agree well with the empirical standard errors (ESEs), irrespective of the degree of the misclassification. The proposed methods yield the estimators with larger standard errors than naive estimators, showing the trade-off between the bias correction and the variance inflation as commonly observed in the context of measurement error. On the other hand, the standard errors of MPLEs are generally larger than those of MLEs regardless of the degree of the misclassification, suggesting that the robustness of the pairwise likelihood is achieved at the price of some loss of efficiency. When the degree of the misclassification increases, the ASEs and ESEs for the proposed estimators become larger, as expected. In conclusion, the proposed methods perform satisfactorily

to correct the misclassification effects.

Table 2.1: Simulation results for progressive models with misclassification based on the EM algorithm

	True States				5% Misclassification				15% Misclassification				30% Misclassification			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
								Naive MLEs								
β_{10}	.000	.048	.047	95.2	.019	.046	.050	90.1	.068	.043	.049	62.9	.178	.037	.048	2.2
β_{11}	.001	.048	.048	94.2	−.028	.047	.052	86.4	−.095	.045	.052	43.4	−.210	.045	.054	1.2
β_{20}	.000	.056	.054	95.2	.023	.055	.059	90.9	.042	.051	.056	84.4	−.026	.041	.058	77.7
β_{21}	.001	.055	.055	94.9	−.049	.054	.060	80.8	−.154	.053	.058	18.0	−.298	.053	.062	0.2
								Naive MPLEs								
β_{10}	.002	.052	.052	95.4	−.020	.051	.051	93.0	−.045	.048	.048	84.5	−.035	.044	.044	87.4
β_{11}	.002	.057	.058	94.2	−.017	.056	.057	93.2	−.053	.052	.053	81.8	−.115	.047	.048	31.1
β_{20}	.001	.061	.060	95.3	−.072	.057	.058	75.8	−.231	.053	.053	0.7	−.515	.051	.051	0.0
β_{21}	.003	.066	.066	94.9	−.072	.061	.062	75.9	−.170	.053	.053	12.1	−.247	.049	.049	0.2
								MLEs based on the EM algorithm								
β_{10}	—	—	—	—	.002	.051	.050	95.6	.002	.053	.052	95.1	.002	.061	.060	95.4
β_{11}	—	—	—	—	.002	.055	.053	95.7	.001	.058	.057	95.5	.001	.067	.065	95.6
β_{20}	—	—	—	—	.000	.060	.060	95.0	.000	.070	.070	95.2	.002	.105	.104	95.5
β_{21}	—	—	—	—	.001	.066	.065	95.3	.002	.078	.075	95.7	.006	.109	.104	96.1
α_0	—	—	—	—	−.037	.205	.214	93.8	−.016	.137	.146	92.2	−.003	.118	.127	92.4
α_1	—	—	—	—	−.009	.441	.459	93.6	−.004	.262	.281	92.2	−.006	.208	.224	92.4
								MPLEs based on the pairwise EM algorithm								
β_{10}	—	—	—	—	.002	.051	.050	95.4	.002	.054	.053	95.3	.002	.068	.064	96.3
β_{11}	—	—	—	—	.002	.056	.055	94.9	.001	.060	.058	95.5	.003	.073	.070	95.6
β_{20}	—	—	—	—	.000	.064	.062	95.4	.002	.078	.076	95.6	.005	.141	.130	95.7
β_{21}	—	—	—	—	.001	.069	.068	95.3	.002	.084	.082	95.4	.005	.129	.121	95.5
α_0	—	—	—	—	−.010	.249	.243	96.0	−.001	.169	.168	95.1	−.002	.142	.135	94.9
α_1	—	—	—	—	−.014	.542	.533	95.3	−.010	.328	.322	95.4	−.012	.251	.241	95.1

500 individuals × 5000 replicates; $X \sim \mathrm{N}(0,1)$, $C(t) \sim \mathrm{U}(0,1)$
parameter values in the transition intensity model: $(\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}) = (-1.0, 0.6, -0.7, 0.4)$
5% misclassification rate: $(\alpha_0, \alpha_1) = (-2.50, -1.50)$
15% misclassification rate: $(\alpha_0, \alpha_1) = (-1.50, -0.90)$
30% misclassification rate: $(\alpha_0, \alpha_1) = (-0.75, -0.55)$

2.6 Application to post-heart-transplant cardiac allograft vasculopathy

We apply the proposed methods to analyze the coronary allograft vasculopathy (CAV) data (Sharples *et al.*, 2003). The data contains 662 heart transplant recipients who survived at least one year after transplant and had at least one coronary angiography. Each recipient underwent the angiogram approximately annually after transplant or biennially after the first angiogram, at which CAV can be diagnosed. Recipients were followed up until death or until their most recent coronary angiography if alive. Three hundred and twelve males (84.10%) out of 371 heart transplant recipients survived at the end of the study. The unbalance between the number of males and that of females causes the problem when we study the effect of the gender. To avoid this problem and the heterogeneity caused by the gender, we restrict our attention to 312 male survivors. We also drop 6 male recipients with missing primary diagnosis (reason for transplantation, variable `pdiag`) from the data in order to investigate the effect of `pdiag`. Therefore, there are 1321 state observations from 306 individuals in the data set we analyze.

CAV is a chronic disease which is regarded as irreversible. The gold standard for the diagnosis of CAV, intravascular ultrasound, is prohibitively expensive. Therefore, coronary angiography is commonly used to diagnose CAV. Based on the coronary angiography, each recipient was classified as CAV-free, mild CAV and moderate or severe CAV, denoted as states 1, 2, and 3, respectively. Since the performance of the angiography is not perfect, the classification of disease states is subject to error. The first observation state is assumed to be correctly classified, since each recipient is assumed to be CAV free at the beginning of the study. It is of interest to simultaneously estimate the diagnostic accuracy of coronary angiography and explore the effects of risk factors on CAV onset and progression. The following factors were assessed as risk factors

for CAV onset and progression: recipient and donor age, and preoperative ischemic heart disease (IHD). Whether the recipient experienced IHD before the heart transplantation is recorded in the primary diagnosis for transplantation (`pdiag`). The recipient age (variable `age`) is a time-dependent variable, and therefore we use the recipient age at the first observation as the covariate for convenience. Both continuous covariates, the recipient and donor age (variable `dage`), are standardized by the transformation $(x - \bar{x})/\mathrm{sd}(x)$, where $\bar{x}$ is the sample mean of the covariate and $\mathrm{sd}(x)$ is the standard deviation of the covariate.

2.6.1 Sensitivity analysis

We first conduct the analysis to evaluate the sensitivity of the likelihood and pairwise likelihood methods. In particular, the parameters in the transition intensity models are estimated at three different degrees of misclassification, in which the misclassification rates are fixed according to the lower and upper bounds of 95% confidence intervals and MLEs obtained from the complete data by Jackson (2011). Table 2.2 presents the results for the sensitivity analysis for the progressive model with all three covariates. The values of the specified misclassification rates are listed at the end of the table. Both likelihood and pairwise likelihood methods suggest that the recipient age has no significant effect on the CAV onset or progression at 5% significance level. Therefore, we drop the variable `age` in the model and then investigate the effects of `IHD` and `dage`. Table 2.3 summarizes the results for the sensitivity analysis of the progressive model with covariates `IHD` and `dage`. The effects of both `IHD` and `dage` are significant on the CAV onset according to the outputs of both methods. The pairwise likelihood approach gives the similar results compared to the likelihood method in terms of estimates, standard errors and p-values at the same degree of misclassification. As expected, the standard errors of the estimates obtained by the pairwise likelihood method are consistently larger than those obtained by the likelihood

method, irrespectively of the degree of misclassification. Furthermore, although the standard errors for each parameters obtained by both methods become larger with the increment of the misclassification, the difference among the estimates and the test results at different levels of misclassification rates is not substantial, suggesting that the parameter estimates are not sensitive to the misclassification rates.

2.6.2 Progressive models with constant misclassification

We utilize the forward selection of risk factors for the progressive models with constant misclassification, in which we assume that the misclassification is independent of covariates. The results in the sensitivity analysis reveal that the IHD and the donor age are significant factors for the CAV onset, which agree with the conclusion drawn by Sharples *et al.* (2003). We start with the model including IHD and donor age in the initial analysis, and then add recipient age. The results, presented in Table 2.4, show that both `IHD` and `dage` have significant effects on the CAV onset but the effect of recipient age is not significant on either CAV onset or progression. In particular, recipients who were transplanted for IHD have a higher chance of CAV onset than recipients transplanted for other reasons; the older donor increases the risk of CAV. The estimates of misclassification probabilities are given in Table 2.5. The results obtained by both likelihood and pairwise likelihood methods agree that angiography is highly specific with CAV-free recipients. However, the estimated accuracy of coronary angiography for classifying two other disease states is different for different methods.

Table 2.2: Sensitivity Analysis for Progressive Model with `IHD`, `dage`, and `age`

			Likelihood			Pairwise Likelihood		
Transtion	Covariates		MLE	SE	p-value	MPLE	SE	p-value
					Mild Misclassification			
$1 \to 2$	Intercept	β_{10}	-3.220	0.217	0.000	-2.946	0.218	0.000
	IHD	β_{11}	0.739	0.269	0.006	0.408	0.303	0.179
	Donor age	β_{12}	0.517	0.125	0.000	0.296	0.157	0.060
	Recipient age	β_{13}	-0.215	0.156	0.166	0.206	0.153	0.179
$2 \to 3$	Intercept	β_{20}	-2.520	0.460	0.000	-2.914	0.563	0.000
	IHD	β_{21}	0.448	0.544	0.411	0.553	0.641	0.388
	Donor age	β_{22}	-0.089	0.239	0.711	0.251	0.258	0.330
	Recipient age	β_{23}	0.031	0.275	0.911	0.045	0.264	0.865
					Moderate Misclassification			
$1 \to 2$	Intercept	β_{10}	-3.205	0.218	0.000	-2.925	0.231	0.000
	IHD	β_{11}	0.733	0.275	0.008	0.421	0.322	0.192
	Donor age	β_{12}	0.537	0.126	0.000	0.304	0.170	0.074
	Recipient age	β_{13}	-0.220	0.141	0.117	0.210	0.154	0.171
$2 \to 3$	Intercept	β_{20}	-2.618	0.478	0.000	-3.122	0.673	0.000
	IHD	β_{21}	0.500	0.555	0.368	0.672	0.746	0.368
	Donor age	β_{22}	-0.093	0.284	0.742	0.289	0.290	0.318
	Recipient age	β_{23}	0.109	0.538	0.839	0.063	0.572	0.912
					Severe Misclassification			
$1 \to 2$	Intercept	β_{10}	-3.178	0.230	0.000	-2.865	0.247	0.000
	IHD	β_{11}	0.729	0.289	0.011	0.431	0.338	0.203
	Donor age	β_{12}	0.562	0.133	0.000	0.311	0.183	0.089
	Recipient age	β_{13}	-0.224	0.146	0.124	0.217	0.163	0.182
$2 \to 3$	Intercept	β_{20}	-2.713	0.550	0.000	-3.582	1.242	0.004
	IHD	β_{21}	0.564	0.631	0.371	1.009	1.354	0.456
	Donor age	β_{22}	-0.079	0.290	0.784	0.391	0.428	0.361
	Recipient age	β_{23}	0.171	0.371	0.644	0.077	0.670	0.909

Mild Misclassification: $\pi_{12} = 1.5\%$; $\pi_{21} = 10.1\%$, $\pi_{23} = 3.6\%$; $\pi_{32} = 5.7\%$

Moderate Misclassification: $\pi_{12} = 2.7\%$; $\pi_{21} = 17.5\%$, $\pi_{23} = 6.3\%$; $\pi_{32} = 11.5\%$

Severe Misclassification: $\pi_{12} = 4.5\%$; $\pi_{21} = 28.7\%$, $\pi_{23} = 10.7\%$; $\pi_{32} = 21.8\%$

Table 2.3: Sensitivity Analysis for Progressive Model with `IHD` and `dage`

			Likelihood			Pairwise Likelihood		
Transtion	Covariates		MLE	SE	p-value	MPLE	SE	p-value
					Mild Misclassification			
$1 \to 2$	Intercept	β_{10}	-3.100	0.188	0.000	-3.081	0.205	0.000
	IHD	β_{11}	0.568	0.232	0.015	0.566	0.284	0.046
	Donor age	β_{12}	0.472	0.117	0.000	0.346	0.156	0.027
$2 \to 3$	Intercept	β_{20}	-2.529	0.429	0.000	-2.934	0.560	0.000
	IHD	β_{21}	0.460	0.500	0.358	0.570	0.638	0.371
	Donor age	β_{22}	-0.086	0.235	0.714	0.253	0.266	0.343
					Moderate Misclassification			
$1 \to 2$	Intercept	β_{10}	-3.084	0.216	0.000	-3.065	0.218	0.000
	IHD	β_{11}	0.560	0.262	0.033	0.584	0.299	0.051
	Donor age	β_{12}	0.493	0.123	0.000	0.355	0.164	0.030
$2 \to 3$	Intercept	β_{20}	-2.633	0.491	0.000	-3.145	0.693	0.000
	IHD	β_{21}	0.521	0.575	0.364	0.693	0.772	0.370
	Donor age	β_{22}	-0.084	0.255	0.740	0.290	0.301	0.334
					Severe Misclassification			
$1 \to 2$	Intercept	β_{10}	-3.055	0.210	0.000	-3.013	0.239	0.000
	IHD	β_{11}	0.556	0.277	0.045	0.604	0.326	0.064
	Donor age	β_{12}	0.521	0.130	0.000	0.364	0.179	0.042
$2 \to 3$	Intercept	β_{20}	-2.737	0.566	0.000	-3.597	1.146	0.002
	IHD	β_{21}	0.595	0.664	0.370	1.020	1.223	0.404
	Donor age	β_{22}	-0.063	0.296	0.831	0.385	0.411	0.349

Mild Misclassification: $\pi_{12} = 1.5\%$; $\pi_{21} = 10.1\%$, $\pi_{23} = 3.6\%$; $\pi_{32} = 5.7\%$

Moderate Misclassification: $\pi_{12} = 2.7\%$; $\pi_{21} = 17.5\%$, $\pi_{23} = 6.3\%$; $\pi_{32} = 11.5\%$

Severe Misclassification: $\pi_{12} = 4.5\%$; $\pi_{21} = 28.7\%$, $\pi_{23} = 10.7\%$; $\pi_{32} = 21.8\%$

Table 2.4: Progressive Model with constant misclassification

			Likelihood			Pairwise Likelihood		
State	Covariates		MLE	SE	p-value	MPLE	SE	p-value
$1 \to 2$	Intercept	β_{10}	-3.076	0.252	0.000	-2.642	0.227	0.000
	IHD	β_{11}	0.561	0.255	0.028	0.525	0.275	0.056
	Donor age	β_{12}	0.507	0.126	0.000	0.319	0.152	0.036
$2 \to 3$	Intercept	β_{20}	-2.646	0.498	0.000	-3.273	0.706	0.000
	IHD	β_{21}	0.488	0.527	0.354	0.752	0.699	0.282
	Donor age	β_{22}	-0.095	0.242	0.694	0.337	0.296	0.254
Misclassification								
$1 \mapsto 2$	Intercept	α_{120}	-3.280	0.342	0.000	-18.476	13.743	0.179
$2 \mapsto 1$	Intercept	α_{210}	-1.207	0.469	0.010	-0.341	0.324	0.293
$2 \mapsto 3$	Intercept	α_{230}	-2.812	0.487	0.000	-2.933	0.885	0.001
$3 \mapsto 2$	Intercept	α_{320}	-2.895	1.037	0.005	-1.159	1.120	0.301
$1 \to 2$	Intercept	β_{10}	-3.191	0.302	0.000	-2.506	0.229	0.000
	IHD	β_{11}	0.730	0.292	0.012	0.366	0.285	0.198
	Donor age	β_{12}	0.550	0.134	0.000	0.269	0.155	0.083
	Recipient age	β_{13}	-0.220	0.168	0.191	0.201	0.146	0.170
$2 \to 3$	Intercept	β_{20}	-2.644	0.542	0.000	-3.231	0.732	0.000
	IHD	β_{21}	0.475	0.558	0.394	0.722	0.796	0.364
	Donor age	β_{22}	-0.099	0.228	0.663	0.329	0.309	0.287
	Recipient age	β_{23}	0.080	0.316	0.801	0.066	0.511	0.898
Misclassification								
$1 \mapsto 2$	Intercept	α_{120}	-3.298	0.416	0.000	-18.371	14.879	0.217
$2 \mapsto 1$	Intercept	α_{210}	-1.197	0.563	0.033	-0.331	0.345	0.336
$2 \mapsto 3$	Intercept	α_{230}	-2.802	0.490	0.000	-2.949	0.948	0.002
$3 \mapsto 2$	Intercept	α_{320}	-2.898	1.037	0.005	-1.135	1.224	0.354

$a \mapsto b$: true state a is misclassified to observed state b, $a, b = 1, 2, 3$.

Table 2.5: Estimated misclassification rates (in percent) for CAV states diagnosed by coronary angiography

	Observed state			Observed state		
True state	1	2	3	1	2	3
	MLE			MPLE		
1	96.4	3.6	0.0	100.0	0.0	0.0
2	22.0	73.6	4.4	40.3	56.7	3.0
3	0.0	5.2	94.8	0.0	23.9	76.1

2.7 Discussion

This chapter focuses on modelling the progressive multi-state model with misclassified states to understand the nature of the disease progression. Individuals under the disease progression study are often under irregular and not equal-spaced observation, and the true disease states may be subject to the classification error. In this chapter, we employ the continuous-time progressive HMM to simultaneously estimate the transition rates and account for state misclassification. The study of the relationship between the observed states and true states offers us a way to estimate the sensitivity and specificity of the diagnostic test for the disease. The proposed model is not limited to the scenario with the misclassified states; it can be applied to the case with discrete-valued surrogates observed for true states, such as the multiple sclerosis/magnetic resonance imaging lesion count data (Altman and Petkau, 2005).

To ensure the validity of the likelihood inference, the condition is derived for the interrelationship between the observation process and the sampling scheme, in addition to two conditions of the interrelationship between the disease process and the sampling scheme, discussed by Grüger *et al.* (1991). We develop the EM algorithm to obtain MLEs under the progressive model with misclassification. In the EM algorithm, Louis's formula (Louis, 1982) is usually employed for

the variance estimation. However, for hidden Markov models, it is difficult to evaluate the conditional expectation of the outer product of the score vector for the complete data likelihood due to the involvement of the conditional probabilities of pairs, triples and quadruples for the underlying states given the observed states. Therefore, we introduce an approximate formula for the Fisher information matrix (Jamshidian and Jennrich, 2000) and the sandwich-type robust variance estimation to protect against the potential correlation among observations. In addition, the pairwise EM algorithm, which requires the conditional probabilities given the pair of observed states instead of all observed states in the E-step, is proposed to obtain the MPLEs. The conditions of the non-informative sampling scheme is also derived for the pairwise likelihood approach. Although the pairwise likelihood method may incur certain efficiency loss due to the minimal model assumption, simulation studies demonstrate that this method gives reasonably comparable results to those obtained from the likelihood method.

Finally, identifiability can be a potential issue whenever fitting hidden Markov models. With continuous-time hidden Markov models, Bureau *et al.* (2003) and Rosychuk and Thompson (2004) discussed this problem for the two-state bidirectional model, while van den Hout *et al.* (2009) investigated this issue for the illness-death model. To ensure identifiability, Jackson *et al.* (2003) suggested that a rich source of data is helpful for fitting complex HMMs. The discussion of identifiability issues for general HMMs can be found in Cappé *et al.* (2005, Section 12.4). In the numerical studies we conducted, such an issue did not arise.

2.8 Technical Details

2.8.1 The complete-data likelihood for the progressive model with misclassification

The likelihood of the complete data with the true states observed for individual ℓ is

$$\begin{aligned}
\mathcal{L}_{\ell}^{\mathrm{c}}(\boldsymbol{\theta}) &= \Pr\left(S_{\ell 1}, \ldots, S_{\ell m_\ell}, S_{\ell 1}^*, \ldots, S_{\ell m_\ell}^* \;\middle|\; H_{\ell, m_\ell+1}^c, \mathbf{x}_\ell; \boldsymbol{\theta}\right) \\
&= \Pr\left(S_{\ell 1}^*, \ldots, S_{\ell m_\ell}^* \;\middle|\; S_{\ell 1}, \ldots, S_{\ell m_\ell}, H_{\ell, m_\ell+1}^c; \boldsymbol{\theta}\right) \Pr\left(S_{\ell 1}, \ldots, S_{\ell m_\ell}, \mid \mathbf{x}_\ell; \boldsymbol{\theta}\right) \\
&= \prod_{r=1}^{m_\ell} \left[\Pr\left(S_{\ell r}^* \mid S_{\ell r}, \mathbf{c}_{\ell r}; \boldsymbol{\alpha}_{s_{\ell r}}\right)\right] \Pr\left(S_{\ell 1}\right) \prod_{s=1}^{m_\ell - 1} \Pr\left(S_{\ell, s+1} \mid S_{\ell s}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \\
&\propto \prod_{r=1}^{m_\ell} \left[\Pr\left(S_{\ell r}^* \mid S_{\ell r}, \mathbf{c}_{\ell r}; \boldsymbol{\alpha}_{s_{\ell r}}\right)\right] \prod_{s=1}^{m_\ell - 1} \Pr\left(S_{\ell, s+1} \mid S_{\ell s}, \mathbf{x}_\ell; \boldsymbol{\beta}\right).
\end{aligned}$$

Then, the log-likelihood of the complete data with the true states observed for individual ℓ is

$$\log \mathcal{L}_{\ell}^{\mathrm{c}}(\boldsymbol{\theta}) = \sum_{r=1}^{m_\ell} \log \Pr\left(S_{\ell r}^* \mid S_{\ell r}, \mathbf{c}_{\ell r}; \boldsymbol{\alpha}_{s_{\ell r}}\right) + \sum_{s=1}^{m_\ell - 1} \log \Pr\left(S_{\ell, s+1} \mid S_{\ell s}, \mathbf{x}_\ell; \boldsymbol{\beta}\right).$$

2.8.2 Pairwise EM algorithm for the progressive model with misclassification

Complete data pairwise likelihood

The complete data pairwise likelihood with each pair of true states observed for individual ℓ is

$$\begin{aligned}
\mathcal{L}_{p\ell}^{\mathrm{c}}(\boldsymbol{\theta}) &= \prod_{r=1}^{m_\ell - 1} \prod_{s=r+1}^{m_\ell} \Pr\left(S_{\ell r}^*, S_{\ell s}^*, S_{\ell r}, S_{\ell s} \mid \mathbf{c}_{\ell r}, \mathbf{c}_{\ell s}, \mathbf{x}_\ell; \boldsymbol{\theta}\right) \\
&= \prod_{s=2}^{m_\ell} \Pr\left(S_{\ell 1}^*, S_{\ell s}^*, S_{\ell 1}, S_{\ell s} \mid \mathbf{c}_{\ell 1}, \mathbf{c}_{\ell s}, \mathbf{x}_\ell; \boldsymbol{\theta}\right)
\end{aligned}$$

$$
\begin{aligned}
&\times \prod_{r=2}^{m_\ell - 1} \prod_{s=r+1}^{m_\ell} \Pr\left(S^*_{\ell r}, S^*_{\ell s}, S_{\ell r}, S_{\ell s} \mid \mathbf{c}_{\ell r}, \mathbf{c}_{\ell s}, \mathbf{x}_\ell; \boldsymbol{\theta}\right) \\
= &\prod_{s=2}^{m_\ell} \Big[\Pr\left(S^*_{\ell 1} \mid S_{\ell 1}, \mathbf{c}_{\ell 1}; \boldsymbol{\alpha}_{s_{\ell 1}}\right) \Pr\left(S^*_{\ell s} \mid S_{\ell s}, \mathbf{c}_{\ell s}; \boldsymbol{\alpha}_{s_{\ell s}}\right) \\
&\times \Pr\left(S_{\ell s} \mid S_{\ell 1}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \Pr\left(S_{\ell 1}\right) \Big] \\
&\times \prod_{r=2}^{m_\ell - 1} \prod_{s=r+1}^{m_\ell} \Bigg\{ \Pr\left(S^*_{\ell r} \mid S_{\ell r}, \mathbf{c}_{\ell r}; \boldsymbol{\alpha}_{s_{\ell r}}\right) \Pr\left(S^*_{\ell s} \mid S_{\ell s}, \mathbf{c}_{\ell s}; \boldsymbol{\alpha}_{s_{\ell s}}\right) \\
&\times \Pr\left(S_{\ell s} \mid S_{\ell r}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \sum_{s_{\ell 1}} \Big[\Pr\left(S_{\ell r} \mid S_{\ell 1}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \Pr\left(S_{\ell 1}\right) \Big] \Bigg\} \\
\propto &\prod_{s=2}^{m_\ell} \Big[\Pr\left(S^*_{\ell 1} \mid S_{\ell 1}, \mathbf{c}_{\ell 1}; \boldsymbol{\alpha}_{s_{\ell 1}}\right) \Pr\left(S^*_{\ell s} \mid S_{\ell s}, \mathbf{c}_{\ell s}; \boldsymbol{\alpha}_{s_{\ell s}}\right) \Pr\left(S_{\ell s} \mid S_{\ell 1}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \Big] \\
&\times \prod_{r=2}^{m_\ell - 1} \prod_{s=r+1}^{m_\ell} \Bigg\{ \Pr\left(S^*_{\ell r} \mid S_{\ell r}, \mathbf{c}_{\ell r}; \boldsymbol{\alpha}_{s_{\ell r}}\right) \Pr\left(S^*_{\ell s} \mid S_{\ell s}, \mathbf{c}_{\ell s}; \boldsymbol{\alpha}_{s_{\ell s}}\right) \\
&\times \Pr\left(S_{\ell s} \mid S_{\ell r}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \sum_{s_{\ell 1}} \Big[\Pr\left(S_{\ell r} \mid S_{\ell 1}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \Pr\left(S_{\ell 1}\right) \Big] \Bigg\}.
\end{aligned}
$$

Then, the complete data pairwise log-likelihood for individual ℓ is

$$
\begin{aligned}
\log \mathcal{L}^{\mathrm{c}}_{p\ell}\left(\boldsymbol{\theta}\right) = &\sum_{s=2}^{m_\ell} \Big[\log \Pr\left(S^*_{\ell 1} \mid S_{\ell 1}, \mathbf{c}_{\ell 1}; \boldsymbol{\alpha}_{s_{\ell 1}}\right) + \log \Pr\left(S^*_{\ell s} \mid S_{\ell s}, \mathbf{c}_{\ell s}; \boldsymbol{\alpha}_{s_{\ell s}}\right) \\
&+ \log \Pr\left(S_{\ell s} \mid S_{\ell 1}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \Big] \\
&+ \sum_{r=2}^{m_\ell - 1} \sum_{s=r+1}^{m_\ell} \Bigg\{ \log \Pr\left(S^*_{\ell r} \mid S_{\ell r}, \mathbf{c}_{\ell r}; \boldsymbol{\alpha}_{s_{\ell r}}\right) + \log \Pr\left(S^*_{\ell s} \mid S_{\ell s}, \mathbf{c}_{\ell s}; \boldsymbol{\alpha}_{s_{\ell s}}\right) \\
&+ \log \Pr\left(S_{\ell s} \mid S_{\ell r}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) + \log \Bigg\{ \sum_{s_{\ell 1}} \Big[\Pr\left(S_{\ell r} \mid S_{\ell 1}, \mathbf{x}_\ell; \boldsymbol{\beta}\right) \Pr\left(S_{\ell 1}\right) \Big] \Bigg\} \Bigg\}.
\end{aligned}
$$

Godambe Information Matrix

Let $\tilde{\boldsymbol{\theta}}$ denote the maximum pairwise likelihood estimator for $\boldsymbol{\theta}$ and $\mathbf{S}^*$ be all m observations for one individual, where S_r and S_s denote the rth and the sth observation. The Godambe information matrix takes the form

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta})\,\mathbf{J}(\boldsymbol{\theta})^{-1}\,\mathbf{H}(\boldsymbol{\theta}),$$

where

$$\mathbf{H}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log \mathcal{L}_p(\boldsymbol{\theta};\mathbf{S}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\right] \quad \text{and} \quad \mathbf{J}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}\left[\frac{\partial \log \mathcal{L}_p(\boldsymbol{\theta};\mathbf{S}^*)}{\partial\boldsymbol{\theta}}\right].$$

Note that

$$\begin{aligned}\mathrm{Var}\left[\frac{\partial \log \mathcal{L}_p(\boldsymbol{\theta};\mathbf{S}^*)}{\partial\boldsymbol{\theta}}\right] &= E\left\{\left[\frac{\partial \log \mathcal{L}_p(\boldsymbol{\theta};\mathbf{S}^*)}{\partial\boldsymbol{\theta}}\right]^{\otimes 2}\right\} - \left\{E\left[\frac{\partial \log \mathcal{L}_p(\boldsymbol{\theta};\mathbf{S}^*)}{\partial\boldsymbol{\theta}}\right]\right\}^{\otimes 2} \\ &= E\left\{\left[\frac{\partial \log \mathcal{L}_p(\boldsymbol{\theta};\mathbf{S}^*)}{\partial\boldsymbol{\theta}}\right]^{\otimes 2}\right\}, \end{aligned} \tag{2.8}$$

and the expectation of the outer product of the score vector can be estimated by

$$\widehat{E}\left\{\left[\frac{\partial \log \mathcal{L}_p(\boldsymbol{\theta};\mathbf{S}^*)}{\partial\boldsymbol{\theta}}\right]^{\otimes 2}\right\} = \frac{1}{N}\sum_{\ell=1}^{N}\left[\left.\frac{\partial \log \mathcal{L}_p(\boldsymbol{\theta};\mathbf{S}^*_\ell)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\right]^{\otimes 2}. \tag{2.9}$$

Therefore, by (2.8) and (2.9), the variability matrix can be estimated by

$$\widehat{\mathbf{J}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{\ell=1}^{N}\left[\left.\frac{\partial \log \mathcal{L}_p(\boldsymbol{\theta};\mathbf{S}^*_\ell)}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\right]^{\otimes 2}. \tag{2.10}$$

Furthermore, because

$$\begin{aligned}
\frac{\partial \log \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} &= \frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \frac{\partial \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \\
&= \frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \frac{\partial}{\partial \boldsymbol{\theta}} \left[\sum_{S_r} \sum_{S_s} \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right) \right] \\
&= \frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \sum_{S_r} \sum_{S_s} \frac{\partial}{\partial \boldsymbol{\theta}} \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right) \\
&= \sum_{S_r} \sum_{S_s} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \left[\log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)\right] \frac{\Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \right\} \\
&= E\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right) \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right],
\end{aligned}$$

we have

$$\begin{aligned}
\frac{\partial \log \mathcal{L}_p\left(\boldsymbol{\theta}; \mathbf{S}^*\right)}{\partial \boldsymbol{\theta}} &= \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} \frac{\partial \log \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \\
&= \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} E\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right) \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right] \\
&= \frac{\partial Q_p\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}. \qquad (2.11)
\end{aligned}$$

Thus, by (2.10) and (2.11), we get

$$\widehat{\mathbf{J}}\left(\boldsymbol{\theta}\right) = \frac{1}{N} \sum_{\ell=1}^{N} \left\{ \frac{\partial Q_p\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}} \Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{(k)}=\tilde{\boldsymbol{\theta}}} \right\}^{\otimes 2}.$$

On the other hand,

$$\frac{\partial^2 \log \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} = \frac{\partial}{\partial \boldsymbol{\theta}} \left[\frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \frac{\partial \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^{\mathsf{T}}} \right]$$

$$
\begin{aligned}
&= -\frac{1}{\left[\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)\right]^2} \frac{\partial \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \frac{\partial \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^{\mathsf{T}}} \\
&\quad + \frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \frac{\partial^2 \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \\
&= -\left[\frac{\partial \log \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right]^{\otimes 2} + \frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \frac{\partial^2 \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \qquad (2.12) \\
&= -\left\{E\left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \middle| S_r^*, S_s^*; \boldsymbol{\theta}\right]\right\}^{\otimes 2} \\
&\quad + \frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \frac{\partial^2 \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}, \qquad (2.13)
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \frac{\partial^2 \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} &= \frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \left[\sum_{S_r} \sum_{S_s} \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)\right] \\
&= \frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \sum_{S_r} \sum_{S_s} \frac{\partial^2 \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}. \qquad (2.14)
\end{aligned}
$$

Similar to (2.12), we have

$$
\begin{aligned}
\frac{\partial^2 \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} &= -\left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right]^{\otimes 2} \\
&\quad + \frac{1}{\Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)} \frac{\partial^2 \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}},
\end{aligned}
$$

which yields

$$
\begin{aligned}
\frac{\partial^2 \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} &= \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right) \left\{\frac{\partial^2 \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}\right. \\
&\quad \left. + \left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right]^{\otimes 2}\right\}. \qquad (2.15)
\end{aligned}
$$

Then, by (2.14) and (2.15), we have

$$\begin{aligned}
\frac{1}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \frac{\partial^2 \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} &= \sum_{S_r} \sum_{S_s} \left\{ \left\{ \frac{\partial^2 \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \right.\right. \\
&\quad \left.\left. + \left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \right]^{\otimes 2} \right\} \frac{\Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)} \right\} \\
&= E\left[\frac{\partial^2 \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right] \\
&\quad + E\left\{ \left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \right]^{\otimes 2} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right\}. \qquad (2.16)
\end{aligned}$$

Therefore, by (2.13) and (2.16), we get

$$\begin{aligned}
\frac{\partial^2 \log \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} &= -\left\{ E\left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right] \right\}^{\otimes 2} \\
&\quad + E\left\{ \left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \right]^{\otimes 2} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right\} \\
&\quad + E\left[\frac{\partial^2 \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right] \\
&= E\left[\frac{\partial^2 \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right] \\
&\quad + \mathrm{Var}\left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right]. \qquad (2.17)
\end{aligned}$$

Thus, by (2.17), we have

$$\begin{aligned}
\frac{\partial^2 \log \mathcal{L}_p\left(\boldsymbol{\theta}; \mathbf{S}^*\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} &= \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} \frac{\partial^2 \log \Pr\left(S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \\
&= \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} \left\{ E\left[\frac{\partial^2 \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right] \right. \\
&\quad \left. + \mathrm{Var}\left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right] \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\partial^2 Q_p\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \\
&\quad + \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} E\left\{ \left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \right]^{\otimes 2} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right\} \\
&\quad - \sum_{r=1}^{m-1} \sum_{s=r+1}^{m} \left\{ E\left[\frac{\partial \log \Pr\left(S_r, S_s, S_r^*, S_s^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \middle| S_r^*, S_s^*; \boldsymbol{\theta} \right] \right\}^{\otimes 2}.
\end{aligned}
$$

Hence, the sensitivity matrix can be estimated by

$$
\begin{aligned}
\widehat{\mathbf{H}}\left(\boldsymbol{\theta}\right) &= \frac{1}{N} \sum_{\ell=1}^{N} \frac{\partial^2 Q_p\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{(k)}=\tilde{\boldsymbol{\theta}}} \\
&\quad + \frac{1}{N} \sum_{\ell=1}^{N} \sum_{r=1}^{m_\ell - 1} \sum_{s=r+1}^{m_\ell} E\left\{ \left[\frac{\partial \log \Pr\left(S_{\ell r}, S_{\ell s}, S_{\ell r}^*, S_{\ell s}^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \right]^{\otimes 2} \middle| S_{\ell r}^*, S_{\ell s}^*; \boldsymbol{\theta} \right\} \Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} \\
&\quad - \frac{1}{N} \sum_{\ell=1}^{N} \sum_{r=1}^{m_\ell - 1} \sum_{s=r+1}^{m_\ell} \left\{ E\left[\frac{\partial \log \Pr\left(S_{\ell r}, S_{\ell s}, S_{\ell r}^*, S_{\ell s}^*; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \middle| S_{\ell r}^*, S_{\ell s}^*; \boldsymbol{\theta} \right] \right\}^{\otimes 2} \Bigg|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}.
\end{aligned}
$$

2.8.3 Effects of parameters in simulation studies

Transition intensity model

In the unidirectional progressive model, the sojourn time τ_i is exponentially distributed with mean $1/q_i$, $i = 1, 2$, where $q_1 = \exp\left(\beta_{10} + \beta_{11} X\right)$ and $q_2 = \exp\left(\beta_{20} + \beta_{21} X\right)$ are transition intensities.

If the prognostic covariate X is simulated from the standard normal distribution, then the mean sojourn time is

$$
\begin{aligned}
E\left(\tau_i\right) &= E\left[1/\exp\left(\beta_{i0} + \beta_{i1} X\right)\right] \\
&= \int_{-\infty}^{\infty} \exp\left(-\beta_{i0} - \beta_{i1} u\right) \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \, \mathrm{d}u \\
&= \exp\left(\frac{1}{2}\beta_{i1}^2 - \beta_{i0}\right).
\end{aligned}
$$

The median sojourn time is $M(\tau_i) = \log(2) \cdot E(\tau_i)$.

If we set $\boldsymbol{\beta} = (-1.0, -0.7, 0.6, 0.4)^{\mathsf{T}}$ and $X \sim \mathrm{N}(0,1)$, then the mean and median sojourn times from State 1 to State 2 are 3.25 and 2.26 and the mean and median sojourn times from State 2 to State 3 are 2.18 and 1.51.

Misclassification model

The misclassification probabilities are given by

$$\begin{aligned}
&\Pr(S^*_{\ell r} = 2 \mid S_{\ell r} = 1, c_{\ell r}; \boldsymbol{\alpha}) = \Pr(S^*_{\ell r} = 2 \mid S_{\ell r} = 3, c_{\ell r}; \boldsymbol{\alpha}) = \frac{\exp(\alpha_0 + \alpha_1 c_{\ell r})}{1 + \exp(\alpha_0 + \alpha_1 c_{\ell r})};\\
&\Pr(S^*_{\ell r} = 1 \mid S_{\ell r} = 2, c_{\ell r}; \boldsymbol{\alpha}) = \Pr(S^*_{\ell r} = 3 \mid S_{\ell r} = 2, c_{\ell r}; \boldsymbol{\alpha}) = \frac{\exp(\alpha_0 + \alpha_1 c_{\ell r})}{1 + 2\exp(\alpha_0 + \alpha_1 c_{\ell r})};\\
&\Pr(S^*_{\ell r} = 3 \mid S_{\ell r} = 1, c_{\ell r}; \boldsymbol{\alpha}) = \Pr(S^*_{\ell r} = 1 \mid S_{\ell r} = 3, c_{\ell r}; \boldsymbol{\alpha}) = 0.
\end{aligned}$$

Note that the time-dependent misclassification covariate $c_{\ell r}$ follows a unit uniform distribution. Then, if the true state is 1 or 3, then the misclassification rate is

$$\int_0^1 \frac{\exp(\alpha_0 + \alpha_1 u)}{1 + \exp(\alpha_0 + \alpha_1 u)}\,\mathrm{d}u = \frac{1}{\alpha_1}\Big\{\log\big[1 + \exp(\alpha_0 + \alpha_1)\big] - \log\big[1 + \exp(\alpha_0)\big]\Big\};$$

if the true state is 2, then the total misclassification rate is

$$\int_0^1 \frac{2\exp(\alpha_0 + \alpha_1 u)}{1 + 2\exp(\alpha_0 + \alpha_1 u)}\,\mathrm{d}u = \frac{1}{\alpha_1}\Big\{\log\big[1 + 2\exp(\alpha_0 + \alpha_1)\big] - \log\big[1 + 2\exp(\alpha_0)\big]\Big\}.$$

We consider three degrees of the misclassification shown in Table 2.6 to investigate the effects of misclassification. The proportions of misclassification for each state are calculated by the expectation. In Table 2.6, we include the empirical overall proportions of misclassification which are obtained based on the simulated sample with 10,000,000 individuals.

Table 2.6: Proportion of misclassification for simulation study

		Proportion		
α_0	α_1	State 1 or 3	State 2	Overall
-2.50	-1.50	4.05%	7.74%	4.89%
-1.50	-0.90	12.73%	22.47%	14.93%
-0.75	-0.55	26.52%	41.83%	29.99%

Chapter 3

Analysis of panel data under hidden mover-stayer models

3.1 Introduction

Continuous-time, multi-state stochastic models provide a useful framework to analyze the longitudinal data when the interest lies in understanding the influence of risk factors on transitions. Parametric, nonparametric, and semiparametric methods can be used for the case where subjects are under continuous observation, and the exact transition times between states (Andersen *et al.*, 1993) are known. In contrast, when the subjects are observed at a sequence of time points, exact transition times may not be observed and thus interval-censored. In this case, the state occupied at each assessment and the information of risk factors are typically collected. Such data are often referred to as panel data (Kalbfleisch and Lawless, 1985; Cook *et al.*, 2002). In the analysis of panel data, the heterogeneity of the data and the state misclassification are two common issues. For example, in the smoking prevention study (Cameron *et al.*, 1999), a substantial number of

children may be non-smokers and therefore they would never experience smoking; in the study of cardiac allograft vasculopathy (Sharples *et al.*, 2003), the gold standard is prohibitively expensive and the disease is diagnosed by coronary angiography, so that the classification of disease states is subject to error.

In this chapter, we propose continuous-time hidden mover-stayer models to analyze the panel data, in which states may be subject to misclassification. Our proposed models provide a convenient tool to feature a paritcular type of heterogeneity of the data, which results in the invalidity of the time-homogeneous assumption in Markov models. This type of heterogeneity arises when the population consists of two types of subjects: the stayer stays in the initial state, whereas the mover evolves according to a Markov process. For example, in population studies of chronic degenerative diseases, a substantial proportion of subjects may be disease free over the study period so they may not experience degeneration (Cook *et al.*, 2002). For the discrete-time version, Frydman (1984) provided a recursive method for obtaining maximum likelihood estimates (MLEs) and Fuchs and Greenhouse (1988) presented an EM algorithm for estimation of model parameters. In the framework of continuous-time models, Cook *et al.* (2002) developed a generalized mover-stayer Markov model to accommodate heterogeneity in the patterns of movement between states by allowing subject-specific absorbing states. Recently, O'Keeffe *et al.* (2013) considered the use of various random effects distributions in the mover-stayer model.

In addition to handling the heterogeneous data, our proposed models simultaneously account for potential misclassification. Our proposed models are not limited to model the scenario where surrogate measurements are assumed to have the same range of values as the true covariate. We consider a generic setting: the discrete-valued surrogates are observed and the number of levels for surrogates can be different from that for the underlying state.

Hidden Markov models are commonly utilized in the analysis of panel data with misclassi-

fied states. Applications include the estimation of parasitic infection dynamics (Nagelkerke *et al.*, 1990; Rosychuk and Thompson, 2003), the analysis of hairy leukoplakia and cervical human papillomavirus infection (Bureau *et al.*, 2003), and the study of abdominal aortic aneurysms (Jackson and Sharples, 2002). Recently, the *msm* package in `R` was developed by Jackson (2011) to fit continuous-time Markov and hidden Markov models in the analysis of panel data. However, these papers, which mainly focus on estimating transition rates when the state classification is imperfect, do not incorporate the feature of the heterogeneity in the panel data.

The rest of this chapter is organized as follows. In Section 3.2, we describe mover-stayer models with misclassification. Our proposed models allow the effects of risk factors on both the underlying mover-stayer process and the misclassification process. The inference procedure based on the EM algorithm is proposed in Section 3.3. The proposed method is applied to a smoking prevention data in Section 3.4. The performance of the proposed method is investigated through simulation studies in Section 3.5. Discussion is given in Section 3.6. Technical details are presented in Section 3.7.

3.2 Mover-stayer models with misclassification

3.2.1 Mover-stayer models in continuous time

To define a mover-stayer model in continuous time, we first introduce a Markov model in continuous time with state space $\{1, 2, \ldots, K\}$. Let $S(t)$ denote the realization of the Markov process at time t. Let $\mathbf{M}(s, s+t)$ be the $K \times K$ transition probability matrix with (i, j) entry

$$M_{ij}(s, s+t) = \Pr[S(s+t) = j \mid S(s) = i]$$

for $s \geq 0$, $t > 0$, $i, j = 1, 2, \ldots, K$. The transition intensity from state i to j at time t is

$$q_{ij}(t) = \lim_{\Delta t \downarrow 0} \frac{M_{ij}(t, t + \Delta t)}{\Delta t}, \qquad i \neq j,$$

and as a convention, define

$$q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t).$$

Let $\mathbf{Q}(t)$ be the $K \times K$ transition intensity matrix with (i, j) entry $q_{ij}(t)$, $i, j = 1, 2, \ldots, K$.

This chapter is primarily concerned with time-homogeneous Markov models in which transition intensities are independent of t. We therefore let $q_{ij}(t) = q_{ij}$, $i, j = 1, \ldots, K$ and write $\mathbf{Q}(t) = \mathbf{Q}$. It follows that $\mathbf{M}(s, s + t) = \mathbf{M}(0, t)$, which can be written as $\mathbf{M}(t)$. By the result of the time-homogeneous Markov model (e.g. Cox and Miller, 1965, Chapter 4), transition probabilities can be obtained from transition intensities by the matrix exponential, that is

$$\mathbf{M}(t) = \exp(\mathbf{Q}t) = \sum_{r=0}^{\infty} \mathbf{Q}^r \frac{t^r}{r!}, \tag{3.1}$$

where the matrix exponential is defined by the power series of the matrix product and $\mathbf{Q}^0 = \mathbf{I}$. The algorithm for the computation of $\mathbf{M}(t)$ can be referred to Section 1.1.1.

In applications, transitions between states in the movers and the probability of being a mover are affected by certain covariates, and the primary interest lies in understanding the influence of these covariates.

Transition intensity model

Let $\mathbf{x}$ be a $p \times 1$ vector of prognostic variables. Consider the multiplicative models

$$q_{ij}(\mathbf{x}) = q_{ij0} \exp\left(\mathbf{x}^{\mathsf{T}} \boldsymbol{\beta}_{ijx}\right), \qquad i \neq j, i, j = 1, \ldots, K, \tag{3.2}$$

where $\boldsymbol{\beta}_{ijx} = (\beta_{ij1}, \beta_{ij2}, \ldots, \beta_{ijp})^{\mathsf{T}}$ is the vector of regression coefficients of primary interest, and q_{ij0} is the baseline transition intensity from state i to state j, which is reparameterized as $q_{ij0} = \exp(\beta_{ij0})$ with parameter β_{ij0} (e.g. Kalbfleisch and Lawless, 1985; Jackson *et al.*, 2003).

Logistic model for the mover-stayer distribution

Let Z be a Bernoulli variable where $Z = 0$ if the subject is a stayer, and $Z = 1$ if the subject is a mover. Let ω_k denote the conditional distribution of being a mover given the initial state k, i.e. $\Pr[Z = 1 \mid S_0 = k]$. To model the covariate effects on the mover-stayer probability, we employ the logistic model:

$$\log\left(\frac{\omega_k}{1 - \omega_k}\right) = \gamma_{k0} + \mathbf{x}^{\mathsf{T}} \boldsymbol{\gamma}_{kx}, \tag{3.3}$$

where $\boldsymbol{\gamma}_{kx} = (\gamma_{k1}, \ldots, \gamma_{kp})^{\mathsf{T}}$ are regression coefficients.

3.2.2 Misclassification model

It is common that the underlying true state, $S(t)$, can not be observed, but the surrogate state, $S^*(t)$, is observed. Assume that the observed state, $S^*(t)$, takes the value from $\{1, 2, \ldots, J\}$, where J can be identical to or different from K. Suppose a $q \times 1$ vector of predictor variables collected at time t, denoted by $\mathbf{c}(t)$, is associated with the observation process. For $i = 1, \ldots, K$,

and $j = 1, \ldots, J$, let

$$\pi_{ij}(t) = \Pr\left[S^*(t) = j \mid S(t) = i, \mathbf{c}(t)\right]$$

denote misclassification probabilities at time t.

Here we employ the multinomial logistic regression model to portray the relationship between misclassification probabilities and $\mathbf{c}(t)$ (e.g. Agresti, 2002, Chapter 7):

$$\log\left[\frac{\pi_{ij}(t)}{\pi_{ik_i}(t)}\right] = \alpha_{ij0} + \boldsymbol{\alpha}_{ijc}^{\mathsf{T}}\mathbf{c}(t), \qquad j \neq k_i, \tag{3.4}$$

where $k_i = \min\{j : \pi_{ij}(t) > 0\}$ is assumed to be time-independent but a function of the underlying state i, and α_{ij0} and $\boldsymbol{\alpha}_{ijc}$ are state-dependent but time-independent regression coefficients. Thus, misclassification probabilities can be written as

$$\pi_{ij}(t) = \begin{cases} \dfrac{1}{1 + \sum\limits_{k \neq k_i} \exp\left[\alpha_{ik0} + \boldsymbol{\alpha}_{ikc}^{\mathsf{T}}\mathbf{c}(t)\right]}, & j = k_i, \\ \dfrac{\exp\left[\alpha_{ij0} + \boldsymbol{\alpha}_{ijc}^{\mathsf{T}}\mathbf{c}(t)\right]}{1 + \sum\limits_{k \neq k_i} \exp\left[\alpha_{ik0} + \boldsymbol{\alpha}_{ikc}^{\mathsf{T}}\mathbf{c}(t)\right]}, & j \neq k_i. \end{cases} \tag{3.5}$$

In applications, suitable constraints may be imposed on misclassification probabilities to reflect a prior knowledge of the observation process. For example, Jackson *et al.* (2003) suggested that the probability of misclassification may be negligibly small for those states that are far apart, such that misclassification probabilities are non-zero constants only for some adjacent states, in the disease progression studies.

3.3 Maximum likelihood estimation

Suppose n subjects are under study and a randomly selected subject either stays in its initial state with a probability, or with a complementary probability, moves among K states according to a time-homogeneous Markov process. For $i = 1, 2, \ldots, n$, let $0 = t_{i0} < t_{i1} < \cdots < t_{im_i} < \infty$ be the assessment times of subject i, and S_{ij} and S^*_{ij} denote the underlying and observed states of subject i at time t_{ij}, respectively, $j = 0, 1, \ldots, m_i$. Let $\mathbf{x}_i$ and $\mathbf{c}_{ij}$ represent time-independent prognostic variables and time-dependent misclassification predictors for subject i at time t_{ij}, respectively, $i = 1, 2, \ldots, n$, $j = 1, \ldots, m_i$. Assume that the first state S_{i0} is known. We employ models (3.2) and (3.3) to postulate the underlying process and (3.4) to model the misclassification process. Let $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}_1^\mathsf{T}, \ldots, \boldsymbol{\alpha}_K^\mathsf{T}, \boldsymbol{\beta}^\mathsf{T}, \boldsymbol{\gamma}_1^\mathsf{T}, \ldots, \boldsymbol{\gamma}_K^\mathsf{T}\right)^\mathsf{T}$, where

$$\begin{aligned} \boldsymbol{\alpha}_i &= \left(\alpha_{ij0}, \boldsymbol{\alpha}_{ijc}^\mathsf{T} \colon j = 1, \ldots, J, j \neq k_i\right)^\mathsf{T}, \qquad i = 1, \ldots, K \\ \boldsymbol{\beta} &= \left(\beta_{ij0}, \beta_{ij1}, \ldots, \beta_{ijp} \colon i \neq j, i, j = 1, \ldots, K\right)^\mathsf{T}, \\ \text{and} \qquad \boldsymbol{\gamma}_i &= \left(\gamma_{i0}, \boldsymbol{\gamma}_{ix}^\mathsf{T}\right)^\mathsf{T}, \qquad i = 1, \ldots, K. \end{aligned}$$

We are intersted in estimating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_i$'s.

3.3.1 Estimation via an EM algorithm

Both the mover-stayer Bernoulli variable and underlying states can be treated as latent variables, and therefore the EM algorithm can be employed for parameter estimation. We now elaborate how to implement the EM algorithm to deal with parameter estimation pertinent to the mover-stayer models with misclassification. The complete data log-likelihood contributed from subject

i is written as

$$\begin{aligned}
\ell_i^c(\boldsymbol{\theta}) &= \log\Big\{\Pr\left(S_{i0}, S_{i1}, \ldots, S_{im_i}, S_{i1}^*, \ldots, S_{im_i}^*, Z_i \mid \mathbf{x}_i, \mathbf{c}_{i1}, \ldots, \mathbf{c}_{im_i}; \boldsymbol{\theta}\right)\Big\} \\
&= \log\Big\{\Pr\left(Z_i \mid S_{i0}, \mathbf{x}_i; \boldsymbol{\gamma}_{s_{i0}}\right)\Big\} + \sum_{j=1}^{m_i} \log\Big\{\Pr\left(S_{ij}^* \mid S_{ij}, \mathbf{c}_{ij}; \boldsymbol{\alpha}_{s_{ij}}\right)\Big\} \\
&\quad + Z_i \sum_{j=1}^{m_i} \log\Big\{\Pr\left(S_{ij} \mid S_{i,j-1}, \mathbf{x}_i, Z_i = 1; \boldsymbol{\beta}\right)\Big\}, \qquad (3.6)
\end{aligned}$$

where the output independence assumption is required, i.e.,

$$\Pr\left(S_{i1}^*, \ldots, S_{im_i}^* \mid S_{i1}, \ldots, S_{im_i}, Z_i; \mathbf{x}_i, \mathbf{c}_{i1}, \ldots, \mathbf{c}_{im_i}; \boldsymbol{\theta}\right) = \prod_{j=1}^{m_i} \Pr\left(S_{ij}^* \mid S_{ij}, \mathbf{c}_{ij}; \boldsymbol{\alpha}_{s_{ij}}\right).$$

In the expectation step, the expected complete data log-likelihood at the $(k+1)$th iteration is

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right) = \sum_{i=1}^{n} Q_i\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right),$$

where $\boldsymbol{\theta}^{(k)}$ is the maximizer of $\boldsymbol{\theta}$ at the kth iteration, $Q_i\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right) = E\left[\ell_i^c(\boldsymbol{\theta}) \;\middle|\; S_{i0}, \mathbf{S}_i^*; \boldsymbol{\theta}^{(k)}\right]$, and $\mathbf{S}_i^* = (S_{i1}^*, \ldots, S_{im_i}^*)$.

From (3.6), we can see that the parameters $\boldsymbol{\alpha}_k$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}_k$, $k = 1, \ldots, K$, are distinct from each other. Therefore, the maximization can be carried out separately using the following functions

$$\begin{aligned}
Q\left(\boldsymbol{\alpha}_k, \boldsymbol{\theta}^{(k)}\right) &= \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mu_{ij}(k) \log\Big[\Pr\left(S_{ij}^* \mid S_{ij} = k, \mathbf{c}_{ij}; \boldsymbol{\alpha}_k\right)\Big], \\
Q\left(\boldsymbol{\gamma}_k, \boldsymbol{\theta}^{(k)}\right) &= \sum_{i \in \{i:\, s_{i0} = k\}} \sum_{z=0}^{1} \kappa_i(z) \log\Big[\Pr\left(Z_i = z \mid S_{i0} = k, \mathbf{x}_i; \boldsymbol{\gamma}_k\right)\Big], \\
\text{and} \qquad Q\left(\boldsymbol{\beta}, \boldsymbol{\theta}^{(k)}\right) &= \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=1}^{K} \sum_{l=1}^{K} \xi_{ij}(k, l) \log\Big[\Pr\left(S_{ij} = l \mid S_{i,j-1} = k, \mathbf{x}_i, Z_i = 1; \boldsymbol{\beta}\right)\Big],
\end{aligned}$$

where conditional probabilities $\mu_{ij}(k)$, $\kappa_i(z)$, and $\xi_{ij}(k,l)$ are given by

$$\begin{aligned}
\mu_{ij}(k) &= \Pr\left\{S_{ij} = k \;\middle|\; S_{i0}, \mathbf{S}_i^*; \boldsymbol{\theta}^{(k)}\right\}; && (3.7)\\
\kappa_i(z) &= \Pr\left\{Z_i = z \;\middle|\; S_{i0}, \mathbf{S}_i^*; \boldsymbol{\theta}^{(k)}\right\}; && (3.8)\\
\xi_{ij}(k,l) &= \Pr\left\{S_{i,j-1} = k, S_{ij} = l, Z_i = 1 \;\middle|\; S_{i0}, \mathbf{S}_i^*; \boldsymbol{\theta}^{(k)}\right\}. && (3.9)
\end{aligned}$$

The maximum likelihood estimates, $\hat{\boldsymbol{\theta}}$, can be obtained through iterations between E and M steps until convergence of $\boldsymbol{\theta}^{(k)}$.

3.3.2 Forward and backward probabilities

Now we describe the algorithm to calculate the conditional probabilities (3.7) and (3.9). Denote the forward probability $\Pr(S_0, S_1^*, \ldots, S_k^*, S_k = j, Z = 1)$ by $\lambda_k(j)$, $k = 1, \ldots, m$, and the backward probability $\Pr\left(S_{k+1}^*, \ldots S_m^* \mid S_k = j, Z = 1\right)$ by $\phi_k(j)$, $k = 1, \ldots, m-1$, where $j = 1, \ldots, K$. For convenience, we define $\phi_m(j) = 1$ for $j = 1, \ldots, K$.

Then, we have

$$\lambda_1(j) \propto \omega_{s_0} M_{s_0,j}(t_1 - t_0)\, \pi_{j s_1^*}(t_1), \qquad j = 1, \ldots, K,$$

where ω_{s_0} is the conditional probability of being a mover given the initial state s_0, modelled by (3.3); $M_{s_0,j}(t_1 - t_0)$ is the transition probability for the mover, defined by (3.1); $\pi_{j s_1^*}(t_1)$ is the misclassification probability, defined by (3.5). Let $\boldsymbol{\lambda}_k$ be the row vector with the jth component $\lambda_k(j)$. It can be recursively calculated as follows:

$$\boldsymbol{\lambda}_k = \boldsymbol{\lambda}_{k-1} \mathbf{N}_k, \qquad k = 2, \ldots, m,$$

where the jth column of $K \times K$ matrix $\mathbf{N}_k$ is $\{M_{ij}(t_k - t_{k-1})\,\pi_{js_k^*}(t_k), i = 1, \ldots, K\}$.

Similarly, the vector $\boldsymbol{\phi}_k$ with the jth component $\phi_k(j)$ can be recursively calculated as follows:

$$\begin{aligned} \boldsymbol{\phi}_{m-1} &= \mathbf{M}(t_m - t_{m-1})\,\boldsymbol{\pi}_{\cdot,s_m^*}(t_m); \\ \boldsymbol{\phi}_{k-1} &= \mathbf{N}_k \boldsymbol{\phi}_k, \qquad k = 2, \ldots, m-1; \end{aligned}$$

where $\boldsymbol{\pi}_{\cdot,s_m^*}$ is a column vector whose jth component is $\pi_{i,s_m^*}(t_m)$.

Based on the properties of forward and backward probabilities in hidden Markov models (e.g. Zucchini and MacDonald, 2009, Chapter 4), we have

$$\Pr(S_i = j, Z = 1, S_0, \mathbf{S}^*) = \lambda_i(j)\,\phi_i(j), \tag{3.10}$$

and

$$\Pr(S_{i-1} = j, S_i = k, Z = 1, S_0, \mathbf{S}^*) = \lambda_{i-1}(j)\,M_{jk}(t_i - t_{i-1})\,\pi_{ks_i^*}(t_i)\,\phi_i(k), \tag{3.11}$$

where $\phi_m(k) = 1$. Then, the conditional probability $\xi_{ij}(k,l)$ can be calculated by (3.11) and the conditional probability $\Pr\left\{S_{ij} = k, Z_i = 1 \,\middle|\, S_{i0}, \mathbf{S}_i^*; \boldsymbol{\theta}^{(k)}\right\}$ in $\mu_{ij}(k)$ can be calculated by (3.10).

3.3.3 Variance estimation in the EM algorithm

To protect against possibly invalid assumptions of the independence structures among observed states, the sandwich-type robust variance estimation (Huber, 1967; White, 1982; Royall, 1986) is suggested. White (1982) showed that if the likelihood, $\mathcal{L}(\boldsymbol{\theta})$, is constructed from a misspecified model, then $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}^*$ almost surely, where $\boldsymbol{\theta}^*$ is the root of the expectation of $\partial \log \mathcal{L}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ taken with respect to the true distribution. The asymptotic normality of $\hat{\boldsymbol{\theta}}$ from

a misspecified model is

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{*}\right)\xrightarrow{d}\mathbf{N}\left[\mathbf{0},\mathbf{C}\left(\boldsymbol{\theta}^{*}\right)\right],$$

where

$$\mathbf{C}\left(\boldsymbol{\theta}\right)=\mathbf{A}^{-1}\left(\boldsymbol{\theta}\right)\mathbf{B}\left(\boldsymbol{\theta}\right)\mathbf{A}^{-1}\left(\boldsymbol{\theta}\right)$$

with

$$\mathbf{A}\left(\boldsymbol{\theta}\right)=E\left[\frac{\partial^{2}\log\mathcal{L}\left(\boldsymbol{\theta};\mathbf{S}^{*}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\right]\quad\text{and}\quad\mathbf{B}\left(\boldsymbol{\theta}\right)=E\left\{\left[\frac{\partial\log\mathcal{L}\left(\boldsymbol{\theta};\mathbf{S}^{*}\right)}{\partial\boldsymbol{\theta}}\right]^{\otimes2}\right\}.$$

The matrix $\mathbf{C}\left(\boldsymbol{\theta}\right)$ evaluated at $\boldsymbol{\theta}^{*}$ can be estimated by $\widehat{\mathbf{A}}^{-1}\left(\boldsymbol{\theta}\right)\widehat{\mathbf{B}}\left(\boldsymbol{\theta}\right)\widehat{\mathbf{A}}^{-1}\left(\boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, where

$$\begin{aligned}\widehat{\mathbf{B}}\left(\boldsymbol{\theta}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial\log\mathcal{L}_{i}\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right]^{\otimes2}=\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial Q_{i}\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right)}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}'=\boldsymbol{\theta}}\right\}^{\otimes2};\\ \widehat{\mathbf{A}}\left(\boldsymbol{\theta}\right) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^{2}\log\mathcal{L}_{i}\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}.\end{aligned}$$

Let $\mathbf{s}\left(\boldsymbol{\theta}\right)=\left[\partial Q\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right)/\partial\boldsymbol{\theta}\right]_{\boldsymbol{\theta}'=\boldsymbol{\theta}}$. The Hessian of the log-likelihood can be obtained from the first-order Richardson extrapolation (e.g. Press *et al.*, 2007, Section 17.3) of the central difference for the gradient of the expected complete data log-likelihood. Then, the jth column of the Hessian matrix is given by

$$\frac{\mathbf{s}\left(\boldsymbol{\theta}-2h\mathbf{u}_{j}\right)-8\,\mathbf{s}\left(\boldsymbol{\theta}-h\mathbf{u}_{j}\right)+8\,\mathbf{s}\left(\boldsymbol{\theta}+h\mathbf{u}_{j}\right)-\mathbf{s}\left(\boldsymbol{\theta}+2h\mathbf{u}_{j}\right)}{12h},$$

where $\mathbf{u}_{j}$ is the jth co-ordinate vector with the jth element equal to 1 and others equal to 0 and h is a small positive value.

3.4 Application to a smoking prevention study

In this section, we apply our proposed method to analyze the data arising from the Waterloo Smoking Prevention Project 3 (WSPP3) (Cameron *et al.*, 1999). This project is a seven-year longitudinal study designed to investigate smoking behaviour among school children. A total of 100 schools in seven Ontario school boards was randomized to receive either the regular health education program provided by the school, or one of four intensive anti-smoking programs delivered by either a specially trained teacher or a public health nurse. Questionnaires regarding tobacco use and school policies and programs were completed annually from grade 6 to grade 12.

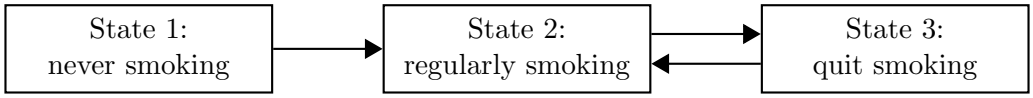

Figure 3.1: Three-state mover-stayer model for the smoking prevention study

Three states are defined to model children's behaviour. Children who have never smoked are classified 'non-smokers', and are represented by state 1. Children who are either 'regular smokers' or are experimenting with smoking are classified in state 2, and children who have smoked but are currently not smoking are classified in state 3. The model is represented by the three-state diagram in Figure 3.1.

Along with the responses, the risk factors that may influence children's smoking behaviour include gender (0–female, 1–male), treatment status (0–control; 1–intervention), social models risk score (0–none of parents, siblings or friends smoke; 1–one or more of parents, siblings or friends smoke). We consider the log-linear model for the transition intensities to incorporate three risk factors $\mathbf{x}_i$,

$$\log(q_{ij}) = \beta_{j0} + \boldsymbol{\beta}_{jx}^{\mathsf{T}}\mathbf{x}_i, \qquad j = 1, 2, 3.$$

The forward selection procedure is implemented to identify the prognostic variables on the transitions.

The data set contains 5,200 subjects who have at least two observations. The total number of state observations is 29,976. There are 1,774 subjects (34.16%) who stayed in state 1 during the study, 503 subjects (9.67%) who stayed in state 2 during the study, and 82 subjects (1.58%) who stayed in state 3 during the study. This motivates us to incorporate the mover-stayer feature in the model. We consider that the subjects starting from state 1 may be a mover with probability ω_{i1}, or a stayer with the complimentary probability. The subjects starting from state 2 or 3 are assumed to be a mover. We use the logistic model for the mover-stayer probability in state 1

$$\log\left(\frac{\omega_{i1}}{1-\omega_{i1}}\right) = \gamma_0 + \boldsymbol{\gamma}_x^{\mathsf{T}}\mathbf{x}_i,$$

where the covariate vector $\mathbf{x}_i$ contains three risk factors: the gender, treatment status, and social models risk score. We utilize the forward selection of risk factors in the logistic model for the mover-stayer distribution.

On the other hand, the self-reported smoking states are subject to misclassification. Table 3.1 presents the frequencies of transitions between observed states at consecutive pairs of times. Note that both the number of transitions from state 2 to state 1 and that of transitions from state 3 to 1 are zero. It implies that the chance of state 2 or 3 being reported as state 1 is negligible. Therefore, we assume that the regular smokers in state 2 may report the smoking status as 'quit smoking in state 3, and the subjects who have quitted smoking in state 3, and the non-smokers in state 1 will honestly report the smoking status. We consider the constant misclassification probabilities, which are assumed to be independent of covariates. The parameter related to

misclassification are reparameterized by the logit transformation,

$$\pi_{23} = \Pr\left(S_{ij}^{*} = 3 \mid S_{ij} = 2\right) = \frac{\exp\left(\alpha_{23}\right)}{1 + \exp\left(\alpha_{23}\right)},$$

where the time t is omitted due to time-independence of the misclassification probability.

Table 3.1: Frequencies of transitions between smoking states

Previous state	Smoking status	Frequencies of transitions to the following states:		
		1	2	3
1	never smoking	12761	1813	985
2	regularly smoking	0	4751	1089
3	quit smoking	0	1383	1994

Table 3.2 presents analysis results for the WSPP3 data. The results show that the gender has significant effects on the transition from never smoking to regularly smoking, but the gender, treatment status and social models risk score have no significant effects on the other transitions or the mover-stayer probabilities. In particular, males have significantly lower transition intensities out of state 1 ($p = 0.038$); there is some evidence that female children are more likely to smoke than male children ($\hat{\beta}_{1x} = -0.082$). The estimate of the proportion of stayers in state 1 is $1/\left\{1 + \exp\left(\hat{\gamma}_0\right)\right\} \approx 2.19\%$. By the delta method, a 95% confidence interval of the 'stayer' probability in state 1 is $(0.00\%, 9.26\%)$, which implies that there may not be essential difference between the inference drawn from a mover-stayer model with misclassification. The estimate of the misclassification probability of state 2 being observed as 3 is $\exp\left(\hat{\alpha}_{23}\right) / \left\{1 + \exp\left(\hat{\alpha}_{23}\right)\right\} \approx 16.56\%$, and the resulting 95% confidence interval is $(0.144, 0.187)$. This appears to be some evidence that state 2 could be mis-reported as state 3, suggesting that there would be significant difference between the inferences drawn from the mover-stayer model with misclassification and an ordinary

mover-stayer model without misclassification.

Table 3.2: Estimates of `gender` effects under the three-state HMSM for the smoking prevention study

	Covariates		EST	ASE	p-value	95% CI
Transition						
$1 \to 2$	Intercept	β_{10}	-1.581	.068	$< .001$	$(-1.715, -1.447)$
	Gender	β_{1x}	-0.082	.039	.038	$(-0.159, -0.005)$
$2 \to 3$	Intercept	β_{20}	-1.913	.074	$< .001$	$(-2.059, -1.767)$
	Gender	β_{2x}	-0.031	.097	.752	$(-0.220, -0.159)$
$3 \to 2$	Intercept	β_{30}	-0.834	.076	$< .001$	$(-0.983, -0.685)$
	Gender	β_{3x}	-0.130	.106	.220	$(-0.338, +0.078)$
Mover-stayer						
State 1	Intercept	γ_0	3.800	1.688	.024	$(+0.492, +7.108)$
Misclassification						
$2 \mapsto 3$	Intercept	α_{23}	-1.617	.044	$< .001$	$(-1.703, -1.531)$

3.5 Simulation studies

In this section, simulations studies are conducted to evaluate the performance of proposed MLEs, as opposed to the naive MLEs, which are obtained with misclassfication ignored. We consider the three-state mover-stayer model with misclassification in Figure 3.1, which may, for example, represents children's smoking behaviour in Section 3.4. The naive MLEs are obtained by ignoring the misclassification among states. In naive analyses, we replace the previous state, state 2, by the lower state, state 1, to adjust the misclassification in the transitions from state 2 to 1.

3.5.1 Simulation setting

The number of subjects is $n = 5000$ and a total of 1000 replicates are used. Each subject is assumed to start from state k with the initial state occupying probabilities

$$\{\Pr(S_{i0} = k), k = 1, 2, 3\} = (0.95, 0.05, 0.05)$$

at the initial time $t_{i0} = 0$. For the mover-stayer distribution, we assume that the subject is a mover if starting from state 2 or 3, i.e. $\omega_{i2} = \omega_{i3} = 1$; if the subject starts from state 1, the probability of being a mover is generated from the logistic model (3.3),

$$\omega_{i1} = \frac{\exp(\gamma_0 + \gamma_x x_i)}{1 + \exp(\gamma_0 + \gamma_x x_i)},$$

where x_i is a Bernoulli variable with probability 0.5. We consider two scenarios:

- $\gamma_0 = 1.0$ and $\gamma_x = 0.2$ such that the proportion of movers among the subjects who start from state 1 is 74.97%;
- $\gamma_0 = 2.5$ and $\gamma_x = 1.2$ such that the proportion of movers among the subjects who start from state 1 is 95.00%.

The effects of $\boldsymbol{\gamma}$ on the mover-stayer distribution are described in Section 3.7.2.

If the subject is a stayer, then the subject stays in state 1 during the study; otherwise, the subject follows the three state time-homogeneous Markov process. The sojourn time in state j follows an exponential distribution with mean $1/q_{jk}(x_i)$, where $q_{jk}(x_i) = \exp(\beta_{j0} + \beta_{jx} x_i)$, $j = 1, 2, 3$. After staying in state j, the subject enters the next state as shown in Figure 3.1. We set $\beta_{10} = -1.0$, $\beta_{1x} = -0.5$, $\beta_{20} = -0.7$, $\beta_{2x} = 0.6$, $\beta_{30} = -0.8$, and $\beta_{3x} = -0.4$. The effects of

$\boldsymbol{\beta}$ on transitions are summerized at Table 3.4. To mimic the data from the Waterloo Smoking Prevention Project, we assume that the observation is carried out each year, but each observation is subject to missingness with probability 0.15; the number of observations for each subject is between two and seven.

For the misclassification process, we assume no misclassification in the initial state or state 1, i.e., $\pi_{s_{i0},s_{i0}}(0) = 1$, and $\pi_{11}(t) = 1$, $t = 1, 2, \ldots, 6$. If the underlying state is 2, we assume it can be misclassified as state 1 or 3 with probabilities

$$\pi_{21}(t) = \frac{\exp(\alpha_{21})}{1 + \exp(\alpha_{21}) + \exp(\alpha_{23})} \quad \text{and} \quad \pi_{23}(t) = \frac{\exp(\alpha_{23})}{1 + \exp(\alpha_{21}) + \exp(\alpha_{23})};$$

if the underlying state is 3, we assume that it can be misclassified as state 1 with probability $\pi_{31}(t) = \exp(\alpha_{31}) / \{1 + \exp(\alpha_{31})\}$. We set $\alpha_{21} = -3.0$, $\alpha_{23} = -2.0$, and $\alpha_{31} = -2.5$ such that 15.62% of state 2 is misclassified and 7.59% of state 3 is misclassified.

3.5.2 Simulation results

We analyze the simulated data using both the naive method and the proposed method. The naive method ignores the feature of misclassification, whereas the proposed method accounts for the misclassification effects. Table 3.3 summarizes the averages of biases of point estimates and their asymptotic and empirical standard errors (ASEs and ESEs), along with coverage rates (CRs) of corresponding 95% confidence intervals.

The results show that the proposed method performs well in finite samples, and illustrate the significant biases and the low coverage rates produced by the naive method. The biases in the proposed MLEs of regression coefficients $\boldsymbol{\beta}$ are relatively small and negligible; the associated ASEs agree well with their empirical counterparts; the resulting coverage rates are close to the

nominal level. For the misclassification parameters $\boldsymbol{\alpha}$, the performance of the MLE of α_{31} is very good while the estimates of α_{21} and α_{23} have slightly larger biases. The ASEs for the MLE of $\boldsymbol{\alpha}$ are slightly overestimated, resulting in coverage rates slightly higher than the nominal level. The biases in the MLEs of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and their SE estimates tend to decrease, as the size of movers increases. For the mover-stayer parameters $\boldsymbol{\gamma}$, as the size of stayers increases, the biases in the proposed MLEs tend to decrease, the ASEs and ESEs become closer, and the coverage rates become closer to the nominal level. However, in the case of 5% stayers, the biases of estimating γ_x by the naive and the proposed methods are large, which may cause relatively large biases in the MLEs of β_{1x}. This is not surprising because estimation of γ_x is more difficult with a smaller sample size.

Table 3.3: Simulation results for the three-state hidden mover-stayer model

	True States				Naive Method				MLE			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
					25% stayers and 75% movers							
β_{10}	.000	.048	.049	93.9	.261	.033	.035	0.0	.002	.050	.049	94.2
β_{1x}	−.001	.093	.094	95.0	.250	.052	.054	0.0	−.003	.096	.095	94.8
β_{20}	−.001	.032	.032	94.9	.455	.034	.037	0.0	−.003	.096	.095	94.8
β_{2x}	.000	.047	.047	94.7	−.133	.050	.056	27.4	.006	.068	.067	95.7
β_{30}	−.001	.044	.043	95.7	.257	.046	.050	0.1	−.003	.065	.062	95.9
β_{3x}	.000	.068	.068	95.5	−.060	.075	.080	85.3	.002	.087	.086	95.1
γ_0	.003	.078	.079	95.8	−.390	.050	.050	0.0	−.001	.080	.078	95.7
γ_x	.021	.212	.208	96.2	−.442	.074	.073	0.0	.023	.217	.206	96.3
α_{21}	–	–	–	–	–	–	–	–	−.014	.175	.170	96.7
α_{23}	–	–	–	–	–	–	–	–	−.009	.200	.189	96.1
α_{31}	–	–	–	–	–	–	–	–	−.002	.076	.073	95.9
					5% stayers and 95% movers							
β_{10}	−.001	.043	.041	96.0	.280	.029	.028	0.0	−.001	.044	.042	96.4
β_{1x}	.036	.081	.065	95.6	.242	.045	.044	0.0	.048	.082	.062	94.9
β_{20}	.001	.029	.030	94.8	.459	.031	.034	0.0	−.001	.079	.079	95.4
β_{2x}	−.001	.042	.041	95.0	−.137	.045	.049	16.6	−.000	.062	.062	94.7
β_{30}	.001	.040	.040	94.9	.266	.042	.045	0.0	−.001	.061	.061	95.3
β_{3x}	−.000	.062	.063	94.2	-.063	.068	.073	82.2	−.001	.080	.080	94.5
γ_0	.023	.218	.203	96.3	−1.008	.067	.065	0.0	.026	.226	.213	96.0
γ_x	−.436	1.218	.625	83.9	−1.601	.095	.099	0.0	−.633	.951	.485	81.8
α_{21}	–	–	–	–	–	–	–	–	−.013	.159	.156	96.0
α_{23}	–	–	–	–	–	–	–	–	−.009	.180	.177	95.0
α_{31}	–	–	–	–	–	–	–	–	−.000	.069	.068	95.6

5000 subjects, 1000 replicates

$\alpha_{21} = -3.0$, $\alpha_{23} = -2.0$, $\alpha_{31} = -2.5$

$\beta_{10} = -1.0$, $\beta_{1x} = -0.5$, $\beta_{20} = -0.7$, $\beta_{2x} = 0.6$, $\beta_{30} = -0.8$, $\beta_{3x} = -0.4$

3.6 Discussion

In this chapter, we propose continuous-time mover-stayer models with misclassification. These models provide a useful framework to simultaneously feature a special type of heterogeneity and account for state misclassification. The underlying mover-stayer model is a mixture of two independent continuous-time Markov processes: one with the identity matrix as the transition probability matrix at any time, and the other with an unspecified transition intensity matrix. This mover-stayer model describes the heterogeneity of the data arising from the existence of two sub-populations, the stayers that stay in the initial state, and the movers that evolve according to a Markov process. The proposed misclassification model is not limited to the scenario with misclassified states; it can be applied to the case with discrete-valued surrogates observed for underlying states, such as the multiple sclerosis/magnetic resonance imaging lesion count data (Altman and Petkau, 2005). Therefore, our proposed models can be viewed as hidden mover-stayer models with discrete observation, which is an extension of hidden Markov models (Zucchini and MacDonald, 2009).

We developed an EM algorithm to obtain maximum likelihood estimates in the analysis of panel data under our proposed models. The forward-backward procedure (Baum and Eagon, 1967; Baum and Sell, 1968) is widely used in the estimation based on the EM algorithm for hidden Markov models due to its efficiency (e.g. Zucchini and MacDonald, 2009, Chapter 4). Therefore, this procedure is also adopted in our E-step to calculate the conditional probabilities. However, special attention on the numerical underflow in forward and backward probabilities is required in the implementation of the forward-backward procedure, especially when the number of assessments is large (Rabiner, 1989). Our simulation studies show that the proposed method performs well in finite samples and illustrate the consequence of ignoring the misclassfication in the naive analyses. The simulation results also indicate that the effect of the risk factor on the

mover-stayer distribution is difficult to estimate for small sample size and may cause considerable biases for estimation of parameters in the transition intensity model.

In the analysis of the WSPP3 data, the probabilities of being a stayer or mover are described by the logistic model. It may be interesting to use the idea of O'Keeffe *et al.* (2013) to introduce a random effect to the mover-stayer distribution in order to account for unobserved heterogeneity among different schools. After data fitting, it is intuitive to examine which model is the most appropriate for these data. The likelihood ratio tests can be utilized to choose our proposed model versus an ordinary mover-stayer model by testing $\mathrm{H}_0 \colon \pi_{23} = 0$ versus $\mathrm{H}_a \colon \pi_{23} > 0$, and our proposed model versus an ordinary Markov model with misclassification by testing $\mathrm{H}_0 \colon \omega_1 = 1$ versus $\mathrm{H}_a \colon \omega_1 < 1$. The asymptotic distribution of the likelihood ratio statistic may not be the chi-square distribution with one degree of freedom in either case, because the parameter value under H_0 is a boundary point of the parameter space. The investigation of this problem dates back to Chernoff (1954), who studied the hypothesis test with the null on the boundary for a multivariate normal distribution. This problem was also reported by Frydman and Kadam (2004) for testing the continuous-time mover-stayer model versus an ordinary Markov model. The study of the asymptotic distribution of the likelihood test statistic under H_0 will be pursued in the future work for our proposed models.

3.7 Technical details

3.7.1 Transition probability matrix for the three-state Markov model

The transition probability matrix for the three-state Markov model in Figure 3.1 can be calculated analytically from the transition intensity matrix

$$\begin{pmatrix} -q & q & 0 \\ 0 & -u & u \\ 0 & v & -v \end{pmatrix},$$

by the function `MatrixExp` in `Mathematica`. The corresponding transition probabilities for gap time t are given by

$$\begin{aligned}
P_{11}(t) &= \exp(-qt); \\
P_{12}(t) &= -\frac{\{1-\exp(-qt)\}v}{q-u-v} \\
&\quad + \frac{q\Big[u[\exp\{-(u+v)t\}-\exp(-qt)]+v\{1-\exp(-qt)\}\Big]}{(q-u-v)(u+v)}; \\
P_{13}(t) &= u\left[\frac{q[1-\exp\{-(u+v)t\}]}{(u+v)(q-u-v)} - \frac{1-\exp(-qt)}{q-u-v}\right]; \\
P_{22}(t) &= \frac{v+\exp\{-(u+v)t\}u}{u+v}; \\
P_{23}(t) &= \frac{u-\exp\{-(u+v)t\}u}{u+v} = 1-P_{22}(t); \\
P_{32}(t) &= \frac{v}{u+v} - \frac{\exp\{-(u+v)t\}v}{u+v}; \\
P_{33}(t) &= \frac{u}{u+v} + \frac{\exp\{-(u+v)t\}v}{u+v};
\end{aligned}$$

$$P_{21}(t) = P_{31}(t) = 0.$$

3.7.2 Effects of parameters in simulation studies

Transition intensity model

In the three-state Markov model in Figure 3.1, the sojourn time τ_i is exponentially distributed with mean $1/q_i$, $i = 1, 2$, where q is the transition intensity modelled by (3.2).

The prognostic covariate X is simulated from the Binomial distribution with probability p. Then, the mean sojourn time is

$$\begin{aligned} E(\tau_i) &= E\left[1/\exp\left(\beta_{i0} + \beta_{ix}X\right)\right] \\ &= (1-p)\exp\left(-\beta_{i0}\right) + p\exp\left(-\beta_{i0} - \beta_{ix}\right). \end{aligned}$$

The median sojourn time is $M(\tau_i) = \log(2) \cdot E(\tau_i)$.

If we set $\boldsymbol{\beta} = (-1.0, -0.5, -0.7, 0.6, -0.8, -0.4)^{\mathsf{T}}$ and $X \sim \text{BIN}(0.5)$, then the mean and median sojourn times from state 1 to 2 are 3.60 and 2.50, the mean and median sojourn times from state 2 to 3 are 1.56 and 1.08, and the mean and median sojourn times from state 3 to 2 are 2.77 and 1.29.

Table 3.4: Parameter effects on transitions

Transition	β_{i0}	β_{ix}	$E(\tau_i)$	$M(\tau_i)$
$1 \to 2$	-1.0	-0.5	3.60	2.50
$2 \to 3$	-0.7	0.6	1.56	1.08
$3 \to 2$	-0.8	-0.4	2.77	1.29

Logistic model for the mover-stayer distribution

The stayer probability is given by

$$\Pr\left(Z_i = 0 \mid S_{i0} = 1, x_i; \boldsymbol{\gamma}\right) = \frac{1}{1 + \exp\left(\gamma_0 + \gamma_x x_i\right)}.$$

If the prognostic covariate X is simulated from the Binomial distribution with probability p, then the mean stayer probability is

$$E\left\{\Pr\left(Z_i = 0 \mid S_{i0} = 1, x_i; \boldsymbol{\gamma}\right)\right\} = \frac{1-p}{1 + \exp\left(\gamma_0\right)} + \frac{p}{1 + \exp\left(\gamma_0 + \gamma_x\right)}.$$

If we set $p = 0.5$, $\gamma_0 = 1.0$, and $\gamma_x = 0.2$, the proportional of stayers among the subjects who start from state 1 is 25.02%; if we set $p = 0.5$, $\gamma_0 = 2.5$, and $\gamma_x = 1.2$, the proportional of stayers among the subjects who start from state 1 is 5.00%.

Chapter 4

Analysis of panel data with misclassified discrete covariates

4.1 Introduction

It is not uncommon that discrete covariates are misclassified. In the case with binary outcomes, various approaches have been proposed for the covariate misclassification problem. The investigation of the impact of misclassification on analysis and interpretation dates back to Bross (1954). An overview of the development was given by Kuha *et al.* (2005), who described the effects of misclassification and summarized the methods for adjusting misclassification effects. Given that misclassification parameters are estimated from validation studies or repeated measurements, consistent estimates of the relative risk and related parameters can be obtained from the matrix method (Bross, 1954; Marshall, 1990; Morrissey and Spiegelman, 1999) or the maximum likelihood method (Espeland and Hui, 1987; Spiegelman *et al.*, 2000). In addition, the regression calibration approach (Carroll and Stefanski, 1990; Gleser, 1990) provides a simple and

convenient way to reduce measurement error effects, in which the unobserved covariate is replaced by its estimated value and then the standard analysis method can be performed. With internal validation data, Spiegelman *et al.* (2001) proposed an efficient regression calibration method for logistic regression, which combines the estimates from the regression calibration and the estimates from the internal validation study using the true covariate value by a generalized inverse-variance weighted average.

The simulation extrapolation (SIMEX) method (Cook and Stefanski, 1994; Stefanski and Cook, 1995) is another useful approach to deal with measurement error problems. Recently, Küchenhoff *et al.* (2006) developed the misclassification SIMEX method for parameter estimation in the presence of misclassification in discrete covariates or responses in regression models and Küchenhoff *et al.* (2007) derived the asymptotic variance estimation for the misclassification SIMEX approach. However, the regression calibration is an approximate method, and the SIMEX method relies on an extrapolation scheme that is uncertain. Therefore, the consistency of estimators can not be guaranteed by either method.

There has been relatively less attention paid to the covariate misclassification in the analysis of panel data. For the three-state progressive Markov model, White (2007, Chapter 3) investigated the impact of misclassification of a binary covariate and proposed the correction methods based on the likelihood and the SIMEX methods. However, transition intensities are required to remain constant through time in the time-homogeneous Markov model. In practice, transition intensities may be time-dependent. A common approach of fitting time-dependent models to panel data is to use Markov models with piecewise constant intensities. This idea dates back to Faddy (1976) and was suggested by Kalbfleisch and Lawless (1985). The discussion of Markov models with piecewise constant intensities includes Lindsey and Ryan (1993), Gentleman *et al.* (1994), Chen and Sen (1999), Hsieh *et al.* (2002), Saint-Pierre *et al.* (2003), van den Hout and Matthews (2009), Chen

et al. (2010), and Tom and Farewell (2011). One issue of fitting models with piecewise constant intensities is to determine cut points properly. Various criteria for the choice of change points have been presented. For example, clinical reasons (Sharples *et al.*, 2001; Alioum *et al.*, 2005), the change in the trend of transition intensities (Pérez-Ocón *et al.*, 2001), division of the number of data points into equal groups (Kay, 1986), selection of the only change point based on the likelihood value (Mathieu *et al.*, 2005), or the merging algorithm proposed by Ocañ-Riola (2005).

In this chapter, we present the maximum likelihood estimation procedure to analyze the panel data under Markov assumption with misclassified discrete covariates. To highlight the idea, the discussion is directed to binary covariates where extensions to accommodating discrete covariates are straightforward. In Section 4.2, the Markov models with piecewise constant intensities are described to account for the time-inhomogeneity. In Section 4.3, we show that the Markov models with misclassified binary covariates are not identifiable. The maximum likelihood estimation procedures are developed in Section 4.4, where two scenarios, known reclassification probabilities and main study/validation study design, are considered. Simulation studies are conducted in Section 4.5 to demonstrate the performance of the proposed method. Data arising from the psoriatic arthritic (PsA) study are analyzed using the proposed methods in Section 4.6. A general discussion is given in Section 4.7, and technical notes are presented in Section 4.8.

4.2 Model formulation

4.2.1 Piecewise constant Markov models

Suppose an individual moves among K states, denoted by integers $1, 2, \ldots, K$. Let $S(t)$ denote the true state at time t occupied by an individual. The process $\{S(t), t \geq 0\}$ is assumed to follow a continuous-time Markov process. Let $\mathbf{P}(s, s+t)$ be the $K \times K$ transition probability matrix

with (i, j) entry

$$P_{ij}(s, s+t) = \Pr[S(s+t) = j \mid S(s) = i]$$

for $s \geq 0$, $t > 0$, $i, j = 1, 2, \ldots, K$. The transition intensity from state i to j at time t is

$$q_{ij}(t) = \lim_{\Delta t \downarrow 0} \frac{P_{ij}(t, t+\Delta t)}{\Delta t}, \qquad i \neq j,$$

and as a convention, define

$$q_{ii}(t) = -\sum_{j \neq i} q_{ij}(t).$$

Let $\mathbf{Q}(t)$ be the $K \times K$ transition intensity matrix with (i, j) entry $q_{ij}(t)$, $i, j = 1, 2, \ldots, K$. In the piecewise constant framework, a sequence of times $0 = b_0 < b_1 < \cdots < b_M < b_{M+1} = \infty$ is pre-specified and transition intensities are assumed to be constant within each interval $B_k = (b_k, b_{k+1}]$, where $k = 0, \ldots, M$. That is,

$$q_{ij}(t) = q_{ijk}, \qquad \text{if } t \in B_k.$$

Then, the transition intensity matrix $\mathbf{Q}(t)$ is written as

$$\mathbf{Q}(t) = \mathbf{Q}_k, \qquad \text{for } t \in B_k.$$

Therefore, transition intensities can be defined as

$$q_{ij}(t) = \prod_{k=0}^{M} q_{ijk}^{I(t \in B_k)} = \sum_{k=0}^{M} q_{ijk} I(t \in B_k),$$

where $I(\cdot)$ is the indicator function and $k = 0, \ldots, M$. Models with piecewise constant intensities are weakly parametric and provide flexible estimation of transition intensities and robust estimation of regression coefficients, though other specification such as splines (e.g. He and Lawless,

2003; Titman, 2011) can also be used for smooth estimates of intensities.

By the result of the time-homogeneous Markov model (e.g. Cox and Miller, 1965, Chapter 4), transition probabilities between two time points s and $s+t$ within the interval of constant transition intensities, where $b_k < s \leq s+t \leq b_{k+1}$, are given by

$$\mathbf{P}(s, s+t) = \exp(\mathbf{Q}_k t).$$

If time points s and $s+t$ are in different intervals such that $b_i < s \leq b_{i+1}$ and $b_j < s+t \leq b_{j+1}$, then by the Chapman-Kolmogorov equation, transition probabilities are given by

$$\mathbf{P}(s, s+t) = \mathbf{P}(s, b_{i+1}) \prod_{k=i+1}^{j-1} [\mathbf{P}(b_k, b_{k+1})] \mathbf{P}(b_j, s+t),$$

where $0 \leq i < j \leq M$ and $\mathbf{P}(s, s) = \mathbf{I}$.

Each transition probability matrix within the time interval of constant transition intensities can be calculated by the closed-form expression for the simple models (Tuma *et al.*, 1979; Chiang, 1980; Longini *et al.*, 1989; Omar *et al.*, 1995; Satten, 1999; Jackson, 2011) or the decomposition methods of evaluating the matrix exponential (Cox and Miller, 1965; Kalbfleisch and Lawless, 1985; Moler and van Loan, 2003; Jackson, 2011) discussed in Section 1.1.1.

4.2.2 Regression model for covariates

Transition intensity model

Let X be a discrete error-free time-independent covariate with two numerical levels x_1 and x_2 and $\mathbf{z}$ be a $p \times 1$ vector of perfectly measured time-independent covariates. To model the effects

of covariates on transitions, we consider the multiplicative intensity model

$$q_{ij}\left(t \mid X, \mathbf{z}\right) = q_{ij0}\left(t\right) \exp\left(X\beta_{ijx} + \mathbf{z}^{\mathsf{T}}\boldsymbol{\beta}_{ijz}\right), \qquad i \neq j,\, i, j = 1, \ldots, K, \tag{4.1}$$

where $q_{ij0}\left(t\right)$ is the baseline transition intensity out of state i to j at time t, and $\left(\beta_{ijx}, \beta_{ijz_1}, \ldots, \beta_{ijz_p}\right)^{\mathsf{T}}$ are vectors of regression coefficients of primary interest. For Markov models with piecewise constant intensities, the baseline transition intensities can be re-parameterized as

$$q_{ij0}\left(t\right) = \exp\left(\beta_{ijk0}\right), \qquad \text{for } t \in B_k = \left(b_k, b_{k+1}\right), k = 0, \ldots, M,$$

where β_{ijk0} is a parameter.

Reclassification model

Let X^* be a surrogate measure of X, which takes the same range of possible values as X. We assume that misclassification is non-differential, that is, the observed measurement is independent of the outcome given the true measurement. Let

$$\lambda_{ij}\left(\mathbf{z}\right) = \Pr\left(X = x_j \mid X^* = x_i, \mathbf{z}\right), \qquad i \neq j,\, i, j = 1, 2$$

be the reclassification probability of the true covariate given the observed surrogate; this can be regarded as a discrete version of the Berkson model (Berkson, 1950). The logistic models

$$\log\left[\frac{\lambda_{12}\left(\mathbf{z}\right)}{1 - \lambda_{12}\left(\mathbf{z}\right)}\right] = \alpha_{10} + \mathbf{z}^{\mathsf{T}}\boldsymbol{\alpha}_{1z} \quad \text{and} \quad \log\left[\frac{\lambda_{21}\left(\mathbf{z}\right)}{1 - \lambda_{21}\left(\mathbf{z}\right)}\right] = \alpha_{20} + \mathbf{z}^{\mathsf{T}}\boldsymbol{\alpha}_{2z} \tag{4.2}$$

can be used to model the effects of covariates on reclassification probabilities, where α_{i0} and $\boldsymbol{\alpha}_{iz} = \left(\alpha_{i1}, \ldots, \alpha_{ip}\right)^{\mathsf{T}}$ are regression coefficients and $i = 1, 2$.

Therefore, reclassification probabilities are given by

$$\lambda_{12}(\mathbf{z}) = \frac{\exp\left(\alpha_{10} + \mathbf{z}^{\mathsf{T}}\boldsymbol{\alpha}_{1z}\right)}{1 + \exp\left(\alpha_{10} + \mathbf{z}^{\mathsf{T}}\boldsymbol{\alpha}_{1z}\right)}, \tag{4.3}$$

$$\lambda_{21}(\mathbf{z}) = \frac{\exp\left(\alpha_{20} + \mathbf{z}^{\mathsf{T}}\boldsymbol{\alpha}_{2z}\right)}{1 + \exp\left(\alpha_{20} + \mathbf{z}^{\mathsf{T}}\boldsymbol{\alpha}_{2z}\right)}, \tag{4.4}$$

$$\lambda_{11}(\mathbf{z}) = 1 - \lambda_{12}(\mathbf{z}), \tag{4.5}$$

$$\text{and} \quad \lambda_{22}(\mathbf{z}) = 1 - \lambda_{21}(\mathbf{z}). \tag{4.6}$$

4.3 Model identifiability

In this section, we show that the joint model for the state process and the reclassification process is not identifiable.

Theorem 4.3.1. *Consider a time-homogenous Markov process* $\{S(t) : t \geq 0\}$ *with the constant transition intensity matrix* $\mathbf{Q}$*. Suppose* $\mathbf{Q}$ *is modelled by* (4.1)*,* $\mathbf{z}$ *is a vector of perfectly measured time-independent covariates, and* X *is subject to misclassification with* X^* *being a surrogate measurement. Let* $\mathbf{S} = (S_1, \ldots, S_m)$ *denote the states of the process* $\{S(t) : t \geq 0\}$ *observed at time points* $0 \leq t_1 < t_2 < \cdots < t_m$*. Suppose the reclassification model is given by* (4.2)*. Let* $\boldsymbol{\alpha} = \left(\alpha_{i0}, \boldsymbol{\alpha}_{iz}^{\mathsf{T}} : i = 1, 2\right)^{\mathsf{T}}$ *and*

$$\boldsymbol{\beta} = \left(\beta_{ij0}, \beta_{ijx}, \beta_{ijz_1}, \ldots, \beta_{ijz_p} : i \neq j, i, j = 1, \ldots, K\right)^{\mathsf{T}}$$

be any given parameters associated with (4.2) *and* (4.1)*, respectively. Assume that the distribution of the initial state,* $\Pr(S_1 \mid X^*, \mathbf{z})$*, is free of* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$*. Define*

$$\boldsymbol{\alpha}^* = -\boldsymbol{\alpha} \quad \text{and} \quad \boldsymbol{\beta}^* = (\beta^*_{ij0}, \beta^*_{ijx}, \beta^*_{ijz_1}, \ldots, \beta^*_{ijz_p} : i \neq j, i, j = 1, \ldots, K)^{\mathsf{T}},$$

with

$$\beta^*_{ij0} = \beta_{ij0} + \beta_{ijx}\left(x_1 + x_2\right), \quad \beta^*_{ijx} = -\beta_{ijx}, \quad \text{and} \quad \boldsymbol{\beta}^*_{ijz} = \boldsymbol{\beta}_{ijz}.$$

Then, we have the probability identity:

$$\Pr\left(\mathbf{S} \mid X^*, \mathbf{z}; \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*\right) = \Pr\left(\mathbf{S} \mid X^*, \mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}\right).$$

Proof. If $\boldsymbol{\alpha}^* = -\boldsymbol{\alpha}$, then

$$\alpha_{i0} + \mathbf{z}^\mathsf{T}\boldsymbol{\alpha}_{iz} + \alpha^*_{i0} + \mathbf{z}^\mathsf{T}\boldsymbol{\alpha}^*_{iz} = 0, \qquad i = 1, 2,$$

which is equivalent to

$$\frac{\exp\left(\alpha_{i0} + \mathbf{z}^\mathsf{T}\boldsymbol{\alpha}_{iz}\right)}{1 + \exp\left(\alpha_{i0} + \mathbf{z}^\mathsf{T}\boldsymbol{\alpha}_{iz}\right)} = \frac{1}{1 + \exp\left(\alpha^*_{i0} + \mathbf{z}^\mathsf{T}\boldsymbol{\alpha}^*_{iz}\right)}, \qquad i = 1, 2.$$

By (4.3)–(4.6), we have

$$\Pr\left(X = x_1 \mid X^*, \mathbf{z}; \boldsymbol{\alpha}\right) = \Pr\left(X = x_2 \mid X^*, \mathbf{z}; \boldsymbol{\alpha}^*\right). \tag{4.7}$$

By the condition $\beta^*_{ij0} = \beta_{ij0} + \beta_{ijx}\left(x_1 + x_2\right)$, $\beta^*_{ijx} = -\beta_{ijx}$, and $\boldsymbol{\beta}^*_{ijz} = \boldsymbol{\beta}_{ijz}$, we know

$$\begin{aligned} & \beta^*_{ij0} + \beta^*_{ijx}x_2 + \mathbf{z}^\mathsf{T}\boldsymbol{\beta}^*_{ijz} &=& \beta_{ij0} + \beta_{ijx}x_1 + \mathbf{z}^\mathsf{T}\boldsymbol{\beta}_{ijz}, \\ \text{and} \qquad & \beta^*_{ij0} + \beta^*_{ijx}x_1 + \mathbf{z}^\mathsf{T}\boldsymbol{\beta}^*_{ijz} &=& \beta_{ij0} + \beta_{ijx}x_2 + \mathbf{z}^\mathsf{T}\boldsymbol{\beta}_{ijz}. \end{aligned}$$

By (4.1), we have

$$q_{ij}\left(x_1, \mathbf{z}; \boldsymbol{\beta}\right) = q_{ij}\left(x_2, \mathbf{z}; \boldsymbol{\beta}^*\right) \quad \text{and} \quad q_{ij}\left(x_2, \mathbf{z}; \boldsymbol{\beta}\right) = q_{ij}\left(x_1, \mathbf{z}; \boldsymbol{\beta}^*\right). \tag{4.8}$$

Thus, by (4.8) and the assumption that $\Pr(S_1 \mid X^*, \mathbf{z})$ does not contain $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$, we know

$$\Pr(\mathbf{S} \mid X = x_1, \mathbf{z}; \boldsymbol{\beta}) = \Pr(\mathbf{S} \mid X = x_2, \mathbf{z}; \boldsymbol{\beta}^*); \tag{4.9}$$

$$\Pr(\mathbf{S} \mid X = x_2, \mathbf{z}; \boldsymbol{\beta}) = \Pr(\mathbf{S} \mid X = x_1, \mathbf{z}; \boldsymbol{\beta}^*). \tag{4.10}$$

Note that

$$\begin{aligned}\Pr(\mathbf{S} \mid X^*, \mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \Pr(\mathbf{S}, X = x_1 \mid X^*, \mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) + \Pr(\mathbf{S}, X = x_2 \mid X^*, \mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \Pr(\mathbf{S} \mid X = x_1, \mathbf{z}; \boldsymbol{\beta}) \Pr(X = x_1 \mid X^*, \mathbf{z}; \boldsymbol{\alpha}) \\ &\quad + \Pr(\mathbf{S} \mid X = x_2, \mathbf{z}; \boldsymbol{\beta}) \Pr(X = x_2 \mid X^*, \mathbf{z}; \boldsymbol{\alpha}). \end{aligned} \tag{4.11}$$

By (4.7), (4.9), (4.10), and (4.11), we know that there exist $\boldsymbol{\alpha}^* = -\boldsymbol{\alpha}$ and $\boldsymbol{\beta}^*$ satisfying

$$\beta^*_{ij0} = \beta_{ij0} + \beta_{ijx}(x_1 + x_2), \quad \beta^*_{ijx} = -\beta_{ijx}, \quad \text{and} \quad \boldsymbol{\beta}^*_{ijz} = \boldsymbol{\beta}_{ijz},$$

such that

$$\Pr(\mathbf{S} \mid X^*, \mathbf{z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \Pr(\mathbf{S} \mid X^*, \mathbf{z}; \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*). \qquad \square$$

This theorem says that two distinct sets of parameters can lead to the same probability mass function, and thus, suggesting that the model is non-identifiable in the presence of misclassified binary covariates. Consequently, in developing valid inference methods to account for covariate misclassification effects, one needs to carefully address non-identifiability. One approach is to impose some restrictions on the parameter space to ensure the identifiability of the model. In particular, specifying reclassification probabilities to be known values can guarantee the model identifiability. On the other hand, the availability of a validation data set is also helpful to overcome the non-identifiability problem arising from the misclassification in binary covariates.

4.4 Maximum likelihood methods

Suppose that the data are obtained from n independent individuals. Let $S_i(t)$ denote the state for individual i at time $t \geq 0$, $i = 1, \ldots, n$. $\{S_i(t), t \geq 0\}$ is assumed to follow a common continuous-time Markov process for each individual. Let $t_{i0} < t_{i1} < \cdots < t_{im_i}$ denote the $(m_i + 1)$ times at which individual i is observed. For simplicity, let S_{ij} denote the state at the jth observation for individual i. Let X_i represent the unobserved true covariate, X_i^* denote the surrogate measure of X_i, and $\mathbf{z}_{i,\,p\times 1}$ be the other precisely measured time-independent covariates for individual i.

In order to estimate parameters of interest in the transition intensity model, we propose the likelihood inference methods for two practical situations: one is that the parameters in reclassification probabilities are known from empirical studies; the other is in the presence of an internal or external validation data.

4.4.1 Known reclassification probabilities

Conditional on known $\boldsymbol{\alpha}$, the likelihood function contributed from individual i is

$$\begin{aligned}
\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \Pr(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \Pr(X_i = x_1 \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}) \Pr(S_{i0} \mid X_i, \mathbf{z}_i) \prod_{j=1}^{m_i} \Pr(S_{ij} \mid S_{i,j-1}, X_i = x_1, \mathbf{z}_i; \boldsymbol{\beta}) \\
&\quad + \Pr(X_i = x_2 \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}) \Pr(S_{i0} \mid X_i, \mathbf{z}_i) \prod_{j=1}^{m_i} \Pr(S_{ij} \mid S_{i,j-1}, X_i = x_2, \mathbf{z}_i; \boldsymbol{\beta}) \\
&\propto \Pr(X_i = x_1 \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}) \prod_{j=1}^{m_i} \Pr(S_{ij} \mid S_{i,j-1}, X_i = x_1, \mathbf{z}_i; \boldsymbol{\beta}) \\
&\quad + \Pr(X_i = x_2 \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}) \prod_{j=1}^{m_i} \Pr(S_{ij} \mid S_{i,j-1}, X_i = x_2, \mathbf{z}_i; \boldsymbol{\beta}), \qquad (4.12)
\end{aligned}$$

where $\Pr(X_i = x_j \mid X_i^* = x_i, \mathbf{z}_i; \boldsymbol{\alpha})$ is the reclassification probability $\lambda_{ij}(\mathbf{z}_i)$ defined by (4.3)–(4.6), $\Pr(S_{ij} \mid S_{i,j-1}, X_i, \mathbf{z}_i; \boldsymbol{\beta})$ is the transition probability $P_{s_{i,j-1},s_{ij}}(t_{i,j-1}, t_{ij} \mid X_i, \mathbf{z}_i; \boldsymbol{\beta})$ defined in Section 4.2.1, and $\Pr(S_{i0} \mid X_i, \mathbf{z}_i)$ is the initial state occupation probability which is assumed to be independent of X_i given $\mathbf{z}_i$. The overall likelihood is the product of all the contributions

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

The maximum likelihood estimates, denoted by $\hat{\boldsymbol{\beta}}$, can be obtained by maximizing the log-likelihood with respect to $\boldsymbol{\beta}$. The gradient and Hessian of the log-likelihood function are described in Section 4.8.1. When the explicit analytic expression of transition probabilities is available, it is not difficult to obtain the analytic expression for the Hessian matrix. Then, the Newton-Raphson algorithm can be directly used to maximize the log-likelihood. On the other hand, although the second derivatives of transition probabilities obtained by the canonical decomposition are available in Kosorok and Chao (1995, 1996), they are so complex that it does not seem worth the effort to supply for the optimization. The quasi-Newton algorithm, such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method incorporating the first derivatives, is then used.

From standard likelihood theory, under regularity conditions, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ and asymptotically normally distributed:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathbf{N}\left[\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\beta})\right], \qquad \text{as } n \to \infty,$$

where $\mathcal{I}(\boldsymbol{\beta}) = E\left[-\partial^2 \log \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{S}) / \left(\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathsf{T}}\right)\right] = E\left\{\left[\partial \log \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{S}) / \partial\boldsymbol{\beta}\right]^{\otimes 2}\right\}$, $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{S})$ is the likelihood based on the states $\mathbf{S}$ of one individual, and $\mathbf{x}^{\otimes 2}$ is the out production of the column vector $\mathbf{x}$, i.e. $\mathbf{x}^{\otimes 2} = \mathbf{x}\mathbf{x}^{\mathsf{T}}$. By Bartlett's identity and the Law of Larger Numbers, $\mathcal{I}(\boldsymbol{\beta})$ can be consistently estimated by $n^{-1}\sum_{i=1}^{n}\left[\partial \log \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \partial\boldsymbol{\beta}\right]^{\otimes 2}$.

4.4.2 Main study/validation study

In addition to the main study data $\{(X_i^*, \mathbf{z}_i, \mathbf{S}_i), i = 1, \ldots, n_1\}$, suppose the internal validation study with data $(X_i, X_i^*, \mathbf{z}_i, \mathbf{S}_i)$ where $i = n_1 + 1, \ldots, n_1 + n_2$, n_1 is the sample size of the main study, and n_2 is the sample size of the validation study. Typically, n_1 is much greater than n_2 due to the high cost of validating error-prone measurements. Let Δ_i denote a selection indicator for individual i where $\Delta_i = 1$ if individual i is in the validation sample and $\Delta_i = 0$ otherwise. The likelihood for a main study/internal validation study design is

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n_1+n_2} \left(\mathcal{L}_i^{\mathrm{C}}\right)^{\Delta_i} \left(\mathcal{L}_i^{\mathrm{I}}\right)^{1-\Delta_i}$$

where $\mathcal{L}_i^{\mathrm{C}}$ is the likelihood contributed from an individual in the internal validation study, and $\mathcal{L}_i^{\mathrm{I}}$ is the likelihood contributed from an individual in the main study. In order to overcome the identifiability issue, the information of reclassification process in the validation study is used such that the covariate X_i is treated as a random variable instead of a constant. Specifically, we calculate $\mathcal{L}_i^{\mathrm{C}}$ and $\mathcal{L}_i^{\mathrm{I}}$ as follows:

$$\begin{aligned}
\mathcal{L}_i^{\mathrm{C}} &= \Pr(\Delta_i = 1, X_i, \mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \pi(\Delta_i = 1 \mid X_i, X_i^*, \mathbf{z}_i; \boldsymbol{\phi}) \Pr(\mathbf{S}_i \mid X_i, \mathbf{z}_i; \boldsymbol{\beta}) \Pr(X_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}) \\
&\propto \Pr(\mathbf{S}_i \mid X_i, \mathbf{z}_i; \boldsymbol{\beta}) \Pr(X_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}),
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{L}_i^{\mathrm{I}} &= \Pr(\Delta_i = 0, \mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \pi(\Delta_i = 0 \mid X_i = x_1, X_i^*, \mathbf{z}_i; \boldsymbol{\phi}) \Pr(\mathbf{S}_i \mid X_i = x_1, \mathbf{z}_i; \boldsymbol{\beta}) \Pr(X_i = x_1 \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha})
\end{aligned}$$

$$
\begin{aligned}
&+ \pi\left(\Delta_i = 0 \mid X_i = x_2, X_i^*, \mathbf{z}_i; \boldsymbol{\phi}\right) \Pr\left(\mathbf{S}_i \mid X_i = x_2, \mathbf{z}_i; \boldsymbol{\beta}\right) \Pr\left(X_i = x_2 \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right) \\
\propto\ & \Pr\left(\mathbf{S}_i \mid X_i = x_1, \mathbf{z}_i; \boldsymbol{\beta}\right) \Pr\left(X_i = x_1 \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right) \\
&+ \Pr\left(\mathbf{S}_i \mid X_i = x_2, \mathbf{z}_i; \boldsymbol{\beta}\right) \Pr\left(X_i = x_2 \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right),
\end{aligned}
$$

where $\pi\left(\cdot; \boldsymbol{\phi}\right)$ is the internal validation study selection model. We assume that

$$
\pi\left(\Delta_i = 0 \mid X_i, X_i^*, \mathbf{z}_i; \boldsymbol{\phi}\right) = \pi\left(\Delta_i = 0 \mid \mathbf{z}_i; \boldsymbol{\phi}\right),
$$

i.e., the selection of individual i into the internal validation study does not depend on either true or observed covariates given perfectly measured covariates. The log-likelihood for the main study/internal validation study design takes the form of

$$
\begin{aligned}
\log \mathcal{L}_{\mathrm{IVS}}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right) \quad = \quad & \sum_{i=1}^{n_1} \log\left[\Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right] + \sum_{i=n_1+1}^{n_1+n_2} \log\left[\Pr\left(X_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right)\right] \\
& + \sum_{i=n_1+1}^{n_1+n_2} \sum_{j=1}^{m_i} \log\left[\Pr\left(S_{ij} \mid S_{i,j-1}, X_i, \mathbf{z}_i; \boldsymbol{\beta}\right)\right] \qquad (4.13)
\end{aligned}
$$

where $\Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)$ is given by (4.12), $\Pr\left(X_i = x_j \mid X_i^* = x_i, \mathbf{z}_i; \boldsymbol{\alpha}\right)$ is the reclassification probability $\lambda_{ij}\left(\mathbf{z}_i\right)$ in (4.3)–(4.6), $\Pr\left(S_{ij} \mid S_{i,j-1}, X_i, \mathbf{z}_i; \boldsymbol{\beta}\right)$ is the transition probability $P_{s_{i,j-1}, s_{ij}}\left(t_{i,j-1}, t_{ij} \mid X_i, \mathbf{z}_i; \boldsymbol{\beta}\right)$ defined in Section 4.2.1.

Suppose the external validation study with data $\left(X_i, X_i^*, \mathbf{z}_i\right)$ is available, $i = n_1+1, \ldots, n_1+n_2$. The likelihood contributed from an individual in the external validation study is

$$
\mathcal{L}_i^{\mathrm{C}} = \pi^*\left(\Delta_i = 1 \mid X_i, X_i^*, \mathbf{z}_i; \boldsymbol{\phi}\right) \Pr\left(X_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right)
$$

where $\pi^*\left(\cdot; \boldsymbol{\phi}\right)$ is the external validation study selection model, which is assumed ignorable. If the terms in the third summation in (4.13) are deleted, the log-likelihood for the main study/external

validation study design can be obtained as follows

$$\log \mathcal{L}_{\text{EVS}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n_1} \log \Big[\Pr(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}) \Big] + \sum_{i=n_1+1}^{n_1+n_2} \log \Big[\Pr(X_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}) \Big].$$

The maximum likelihood estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ can be obtained by the Newton-Raphson method when both gradient and Hessian of the log-likelihood function are available; alternatively, they are obtained by the quasi-Newton method which is only required to specify the gradient. The gradient and Hessian of the log-likelihood function are presented in Section 4.8.1.

Under suitable regularity conditions, the maximum likelihood estimator $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is consistent for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and asymptotically normally distributed:

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \end{pmatrix} \xrightarrow{d} \mathbf{N}\left[\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\alpha}, \boldsymbol{\beta})\right], \quad \text{as } n \to \infty,$$

where

$$\begin{aligned} \mathcal{I}(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= E \begin{bmatrix} -\partial^2 \log \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \left(\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^{\mathsf{T}}\right) & -\partial^2 \log \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \left(\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}^{\mathsf{T}}\right) \\ -\partial^2 \log \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \left(\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^{\mathsf{T}}\right) & -\partial^2 \log \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \left(\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathsf{T}}\right) \end{bmatrix} \\ &= E \left\{ \begin{bmatrix} \partial \log \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \partial \boldsymbol{\alpha} \\ \partial \log \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta} \end{bmatrix}^{\otimes 2} \right\}, \end{aligned}$$

$\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the likelihood based on the data of one individual from either main study/internal validation study design or main study/external validation study design, and $\mathbf{x}^{\otimes 2}$ is the out production of the column vector $\mathbf{x}$, i.e. $\mathbf{x}^{\otimes 2} = \mathbf{x}\mathbf{x}^{\mathsf{T}}$. By Bartlett's identity and the law of larger

numbers, $\mathcal{I}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ can be consistently estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}\partial \log \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})/\partial \boldsymbol{\alpha} \\ \partial \log \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}\end{bmatrix}^{\otimes 2}.$$

4.5 Simulation studies

Simulation studies are carried out to evaluate the performance of the proposed MLEs for the Markov models with one misclassified binary covariate and the consequence of the naive MLEs which ignore the covariate misclassification.

4.5.1 Simulation setting

Simulation studies access the three-state progressive time-homogenous Markov models. The numbers of individuals are $n_1 = 500$ in the main study and $n_2 = 50$ or 100 in the validation study, and a total of 1,000 replications are used to evaluate the performance of the proposed methods.

Each individual is assumed to start from state 1 at the initial time $t_{i0} = 0$ and be observed at eleven examination times, $t_{i1}, \ldots, t_{i11}$. The gap between two adjacent examination times, $t_{ij} - t_{i,j-1}$, is uniformly distributed on the interval $[0.5, 1.0]$, where $j = 1, \ldots, 11$. A continuous covariate is generated from the standard normal distribution. One observed binary covariate X^* is generated from the Bernoulli distribution with possible values -1 and 1 and probabilities $2/3$ and $1/3$. Conditional on X^*, the true binary covariate X is generated based on reclassification probabilities

$$\alpha_1 \;=\; \Pr(X = 1 \mid X^* = -1) = 1 - \Pr(X = -1 \mid X^* = -1) = 0.3;$$

$$\alpha_2 \quad = \quad \Pr\left(X = -1 \mid X^* = 1\right) = 1 - \Pr\left(X = 1 \mid X^* = 1\right) = 0.1.$$

In the main study, we set $\Pr\left(X^* = 1\right) = \left(0.5 - \alpha_1\right) / \left(1 - \alpha_1 - \alpha_2\right)$ such that the true binary covariate X is uniformly distributed (see Section 4.8.3). In the validation study, we set the numbers of individuals with different observed binary covariate values are equal, i.e.

$$\#\left(X^* = -1\right) = \#\left(X^* = 1\right) = n_2/2.$$

In the transition intensity model

$$q_{i,i+1} = \exp\left(\beta_{i0} + \beta_{ix}X + \beta_{iz}z\right), \qquad i = 1, 2,$$

we set $\beta_{10} = -1.0$, $\beta_{1x} = -0.2$, $\beta_{1z} = 0.6$, $\beta_{20} = -0.7$, $\beta_{2x} = -0.3$, and $\beta_{2z} = 0.5$, such that the mean sojourn times from state 1 to 2 and from state 2 to 3 are 3.97 and 3.08 if $X = -1$ and 2.66 and 1.69 if $X = 1$ (see Section 4.8.3).

The procedure for generating the panel data for each individual can be described by the following procedure:

1. Simulate the continuous covariate z_i from the standard normal distribution.

2. Simulate the observed binary covariate x_i^* from a Bernoulli trial with probabilities

$$\Pr\left(X^* = -1\right) = 2/3 \quad \text{and} \quad \Pr\left(X^* = 1\right) = 1/3.$$

3. Conditional on the observed binary covariate x_i^*, simulate the true binary covariate x_i from

the Bernoulli trial with probabilities

$$\Pr(X = -1 \mid X^* = -1) = 0.7 \quad \text{and} \quad \Pr(X = 1 \mid X^* = -1) = 0.3, \qquad \text{if } x_i^* = -1,$$

or

$$\Pr(X = -1 \mid X^* = 1) = 0.1 \quad \text{and} \quad \Pr(X = 1 \mid X^* = 1) = 0.9, \qquad \text{if } x_i^* = 1.$$

4. Simulate the gap time $t_{ij} - t_{i,j-1}$ from the uniform distribution on the interval $[0.5, 1.0]$ and calculate the examination times t_{ij}, where $t_{i0} = 0$, $j = 1, \ldots, 11$.

5. Initialize the state $S_{i0} = 1$.

6. For the given covariates x_i and z_i, calculate the transition rates $q_{12}(x_i, z_i)$ and $q_{23}(x_i, z_i)$.

7. Simulate the sojourn times from state 1 to 2 and from state 2 to 3, τ_{i1} and τ_{i2}, by drawing from exponential distributions with mean $1/q_{12}(x_i, z_i)$ and $1/q_{23}(x_i, z_i)$ respectively.

8. According to the time of entering state 2 and state 3, τ_{i1} and $\tau_{i1} + \tau_{i2}$, calculate the underlying states at each observation time: for $j = 1, \ldots, 11$, if $t_{ij} < \tau_{i1}$ then $S_{ij} = 1$; if $\tau_{i1} \leq t_{ij} < \tau_{i1} + \tau_{i2}$ then $S_{ij} = 2$; if $t_{ij} \geq \tau_{i1} + \tau_{i2}$ then $S_{ij} = 3$.

4.5.2 Simulation results

Table 4.1 presents simulation results for the three-state progressive model with one misclassified binary covariate based on known reclassification probabilities. We consider three scenarios for known reclassification proabilities:

- the reclassification probabilities are the same as the values in the simulation setting
- the reclassification probabilities are lower than the values in the simulation setting

- the reclassification probabilities are larger than the values in the simulation setting

The results based on the true and observed binary covariates are also described for comparison; they are obtained by fitting an ordinary three-state progressive model. The estimators based on the exact relassificatoin probablities have negligible biases, and their aymptotic standard error (ASE) estimates agree well with their empirical counterpars. Compared with the results based on the true covariate, the standard error(SE) estimates, related to the intercepts (β_{i0}) and the parameters of the perfectly measured covariate (β_{iz}), increase slightly; but the SE estimates for the parameters of the misclassified covariate (β_{ix}) increase a little. However, the coverage rates of the corresponding 95% condifience intervals are around the nominal level in both situations. On the other hand, the misspecified reclassification probailities yield the biases in the estimates of β_{i0} and β_{ix}, but the biases of the estimates for β_{iz} are still negligible. In the situations of misspecified reclassification probabilities, the ASE and emprical standard error (ESE) estimates agree well with each other, although coverage rates are a little below the nomial level.

Table 4.2 summarizes the results for the three-state progressive model with one misclassified binary covariate based on main study/validation study design. The proposed MLEs have negligible biases and their ASE estimates agree well with the ESE estimates except those for parameters related to reclassification probabilities. As the sample size in the validation study increases, the biases, ASE and ESE estimates of the estimators of the parameters become smaller. The covarage rates of the corresponding 95% condifience intervals are close to the nominal level except those of α_{10} and α_{20}. Combining the results in Table 4.1 and 4.2, we conclude that the methods based on the true reclassification probabilities and the main study/validation study design give comparable results in terms of biases and coverage rates of the confidence intervals. However, the standard error estimates obtained from the main study/internal validation study design are slightly less than those in the sensitivity method based on the true reclassification probabilities, which are

slightly less than the standard error estimates obtained in the main study/external validation study design.

Table 4.1: Simulation results for three-state progressive models with a misclassified binary covariate based on known reclassification probabilities

	True covariate				Observed covariate							
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%				
β_{10}	.005	.047	.047	94.6	−.045	.050	.050	85.4				
β_{1x}	−.001	.047	.045	95.4	.085	.050	.051	59.7				
β_{1z}	.004	.049	.048	95.3	.000	.050	.051	94.5				
β_{20}	.003	.051	.053	93.2	−.071	.054	.055	73.6				
β_{2x}	−.002	.051	.050	95.2	.130	.054	.054	33.6				
β_{2z}	.003	.054	.057	93.6	−.014	.056	.056	93.8				
	$(\alpha_1, \alpha_2) = (0.3, 0.1)$				$(\alpha_1, \alpha_2) = (0.2, 0.05)$				$(\alpha_1, \alpha_2) = (0.4, 0.15)$			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{10}	.006	.049	.050	93.7	−.015	.048	.048	92.7	.032	.052	.054	90.5
β_{1x}	.004	.079	.080	92.9	.038	.069	.070	90.0	−.027	.089	.089	91.7
β_{1z}	.004	.050	.051	94.8	.002	.050	.050	95.0	.004	.051	.051	94.9
β_{20}	.005	.054	.053	94.9	−.027	.052	.052	91.5	.046	.058	.058	89.7
β_{2x}	.003	.084	.082	95.0	.052	.077	.073	89.7	−.037	.089	.087	92.7
β_{2z}	.002	.057	.059	93.5	−.005	.057	.059	93.0	.006	.058	.060	93.7

1000 replicates; $\Pr(X = -1) = \Pr(X = 1) = 0.5$
$(\beta_{10}, \beta_{1x}, \beta_{1z}, \beta_{20}, \beta_{2x}, \beta_{2z}) = (-1.0, -0.2, 0.6, -0.7, -0.3, 0.5)$
$(\alpha_1, \alpha_2) = (0.3, 0.1)$

4.6 Application to the PsA data

In this section, we apply our proposed methods to analyze the data arising from the psoriatic arthritic (PsA) study, which are available in the `msm` package (Jackson, 2011). This data set contains 305 subjects with 806 observations, which represent visits to a psoriatic arthritis (PsA) clinic. Psoriatic arthritis (PsA) is a progressive disease, in which the progerssion is usually reflected in the accumulation and severity of damaged joints. We consider a three-state progressive model shown in Figure 4.1 to model the progression of PsA: subjects in state 1 have no damaged joints, subjects in state 2 have 1 to 4 damaged joints, and subjects in state 3 have 5 or more damaged joints. A risk factor, denoted by X_i, is taken as the presence or absence of five or more effusions (coded by 'hieff', -1 for "no presence", $+1$ for "presence"). This covariate, is time-independent with 48 positive values and 257 negative values among all the subjects.

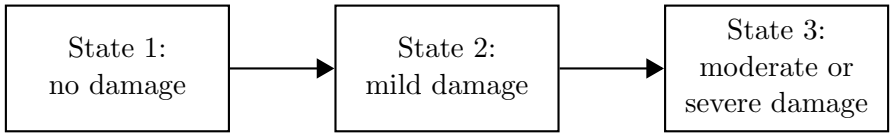

Figure 4.1: Three-state progressive model for the PsA study

In the three-state progression model, transition intensities are modelled by the log-linear model

$$\log\left(q_{i,i+1}\right) = \beta_{i0} + \beta_{ix} \cdot X_i, \qquad i = 1, 2, \tag{4.14}$$

where x_i is the true covariate of subject i.

We conduct the following four analyses for the PsA data:

Analysis 1:

The three-state progressive Markov model (4.14) is fitted to the data $\{(X_i, \mathbf{S}_i), i = 1, \ldots, 305\}$.

Table 4.2: Simulation results for three-state progressive models with a misclassified binary covariate based on the main/validation study

	Naive Analysis							
	True covariate				Observed covariate			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{10}	.005	.047	.047	94.6	−.045	.050	.050	85.4
β_{1x}	−.001	.047	.045	95.4	.085	.050	.051	59.7
β_{1z}	.004	.049	.048	95.3	.000	.050	.051	94.5
β_{20}	.003	.051	.053	93.2	−.071	.054	.055	73.6
β_{2x}	−.002	.051	.050	95.2	.130	.054	.054	33.6
β_{2z}	.003	.054	.057	93.6	−.014	.056	.056	93.8
	Internal validation study							
	$n_2 = 50$				$n_2 = 100$			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{10}	.006	.057	.052	96.5	.006	.049	.049	95.4
β_{1x}	.003	.076	.073	94.3	−.000	.067	.066	95.1
β_{1z}	.003	.048	.050	94.8	.001	.046	.046	94.2
β_{20}	.007	.071	.062	97.2	.004	.058	.052	96.2
β_{2x}	−.003	.083	.079	94.9	−.001	.071	.068	95.3
β_{2z}	.000	.055	.055	94.9	.002	.052	.052	94.9
α_{10}	−.034	.539	.407	98.7	−.016	.395	.276	99.2
α_{20}	−.098	1.018	.574	100.0	−.062	.617	.439	100.0
	External validation study							
	$n_2 = 50$				$n_2 = 100$			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{10}	.004	.064	.057	95.8	.009	.055	.053	95.8
β_{1x}	.003	.088	.083	93.8	.002	.082	.084	93.3
β_{1z}	.004	.051	.051	94.8	.004	.050	.051	94.1
β_{20}	.004	.080	.065	97.4	.006	.065	.064	95.4
β_{2x}	.000	.096	.087	94.2	.003	.089	.086	94.9
β_{2z}	.000	.058	.060	93.3	.001	.058	.061	94.0
α_{10}	−.086	.502	.425	98.0	−.010	.330	.295	97.4
α_{20}	−.091	1.174	.604	100.0	−.109	.633	.457	99.8

Analysis 2:

To illustrate our methods in real application, we consider a scenario that the surrogate measurement, denoted by X_i^*, is available, but X_i is not observed. Specifically, the surrogate measurement is related to the true covariate X_i in a way such that only one type of reclassifications, denoted by $+ \mapsto -$, is present, i.e.

$$\Pr\left(X = +1 \mid X^* = -1\right) = 0 \quad \text{and} \Pr\left(X = -1 \mid X^* = +1\right) > 0.$$

In particular, the surrogate measurement X_i^* is generated by the following procedure:

1. Conditional on the true binary covariate X_i, simulate the surrogate measurement X_i^* from the Bernoulli trial with probabilities

$$\Pr\left(X^* = +1 \mid X = +1\right) = \Pr\left(X^* = -1 \mid X = -1\right) = 0.8.$$

2. If the value of the surrogate measurement is negative, i.e. $X_i^* = -1$, then this value is replaced by the corresponding value of the true covariate, i.e. $X_i^* = X_i$.

The simulated surrogate measurements contain 107 positive values and 198 negative values. There are 59 negative values in the true covariates out of 107 surrogate measurements with the positive value.

The three-state progressive Markov model (4.14) is fitted to the data $\{(X_i^*, \mathbf{S}_i), i = 1, \ldots, 305\}$, with X_i replaced by X_i^*. This is a naive method which ignores the misclassification.

Analysis 3:

The method described in Section 4.4.1 is applied to the data $\{(X_i^*, \mathbf{S}_i), i = 1, \ldots, 305\}$, where

the reclassification probability is reparameterized as

$$\Pr\left(X = -1 \mid X^* = +1\right) = \frac{\exp\left(\alpha\right)}{1 + \exp\left(\alpha\right)},$$

and the parameter α is assumed to be known as 0.5.

Analysis 4:

The method described in Section 4.4.2 is applied to the main/internal validation data, which contain $\{(X_i^*, \mathbf{S}_i), i = 1, \ldots, 305\}$ as the main study and 30 randomly selected subjects with a positive surrogate measurement as the internal validation data.

The analysis results are summarized in Table 4.3. From these results, we have the following findings.

Analysis 1 vs Analysis 2:

The point estimates and standard errors obtained from Analyses 1 and 2 are close, except the estimate of β_{1x}. The estimate $\hat{\beta}_{1x}$ obtained from Analysis 2 is attenuated, compared with $\hat{\beta}_{1x}$ obtained from Analysis 1. The significant effect of `hieff` on the onset of PsA (State $1 \rightarrow 2$) is detected in Analysis 1 but not detected in Analysis 2, showing the consequence of ignoring the misclassification in Analysis 2.

Analysis 3 vs Analysis 4:

The point estimates in Analyses 3 and 4 agree well, and standard errors for the parameters related to the disease progression (State $2 \rightarrow 3$) in both analyses are close. However, standard errors of $\hat{\beta}_{10}$ and $\hat{\beta}_{1x}$ in Analysis 3 are greatly larger than those obtained from Analysis 4. The inflated standard error for $\hat{\beta}_{1x}$ results in the failure of detecting the significant effect of `hieff` on the onset of PsA (State $1 \rightarrow 2$) in Analysis 3.

Analysis 4 vs Analysis 1:

The results obtained based on the main study/internal validation study design (Analysis 4) agree well with the results obtained using the true covariate (Analysis 1). Both methods successfully capture the significant effect of `hieff` on the onset of PsA (State $1 \to 2$), and give comparable estimates and p-values for all the parameters.

Table 4.3: Analyses of PsA data under the three-state progressive model

				Analysis 1			Analysis 2		
		Covariate		EST	ASE	p-value	EST	ASE	p-value
Transition									
State	$1 \to 2$	Intercept	β_{10}	-2.05	0.20	$< .001$	-2.14	0.21	$< .001$
		`hieff`	β_{1x}	0.42	0.20	.036	0.29	0.22	.169
State	$2 \to 3$	Intercept	β_{20}	-1.71	0.16	$< .001$	-1.70	0.17	$< .001$
		`hieff`	β_{2x}	0.23	0.16	.135	0.25	0.17	.148
				Analysis 3			Analysis 4		
		Covariate		EST	ASE	p-value	EST	ASE	p-value
Transition									
State	$1 \to 2$	Intercept	β_{10}	-1.87	0.72	.010	-1.93	0.27	$< .001$
		`hieff`	β_{1x}	0.68	0.84	.418	0.58	0.29	.045
State	$2 \to 3$	Intercept	β_{20}	-1.62	0.23	$< .001$	-1.62	0.21	$< .001$
		`hieff`	β_{2x}	0.38	0.28	.177	0.37	0.23	.113

4.7 Discussion

In this chapter, we develop the maximum likelihood estimation procedure to analyze the panel data with misclassified discrete covariates. The sequence of discrete time points, at which the states occupied by the subjects under study were observed in the panel data, are commonly not equally spaced. In addition, the exact transition times are interval censored under panel/intermittent observation. Therefore, continuous-time Markov models are utlized for the

analysis of panel data, and the scientific interest lies in understanding the influence of variables on transitions between defined states. On the other hand, many variables are difficult to measure precisely and may be subject to measurement error. In this chapter, we restrict our attentions to discrete variables subject to classification error.

To model time-dependent intensities in Markov models, we allow the transition intensity matrix to be a piecewise constant function. This is usually achieved by specifying the baseline intensity functions to be piecewise constant, or by the approximation of time-varying variables as piecewise constant functions. In particular, time-varying variables are assumed to be constant between the time points at which they were observed, and the baseline intensity functions can be specified either to be constant or piecewise constant (Kalbfleisch and Lawless, 1985; Lindsey and Ryan, 1993; Marshall and Jones, 1995; Saint-Pierre *et al.*, 2003; Cook *et al.*, 2008; van den Hout and Matthews, 2008; Tom and Farewell, 2011). Markov models with piecewise constant transition intensities provide considerable flexibility in term of time dependence, compared with time transformation models suggested by (Kalbfleisch and Lawless, 1985), in which all the intensities after transformation must be monotonically increasing or decreasing. Although nonparametric time transformation models proposed by Hubbard *et al.* (2008) allow more flexibility, they are still restrictive due to the requirement of a common time-varying multiplicative change for all the intensities. The general smooth intensity models are developed by Titman (2011) to allow more flexibility than time transformation methods and offer more biologically plausibility in term of time dependency than piecewise constant intensity models. However, it is more computationally intensive of general smooth intensity models than other models, particularly in models with covariates.

The covariate misclassification poses an identifiability problem in the Markov model. We show that the model is not identifiable in the presence of misclassified binary covariates. The

identifiability issue is going to be explored for the discrete covariate subject to misclassification with more than two levels in the Markov model. However, the length and structure of the observed sequences of states, which the joint probability function of the states depends on, bring the challenges in investigating the identifiability of model parameters.

To address the identifiability problem, we propose the likelihood methods to make statistical inference and ensure the model identifiability in two practical situations: one is to conduct the sensitivity analysis based on the known reclassification probabilities; the other one is developed based on the main study/validation study design. The maximum likelihood estimates can be obtained by directly maximizing the log-likelihood function using the Newton-Raphson algorithm if the explicit expression of transition probabilities is available, or using the quasi-Newton algorithm with the first derivatives incorporated if the transition probabilities are calculated based on the matrix exponential. The simulation studies evaluate the performance of our proposed methods. The biases in the proposed estimates are negligible, and the coverage rate of confidence intervals are close to the nominal level, although the asymptotic standard error estimates for the reclassification parameters are larger than the corresponding empirical estimates.

4.8 Technical notes

4.8.1 Gradient and Hessian of the log-likelihood function

Known reclassification probabilities

The uth element of the score vector takes the form of

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \beta_u} = \sum_{i=1}^{n} \left[\frac{1}{\mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})} \frac{\partial \mathcal{L}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \beta_u} \right],$$

where

$$\begin{aligned}
\frac{\partial \mathcal{L}_i(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_u} &= \sum_{i=1}^{2}\left\{ \Pr\left(X_i = x_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right) \frac{\partial}{\partial \beta_u}\left[\prod_{j=1}^{m_i} \Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)\right]\right\} \\
&= \sum_{i=1}^{2}\left\{ \Pr\left(X_i = x_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right) \prod_{j=1}^{m_i}\left[\Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)\right]\right. \\
&\quad \left.\times \sum_{j=1}^{m_i}\left\{ \frac{\partial}{\partial \beta_u} \log\left[\Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)\right]\right\}\right\} \\
&= \sum_{i=1}^{2}\left\{ \Pr\left(X_i = x_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right) \prod_{j=1}^{m_i}\left[\Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)\right]\right. \\
&\quad \left.\times \sum_{j=1}^{m_i}\left[\frac{1}{\Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)} \frac{\partial}{\partial \beta_u} \Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)\right]\right\}.
\end{aligned}$$

The detailed derivation of the first derivatives of transition probabilities in piecewise constant Markov models is presented in Section 4.8.2.

The expression for the (u, v)th entry of the Hessian matrix is also available and given by

$$\begin{aligned}
\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_u \partial \beta_v} &= \sum_{i=1}^{n} \frac{\partial^2 \log \mathcal{L}_i(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_u \partial \beta_v} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_v}\left[\frac{1}{\mathcal{L}_i(\boldsymbol{\alpha},\boldsymbol{\beta})} \frac{\partial \mathcal{L}_i(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_u}\right] \\
&= \sum_{i=1}^{n}\left[\frac{1}{\mathcal{L}_i(\boldsymbol{\alpha},\boldsymbol{\beta})} \frac{\partial^2 \mathcal{L}_i(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_u \partial \beta_v} - \frac{1}{\left[\mathcal{L}_i(\boldsymbol{\alpha},\boldsymbol{\beta})\right]^2} \frac{\partial \mathcal{L}_i(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_u} \frac{\partial \mathcal{L}_i(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_v}\right],
\end{aligned}$$

where

$$\frac{\partial^2 \mathcal{L}_i(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_u \partial \beta_v} = \sum_{i=1}^{2}\left\{ \Pr\left(X_i = x_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right) \frac{\partial^2}{\partial \beta_u \partial \beta_v}\left[\prod_{j=1}^{m_i} \Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)\right]\right\},$$

and

$$\frac{\partial^2}{\partial\beta_u\partial\beta_v}\left[\prod_{j=1}^{m_i}\Pr\left(S_{ij}\mid S_{i,j-1},X_i=x_i,\mathbf{z}_i;\boldsymbol{\beta}\right)\right]=\prod_{j=1}^{m_i}\left[\Pr\left(S_{ij}\mid S_{i,j-1},X_i=x_i,\mathbf{z}_i;\boldsymbol{\beta}\right)\right]$$
$$\times\left\{\sum_{j=1}^{m_i}\left[\frac{\partial^2}{\partial\beta_u\partial\beta_v}\log\Pr\left(S_{ij}\mid S_{i,j-1},X_i=x_i,\mathbf{z}_i;\boldsymbol{\beta}\right)\right]\right.$$
$$\left.+\sum_{j=1}^{m_i}\left[\frac{\partial}{\partial\beta_u}\log\Pr\left(S_{ij}\mid S_{i,j-1},X_i=x_i,\mathbf{z}_i;\boldsymbol{\beta}\right)\right]\sum_{j=1}^{m_i}\left[\frac{\partial}{\partial\beta_v}\log\Pr\left(S_{ij}\mid S_{i,j-1},X_i=x_i,\mathbf{z}_i;\boldsymbol{\beta}\right)\right]\right\}.$$

Main study/validation study

The gradient and Hessian of $\log\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)$ with regard to $\boldsymbol{\beta}$ is given in the previous part. Similarly, the first derivative of $\log\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)$ with regard to α_{uv} can be written as

$$\begin{aligned}&\frac{\partial\log\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)}{\partial\alpha_{uv}}\\ =\;&\frac{1}{\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)}\frac{\partial\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)}{\partial\alpha_{uv}}\\ =\;&\frac{1}{\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)}\sum_{k=1}^{2}\left[\frac{\partial\Pr\left(X_i=x_k\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha}\right)}{\partial\alpha_{uv}}\prod_{j=1}^{m_i}\Pr\left(S_{ij}\mid S_{i,j-1},X_i=x_k,\mathbf{z}_i;\boldsymbol{\beta}\right)\right],\end{aligned}$$

where $u=1,2$, and $v=0,1,\ldots,p$. Note that $\partial\log\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)/\partial\alpha_{uv}=0$ if the value of X_i^* is not x_i. The second derivative of $\log\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)$ with regard to α_{uv} and α_{uw} takes the form of

$$\begin{aligned}&\frac{\partial^2\log\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)}{\partial\alpha_{uv}\partial\alpha_{uw}}\\ =\;&\frac{1}{\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)}\frac{\partial^2\Pr\left(\mathbf{S}_i\mid X_i^*,\mathbf{z}_i;\boldsymbol{\alpha},\boldsymbol{\beta}\right)}{\partial\alpha_{uv}\partial\alpha_{uw}}\end{aligned}$$

$$-\frac{1}{\left[\Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right]^2} \frac{\partial \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{\partial \alpha_{uv}} \frac{\partial \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{\partial \alpha_{uw}},$$

where

$$\frac{\partial^2 \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{\partial \alpha_{uv} \partial \alpha_{uw}} = \sum_{h=1}^{2} \left[\frac{\partial^2 \Pr\left(X_i = x_h \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right)}{\partial \alpha_{uv} \partial \alpha_{uw}} \prod_{j=1}^{m_i} \Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_h, \mathbf{z}_i; \boldsymbol{\beta}\right) \right],$$

$u = 1, 2$ and $v, w = 0, 1, \ldots, p$. Note that $\partial^2 \log \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right) / \left(\partial \alpha_{uv} \partial \alpha_{uw}\right) = 0$ if the value of X_i^* is not x_i and $\partial^2 \log \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right) / \left(\partial \alpha_{uv} \partial \alpha_{u'w}\right) = 0$ if $u \neq u'$. The second derivative of $\log \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)$ with regard to α_{uv} and β_u is given by

$$\begin{aligned}
&\frac{\partial^2 \log \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{\partial \alpha_{uv} \partial \beta_u} \\
&= \frac{1}{\Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)} \frac{\partial^2 \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{\partial \alpha_{uv} \partial \beta_u} \\
&\quad -\frac{1}{\left[\Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)\right]^2} \frac{\partial \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{\partial \alpha_{uv}} \frac{\partial \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{\partial \beta_u},
\end{aligned}$$

where

$$\begin{aligned}
&\frac{\partial^2 \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{\partial \alpha_{uv} \partial \beta_u} \\
&= \sum_{k=1}^{2} \left\{ \frac{\partial \Pr\left(X_i = x_k \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right)}{\partial \alpha_{uv}} \prod_{j=1}^{m_i} \left[\Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_k, \mathbf{z}_i; \boldsymbol{\beta}\right) \right] \right. \\
&\quad \left. \times \sum_{j=1}^{m_i} \left\{ \frac{\partial}{\partial \beta_u} \log \left[\Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_k, \mathbf{z}_i; \boldsymbol{\beta}\right) \right] \right\} \right\} \\
&= \sum_{k=1}^{2} \left\{ \frac{\partial \Pr\left(X_i = x_k \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}\right)}{\partial \alpha_{uv}} \prod_{j=1}^{m_i} \left[\Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_k, \mathbf{z}_i; \boldsymbol{\beta}\right) \right] \right.
\end{aligned}$$

$$\times \sum_{j=1}^{m_i} \left[\frac{1}{\Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_k, \mathbf{z}_i; \boldsymbol{\beta}\right)} \frac{\partial}{\partial \beta_u} \Pr\left(S_{ij} \mid S_{i,j-1}, X_i = x_k, \mathbf{z}_i; \boldsymbol{\beta}\right) \right] \Bigg\}.$$

Note that $\partial^2 \log \Pr\left(\mathbf{S}_i \mid X_i^*, \mathbf{z}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}\right) / \left(\partial \alpha_{uv} \partial \beta_u\right) = 0$ if the value of X_i^* is not x_i.

4.8.2 First derivatives of transition probabilities in piecewise constant Markov models

If transition probabilities can not be analytically calculated from transition intensities, the canonical decomposition for the computation of $\mathbf{P}\left(s, s+t\right)$ is available when $\mathbf{Q}_k$ has distinct eigenvalues (Kalbfleisch and Lawless, 1985). In this case,

$$\mathbf{Q}_k = \mathbf{H}_k \mathbf{D}_k \mathbf{H}_k^{-1}, \qquad k = 0, \ldots, M,$$

where $\mathbf{D}_k = \operatorname{diag}\left(d_{k1}, d_{k2}, \ldots, d_{kK}\right)$ is a diagonal matrix of distinct eigenvalues of $\mathbf{Q}_k$ and $\mathbf{H}_k$ is the $K \times K$ matrix whose jth column is the eigenvector associated with d_{kj}. Then, $\mathbf{P}\left(s, s+t\right)$ is calculated as

$$\mathbf{P}\left(s, s+t\right) = \begin{cases} \mathbf{H}_k \exp\left(\mathbf{D}_k t\right) \mathbf{H}_k^{-1}, & b_k < s \leq s+t \leq b_{k+1}, \\ \mathbf{P}\left(s, b_{i+1}\right) \left\{ \prod_{k=i+1}^{j-1} \mathbf{P}\left(b_k, b_{k+1}\right) \right\} \mathbf{P}\left(b_j, s+t\right), & b_i < s \leq b_{i+1} \leq b_j < s+t \leq b_{j+1}, \end{cases}$$

where the dependence of $\mathbf{Q}_k$, $\mathbf{P}\left(s, s+t\right)$, $\mathbf{H}_k$ and $\mathbf{D}_k$ on $\boldsymbol{\beta}$ is suppressed for notational convenience. On the other hand, if $\mathbf{Q}_k$ has repeated eigenvalues, Kalbfleisch and Lawless (1985) suggested an analogous decomposition of $\mathbf{Q}_k$ to the Jordan canonical form (e.g. Cox and Miller, 1965, Chapter 3). More recently, Jackson (2011) recommended the method based on Padé approximation with scaling and squaring (Moler and van Loan, 2003) for the case with repeated

eigenvalues. However, for most models of interest, $\mathbf{Q}_k(\boldsymbol{\beta})$ has distinct eigenvalues for almost all $\boldsymbol{\beta}$ and therefore it is rarely necessary (Kalbfleisch and Lawless, 1985).

The first derivatives of transition probabilities can be computationally efficiently obtained by

$$\frac{\partial \mathbf{P}(s, s+t)}{\partial \beta_u} = \mathbf{H}_k \mathbf{V}_{ku} \mathbf{H}_k^{-1}, \qquad \text{if } b_k < s \leq s+t \leq b_{k+1}, \tag{4.15}$$

where $\mathbf{V}_{ku}$ is a $K \times K$ matrix with (i,j) entry

$$\begin{cases} g_{kij}^{(u)} \Big[\exp(d_{ki}t) - \exp(d_{kj}t) \Big] / (d_{ki} - d_{kj}), & \text{if } i \neq j, \\ g_{kii}^{(u)} t \exp(d_{ki}t), & \text{if } i = j, \end{cases}$$

and $g_{kij}^{(u)}$ is the (i,j) entry in $\mathbf{G}_k^{(u)} = \mathbf{H}_k^{-1} (\partial \mathbf{Q} / \partial \beta_u) \mathbf{H}_k$. A derivation of this result for the time-homogeneous case appears in Jennrich and Bright (1976) and Kalbfleisch and Lawless (1985). If $b_i < s \leq b_{i+1} \leq b_j < s+t \leq b_{j+1}$, the first derivatives of transition probabilities can be written as

$$\begin{aligned} \frac{\partial \mathbf{P}(s, s+t)}{\partial \beta_u} = \mathbf{P}(s, b_{i+1}) \prod_{k=i+1}^{j-1} [\mathbf{P}(b_k, b_{k+1})] \mathbf{P}(b_j, s+t) \times \Bigg\{ & \frac{1}{\mathbf{P}(s, b_{i+1})} \frac{\partial \mathbf{P}(s, b_{i+1})}{\partial \beta_u} \\ & + \sum_{k=i+1}^{j-1} \left[\frac{1}{\mathbf{P}(b_k, b_{k+1})} \frac{\partial \mathbf{P}(b_k, b_{k+1})}{\partial \beta_u} \right] + \frac{1}{\mathbf{P}(b_j, s+t)} \frac{\partial \mathbf{P}(b_j, s+t)}{\partial \beta_u} \Bigg\}, \end{aligned}$$

where the derivatives of transition probabilities within the interval of constant transition intensities are given by (4.15).

4.8.3 Effects of parameters in simulation studies

Measurement error model

The observed covariate X^* is generated from the Bernoulli distribution with successive probability p for value one. Then the true covariate X is generated based on $\alpha_0 = \Pr(X = 1 \mid X^* = -1)$ and $\alpha_1 = \Pr(X = -1 \mid X^* = 1)$. Note that

$$\begin{aligned}
\Pr(X = 1) &= \Pr(X = 1, X^* = 1) + \Pr(X = 1, X^* = 0) \\
&= \Pr(X = 1 \mid X^* = 1) \Pr(X^* = 1) + \Pr(X = 1 \mid X^* = 0) \Pr(X^* = 0) \\
&= (1 - \alpha_1) p + \alpha_0 (1 - p) \\
&= (1 - \alpha_1 - \alpha_0) p + \alpha_0.
\end{aligned}$$

If we set $\Pr(X = 1) = 0.5$, then $p = \Pr(X^* = 1) = (0.5 - \alpha_0) / (1 - \alpha_0 - \alpha_1)$ in the case that $\alpha_0 + \alpha_1 \neq 1$.

Transition intensity model

In the unidirectional progressive model, the sojourn time τ_i is exponentially distributed with mean $1/q_i$, $i = 1, 2$, where $q_1 = \exp(\beta_{10} + \beta_{1x} X + \beta_{1z} z)$ and $q_2 = \exp(\beta_{20} + \beta_{2x} X + \beta_{2z} z)$ are transition intensities; X is a surrogate binary covariate which follows the discrete uniform distribution with values -1 and 1; z is a standard normally distributed error-free covariate.

Then, the mean sojourn time conditional on X is

$$\begin{aligned}
E(\tau_i \mid X) &= E[1/\exp(\beta_{i0} + \beta_{ix} X + \beta_{iz} z) \mid X] \\
&= \exp(-\beta_{i0} - \beta_{ix} X) E[\exp(-\beta_{iz} z)]
\end{aligned}$$

$$
\begin{aligned}
&= \exp\left(-\beta_{i0} - \beta_{ix}X\right) \int_{-\infty}^{\infty} \exp\left(-\beta_{iz}u\right) \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, \mathrm{d}u \\
&= \frac{\exp\left(-\beta_{i0} - \beta_{ix}X\right)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}u^2 - \beta_{iz}u\right) \mathrm{d}u \\
&= \frac{\exp\left(-\beta_{i0} - \beta_{ix}\right)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(u + \beta_{iz}\right)^2 + \frac{1}{2}\beta_{iz}^2\right] \mathrm{d}u \\
&= \exp\left(\frac{1}{2}\beta_{iz}^2 - \beta_{i0} - \beta_{ix}X\right).
\end{aligned}
$$

If we set $(\beta_{10}, \beta_{1x}, \beta_{1z}, \beta_{20}, \beta_{2x}, \beta_{2z}) = (-1.0, -0.2, 0.6, -0.7, -0.3, 0.5)^{\mathsf{T}}$, then

X	$E(\tau_1 \mid X)$	$E(\tau_2 \mid X)$
-1	3.97	3.08
1	2.66	1.69

Chapter 5

Statistical inference of two-state Markov models for panel data with time-dependent surrogate covariates

5.1 Introduction

A study of disease progression often involves longitudinal follow-up on a group of subjects. Many diseases are measured by a binary outcome where the scientific interest lies in the inference about the rate of transitions between the disease states and about the influence of covariates on transitions. Examples include: (i) chronic bronchitis where subjects may transit between the exacerbation of symptoms and a symptom resolution (Cook *et al.*, 1999); (ii) parasitic infection where subjects transit between the presence or absence of the parasite (Nagelkerke *et al.*, 1990); (iii) unipolar depression where subjects transit between periods of depression and periods of normal mood (Frank *et al.*, 1990); (iv) migraine where subjects transit between migraine attacks

and pain free periods (Tfelt-Hansen and Olesen, 1985), and so on.

A feature of panel data collected from disease progression studies is the irregular spacing in the observation times. Moreover, the observation times may be unique to each subject and the exact times of disease onset or progression are interval censored. That is, the disease information at the intermittent follow-up visits is known, but the information between visits is commonly unavailable. For analyzing of such data, continuous-time Markov models play an important role in handling the irregularly spaced observation times due to the feasibility of constructing the likelihood in such models. A widely used approach to fit a time-homogeneous Markov model is the Fisher-scoring algorithm proposed by Kalbfleisch and Lawless (1985) for obtaining the maximum likelihood estimates and corresponding asymptotic covariance matrix. The applications of this method can be found in HIV/AIDS studies (Gentleman *et al.*, 1994) and rheumatology (Gladman *et al.*, 1995), among many others.

Another feature of panel data collected from disease progression studies is time-dependence on covariates, such as, the blood pressure observed during every clinic visit and the fat and calories intake records taken at every interview. A commonly used method, which allows for time-dependent covariates in continuous-time Markov models, assumes that time-dependent covariates remain the same between two consecutive times, and then the contribution from time-independent covariates is replaced with the contribution from the time-dependent covariates at the specific time (Saint-Pierre *et al.*, 2003). This method yields the piecewise-constant intensities with the change points specified by the observation times of the time dependent covariate in the Markov models. However, the discontinuities of transition intensities determined by the covariate observation times may not be plausible for some applications. On the other hand, the long-term average, instead of the time-dependent covariate, may be the true predictor in the regression model, such as the long-term blood pressure and the long-term diet intake. Therefore, we use the

measurement error model to compensate for the time-dependent covariate, in which the multiple observations of the time-dependent covariate are treated as the surrogates of the unobserved true predictor.

A number of approaches to reduce or correct the effects of measurement error have been discussed previously. Yi and He (2006) proposed methods for bivariate survival data with mismeasured covariates under an accelerated failure time model. Yi and Lawless (2007) developed a corrected likelihood method for the proportional hazards model with covariates subject to measurement error. Yi (2008) developed a simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. Yi (2009) reviewed some analysis methods handling covariate measurement error for life history data. Yi *et al.* (2011a) developed likelihood method to make simultaneous inference for longitudinal data with covariate measurement error and missing responses. Yi and He (2012) developed the simulation-extrapolation method for survival data with covariate measurement error under parametric proportional odds models. Yi *et al.* (2012) developed a functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. Yi and Lawless (2012) developed likelihood-based and marginal inference methods for recurrent event data with covariate measurement error. Yi *et al.* (2015) developed Functional and structural methods for mixed measurement error and mislassification in covariates. Yan and Yi (2015) developed a class of functional methods for error-contaminated survival data under additive hazards models with replicate measurements. However, relatively less attention has been paid to the covariate measurment error in the panel data.

In this chapter, we describe both structural and functional modelling approaches for inference about two-state Markov models where a time-independent covariate is unavailable but its time-dependent surrogate measurements are collected. In Section 5.2, the two-state Markov model and

the classic measurement error model are introduced. In Section 5.3, two functional modelling approaches, simulation extrapolation and regression calibration, are presented. Both approaches make no distributional assumption on the unobserved true covariate. The simulation studies are conducted to evaluate the performance of these methods in Section 5.4. In Section 5.5, the likelihood analysis is proposed via an Monte Carlo EM algorithm through the structural modelling, which assumes a parametric distribution for the unobserved true covariate, and simulation results are also presented. The discussion is given in Section 5.6 and technical details are presented in Section 5.7.

5.2 Model setup

5.2.1 Two-state Markov model

Consider a two-state bidirectional Markov model with the states denoted by 1 and 2. Let u denote the transition intensity from state 1 to 2 and v denote the transition intensity from state 2 to 1. Then, the transition intensity matrix is given by

$$\mathbf{Q} = \begin{pmatrix} -u & u \\ v & -v \end{pmatrix},$$

and the transition probabilities take the following forms:

$$\begin{aligned} P_{12}(t) &= \frac{u}{u+v}\Big[1 - \exp\{-(u+v)\,t\}\Big], \\ P_{21}(t) &= \frac{v}{u+v}\Big[1 - \exp\{-(u+v)\,t\}\Big], \\ P_{11}(t) &= 1 - P_{12}(t), \\ P_{22}(t) &= 1 - P_{21}(t), \end{aligned}$$

where $P_{ij}(t) = \Pr[S(t+s) = j \mid S(s) = i]$, and $S(t)$ is the state occupied at time t.

Let π_i denote the stationary probability, $i = 1, 2$. That is, $\boldsymbol{\pi}\mathbf{P}(t) = \boldsymbol{\pi}$, for all t, where $\boldsymbol{\pi} = (\pi_1, \pi_2)$. Then,

$$\pi_1 = \frac{v}{u+v} \quad \text{and} \quad \pi_2 = \frac{u}{u+v}. \tag{5.1}$$

5.2.2 Transition intensity model

Let X denote an unobserved time-independent continuous covariate and $\mathbf{Z}$ be a $p \times 1$ vector of perfectly measured time-independent covariates. For the time-homogeneous Markov model, we consider regression models

$$u(X, \mathbf{Z}) = u_0 \exp\left(\beta_{ux} X + \boldsymbol{\beta}_{uz}^{\mathsf{T}} \mathbf{Z}\right), \tag{5.2}$$

$$v(X, \mathbf{Z}) = v_0 \exp\left(\beta_{vx} X + \boldsymbol{\beta}_{vz}^{\mathsf{T}} \mathbf{Z}\right), \tag{5.3}$$

where u_0 and v_0 are baseline transition intensities out of state 1 to 2 and state 2 to 1, respectively, and $\left(\beta_{ux}, \beta_{uz_1}, \ldots, \beta_{uz_p}\right)$ and $\left(\beta_{vx}, \beta_{vz_1}, \ldots, \beta_{vz_p}\right)$ are vectors of regression coefficients of primary interest.

For the time-homogeneous model, the parametric form of the baseline transition intensity, i.e.

$$u_0 = \exp(\beta_{u0}) \quad \text{and} \quad v_0 = \exp(\beta_{v0})$$

is considered (e.g. Kalbfleisch and Lawless, 1985; Jackson *et al.*, 2003). The transition intensity matrix $\mathbf{Q}$ incorporating the covariates is then used to calculate the likelihood.

5.2.3 Measurement error model

Let $X^*(t)$ be the time-dependent surrogate measure of X at time t. The replicate measurements $\{X^*(t) : t \geq 0\}$ of X follow the additive error model

$$X^*(t) = X + U(t), \tag{5.4}$$

where $U(t)$ is independent of X and normally distributed with mean zero and variance σ_u^2.

5.3 Functional methods of reducing measurement error effects

In this section, as opposed to the naive analysis which ignores measurement error, we develop functional methods which reduce the effects of measurement error. Suppose n independent subjects are under study. The data for subject i consist of the observed states $\mathbf{s}_i = \{s_{i0}, s_{i1}, \ldots, s_{im_i}\}$ and the error-prone covariates $\mathbf{x}_i^* = \left\{x_{i0}^*, x_{i1}^*, \ldots, x_{im_i}^*\right\}$ at the times $t_{i0} < t_{i1} < \cdots < t_{im_i}$ and the time-independent covariates $\mathbf{z}_i$.

5.3.1 Naive maximum likelihood estimation

If the true covariate value x_i were known for each subject, then the log-likelihood for subject i is

$$\ell_i(\boldsymbol{\beta}) = \log\Big\{\Pr(s_{i0} \mid x_i, \mathbf{z}_i; \boldsymbol{\beta})\Big\} + \sum_{j=1}^{m_i} \log\Big\{\Pr(s_{ij} \mid s_{i,j-1}, x_i, \mathbf{z}_i; \boldsymbol{\beta})\Big\}, \tag{5.5}$$

where $\Pr(s_{i0} \mid x_i, \mathbf{z}_i; \boldsymbol{\beta})$ is defined to be the statinary probability $\pi_{s_{i0}}$, and $\Pr(s_{ij} \mid s_{i,j-1}, x_i, \mathbf{z}_i; \boldsymbol{\beta})$ is the transition probability $P_{s_{i,j-1},s_{ij}}(t_{ij} - t_{i,j-1})$ defined in Section 5.2.1. The log-likelihood over

all the subjects takes the form

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}). \tag{5.6}$$

The first and second order derivatives of the logarithms of stationary probabilities and transition probabilities are presented in Section 5.7.2. The Newton-Raphson algorithm can be used to obtain the maximum likelihood estimates.

5.3.2 Simulation extrapolation

The simulation extrapolation (SIMEX) method (Stefanski and Cook, 1995) is a simulation-based functional method for measurement error problems, in which no distribution assumption is made on the true covariate. The idea of the SIMEX method is to establish the trend of naive estimates towards the variance of the induced measurement error by incorporating additional variability to the observed measurement and then extrapolate the trend back to the case of no measurement error to obtain parameter estimates for the true covariates. When replicate measurements are available for each subject, Devanarayan and Stefanski (2002) developed the empirical SIMEX, which allows for unknown measurement error variance. This method does not require the assumption of homogeneity of error variance and uses the replicate measurements directly to compute pseudo data.

The empirical SIMEX procedure consists of two steps, a SIMulation step and an Extrapolation step. In the simulation step, naive estimates are obtained from the pseudo data generated with the measurement error variance $(1+\xi)\,\sigma_{iu}^2$. Without knowing σ_{iu}^2, the pseudo data are generated by the linear combination of replicate measurements. For subject i, (m_i+1) independent and identically distributed standard normal random numbers, $\{y_{bij} : j = 0, \ldots, m_i\}$, are generated

and the empirical pseudo data are defined to be

$$x_{bi}(\xi) = \bar{x}^*_{i\cdot} + \sqrt{\frac{\xi}{m_i+1}} \sum_{j=0}^{m_i} c_{bij}\, x^*_{ij},$$

where $\bar{x}^*_{i\cdot} = (m_i+1)^{-1}\left\{\sum_{j=0}^{m_i} x^*_{ij}\right\}$, $c_{bij} = \left\{\sum_{j=0}^{m_i}(y_{bij} - \bar{y}_{bi\cdot})^2\right\}^{-1/2}(y_{bij} - \bar{y}_{bi\cdot})$, and $\bar{y}_{bi\cdot} = (m_i+1)^{-1}\left\{\sum_{j=0}^{m_i} y_{bij}\right\}$, $b = 1, \ldots, B$, $i = 1, \ldots, n$. The naive estimates from the simulated data, denoted by $\hat{\boldsymbol{\beta}}(b, \xi)$, are obtained from maximizing the log-likelihood (5.6) by replacing x_i with $x_{bi}(\xi)$. The corresponding covariance estimate, denoted by $\widehat{\boldsymbol{\Omega}}(b, \xi)$, is computed by the matrix

$$\left\{ -\sum_{i=1}^{n} \frac{\partial^2 \ell_i\left\{\boldsymbol{\beta}; \mathbf{s}_i, x_{bi}(\xi), \mathbf{z}_i\right\}}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^{\mathsf{T}}}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(b,\xi)} \right\}^{-1},$$

where $\ell_i\left\{\boldsymbol{\beta}; \mathbf{s}_i, x_{bi}(\xi), \mathbf{z}_i\right\}$ is determined by (5.5) with x_i replaced by $x_{bi}(\xi)$.

To avoid the simulation variability, the estimation procedure is repeated a large number, say B, times and then the following quantities are computed:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(\xi) &= B^{-1}\left\{\sum_{b=1}^{B} \hat{\boldsymbol{\beta}}_b(\xi)\right\}, \\
\widehat{\boldsymbol{\Omega}}(\xi) &= B^{-1}\left\{\sum_{b=1}^{B} \widehat{\boldsymbol{\Omega}}(b, \xi)\right\}, \\
\boldsymbol{\Omega}^*(\xi) &= (B-1)^{-1}\left[\sum_{b=1}^{B}\left\{\hat{\boldsymbol{\beta}}(b,\xi) - \hat{\boldsymbol{\beta}}(\xi)\right\}\left\{\hat{\boldsymbol{\beta}}(b,\xi) - \hat{\boldsymbol{\beta}}(\xi)\right\}^{\mathsf{T}}\right], \\
\text{and} \qquad \widehat{\boldsymbol{\Gamma}}(\xi) &= \widehat{\boldsymbol{\Omega}}(\xi) - \boldsymbol{\Omega}^*(\xi).
\end{aligned}$$

The procedure is also repeated for a sequence of ξ, such as $\{0.0, 0.5, 1.0, 1.5, 2.0\}$, and the average of resulting naive estimates $\hat{\boldsymbol{\beta}}(\xi)$ and $\widehat{\boldsymbol{\Gamma}}(\xi)$ are plotted versus ξ. In the extrapolation step, a regression model (e.g., a quadratic model) is fitted to the average of naive estimates as a function

of ξ. The SIMEX estimates and the associated covariance estimates are obtained by extrapolating the regression model to the value $\xi = -1$. That is,

$$\hat{\boldsymbol{\beta}}_{\text{SIMEX}} = \lim_{\xi \to -1} \hat{\boldsymbol{\beta}}(\xi) \quad \text{and} \quad \widehat{\text{var}}\left(\hat{\boldsymbol{\beta}}_{\text{SIMEX}}\right) = \lim_{\xi \to -1} \widehat{\boldsymbol{\Gamma}}(\xi).$$

5.3.3 Regression calibration

The basis of the regression calibration (RC) method (Prentice, 1982; Rosner *et al.*, 1989; Carroll and Stefanski, 1990; Gleser, 1990) is to replace the true covariate X by the conditinal mean $E(X \mid X^*, \mathbf{Z})$, which can be obtained based on the regression of X on observed covariates and then perform the standard analysis. When replicate measurements of the error-prone covariate exist, the best linear approximation is suggested to estimate the regression calibration function (Carroll *et al.*, 2006, Section 4.4.2).

Suppose that there are m replicate measurements $X_1^*, \ldots, X_m^*$ of X. The best linear approximation to X given $\left(\mathbf{Z}, \bar{X}^*\right)$ is

$$E(X \mid \mathbf{z}, \bar{x}^*) \approx \mu_x + \left(\sigma_x^2, \Sigma_{xz}\right) \begin{bmatrix} \sigma_x^2 + \sigma_u^2/m & \Sigma_{xz} \\ \Sigma_{xz}^{\mathsf{T}} & \Sigma_{zz} \end{bmatrix}^{-1} \begin{pmatrix} \bar{X}^* - \mu_{x^*} \\ \mathbf{z} - \boldsymbol{\mu}_z \end{pmatrix},$$

where μ_a and σ_a^2 denote the mean and variance of random variable A respectively, and Σ_{ab} denotes the covariance matrix between two random variables A and B. Based on observations $(\mathbf{z}_i, \bar{x}_{i\cdot}^*)$ with replicate sample size $m_i + 1$, those quantities can be estimated by

$$\hat{\mu}_x \quad = \quad \hat{\mu}_{x^*} = \left\{ \sum_{i=1}^{n} (m_i + 1) \right\}^{-1} \left\{ \sum_{i=1}^{n} \sum_{j=0}^{m_i} x_{ij}^* \right\},$$

$$
\begin{aligned}
\hat{\boldsymbol{\mu}}_z &= \bar{\mathbf{z}}_{\cdot} = n^{-1}\left\{\sum_{i=1}^{n}\mathbf{z}_i\right\},\\
\hat{\sigma}_u^2 &= \left\{\sum_{i=1}^{n} m_i\right\}^{-1}\left\{\sum_{i=1}^{n}\sum_{j=0}^{m_i}\left(x_{ij}^{*}-\bar{x}_{i\cdot}^{*}\right)^2\right\},\\
\widehat{\Sigma}_{zz} &= (n-1)^{-1}\left\{\sum_{i=1}^{n}\left(\mathbf{z}_i-\bar{\mathbf{z}}_{\cdot}\right)\left(\mathbf{z}_i-\bar{\mathbf{z}}_{\cdot}\right)^{\mathsf{T}}\right\},\\
\nu &= \left\{\sum_{i=1}^{n}(m_i+1)\right\}-\left\{\sum_{i=1}^{n}(m_i+1)\right\}^{-1}\left\{\sum_{i=1}^{n}(m_i+1)^2\right\},\\
\widehat{\Sigma}_{xz} &= \nu^{-1}\left\{\sum_{i=1}^{n}(m_i+1)\left(\bar{x}_{i\cdot}^{*}-\hat{\mu}_{x^*}\right)\left(\mathbf{z}_i-\bar{\mathbf{z}}_{i\cdot}\right)^{\mathsf{T}}\right\},\\
\text{and}\qquad \hat{\sigma}_x^2 &= \nu^{-1}\left[\left\{\sum_{i=1}^{n}(m_i+1)\left(\bar{x}_{i\cdot}^{*}-\hat{\mu}_{x^*}\right)^2\right\}-(n-1)\,\hat{\sigma}_u^2\right].
\end{aligned}
$$

The resulting best linear approximation to the calibration function $E\left(X_i \mid \mathbf{z}_i, \bar{x}_i^{*}\right)$ is

$$
\hat{\mu}_x + \left(\hat{\sigma}_x^2, \widehat{\Sigma}_{xz}\right)\begin{bmatrix}\hat{\sigma}_x^2+\dfrac{\hat{\sigma}_u^2}{m_i+1} & \widehat{\Sigma}_{xz}\\ \widehat{\Sigma}_{xz}^{\mathsf{T}} & \widehat{\Sigma}_{zz}\end{bmatrix}^{-1}\begin{pmatrix}\bar{x}_i^{*}-\hat{\mu}_{x^*}\\ \mathbf{z}_i-\hat{\boldsymbol{\mu}}_z\end{pmatrix}.
$$

After replacing the true covariate x_i by the estimated regression calibration function $\widehat{E}\left(X_i \mid \mathbf{z}_i, \bar{x}_i^{*}\right)$, we can carry out the standard maximum likelihood procedure for parameters estimation. The bootstrap method can be used to obtain standard errors for parameter estimators.

5.4 Simulation studies for functional methods

In this section, simulation studies are conducted to evaluate the performance of the SIMEX and RC methods, and to illustrate the consequence of ignoring the measurement error by the naive method.

5.4.1 Simulation setting

A total of 2000 replicates are used and the number of subjects is $n = 1000$ in each simulated dataset. The number of observation for each subject is generated from the uniform distribution over the set $\{2, \ldots, 6\}$, $i = 1, \ldots, n$. The gap between two adjacent observation times, $t_{ij} - t_{i,j-1}$, is uniformly distributed over the interval $[1, 2]$, where $i = 1, \ldots, n$, $j = 1, \ldots, m_i$.

The initial state for each subject at time $t_{i0} = 0$ is generated from a Bernoulli variable with values 1 and 2 according to the stationary distribution (5.1), where transition intensities are

$$u_i = \exp\left(\beta_{u0} + \beta_{u1} z_i + \beta_{ux} x_i\right) \quad \text{and} \quad v_i = \exp\left(\beta_{v0} + \beta_{v1} z_i + \beta_{vx} x_i\right), \qquad i = 1, \ldots, n,$$

where z_i is generated from $\mathrm{N}(0, 1)$, and x_i is generated from the linear model

$$x_i = \gamma_0 + \gamma_1 z_i + e_{xi} \tag{5.7}$$

with e_{xi} generated from $\mathrm{N}\left(0, \sigma_x^2\right)$.

The time dependent surrogate measurements x_{ij}^* is generated by the measurement error model

$$x_{ij}^* = x_i + u_{ij},$$

where u_{ij} is generated from $\mathrm{N}\left(0, \sigma_u^2\right)$.

The sojourn times from state 1 to 2 and from state 2 to 1, τ_{i1j} and τ_{i2j}, are simulated from exponential distributions with mean $1/u_i$ and $1/v_i$, respectively, where the subscript j denote the jth transition. Then, the states $\{s_{ij} \colon j = 1, \ldots, m_i\}$ can be determined by the exact transition and observation times, where $i = 1, \ldots, n$.

We set $\beta_{u0} = -0.8$, $\beta_{u1} = -0.5$, $\beta_{ux} = 0.3$, $\beta_{v0} = -0.9$, $\beta_{v1} = -0.4$, $\beta_{vx} = -0.5$ in the

transition intensity model, and $\gamma_0 = 0.5$, $\gamma_1 = 1.0$ in the linear model, such that the mean sojourn times from state 1 to 2 and from state 2 to 1 are 2.04 and 5.37, respectively (see Section 5.7.1). In addition, we set $\sigma_u^2 = \sigma_x^2 = 0.5$.

5.4.2 Simulation results

We analyze the simulated data using the naive method and the proposed functional methods. The naive method is carried out by subsituting the average of time-dependent covariates, $\bar{x}_i^*$., for the true covariate x_i. In the SIMEX approach, the estimation procedure is repeated $B = 500$ times; the sequence of ξ is set to be $\{0.0, 0.5, 1.0, 1.5, 2.0\}$; the quadratic function is used in the extrapolation step. In the RC approach, we generate 500 bootstrap samples to obtain the variance estimates.

Table 5.1 summerizes the averages of biases of point estimates and their asymptotic and empirical standard errors (ASEs and ESEs), as well as coverage rates (CRs) of corresponding 95% confidence intervals. For comparison, we also display the results obtained using the true covariate x_i. The results show that functional methods perform well in finite samples, and illustrate the relatively large biases and low coverage rates yielded by the naive method. Compared with the RC method, the SIMEX method has relatively smaller biases. However, the associated ASEs in the SIMEX method are slightly underestimated compared to ESEs, thus resulting in coverage rates slightly lower than the nominal level. In the RC method, the associated ASEs agree well with their empirical counterparts; the resulting coverage rates are close to the nominal level.

5.4.3 Robustness investigation

We further investigate the robustness of the SIMEX and RC methods to illustrate that the functional methods, SIMEX and RC, do not require the correct specification of the model for X.

Table 5.1: Simulation results for the functional methods in the two-state Markov model with a time-dependent covariate

	True Covariate				Naive Method			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{u0}	.006	.079	.078	95.70	.010	.075	.074	96.05
β_{u1}	−.001	.115	.114	94.90	.073	.107	.107	88.20
β_{ux}	−.001	.095	.096	94.40	−.065	.084	.085	87.00
β_{v0}	.003	.079	.078	95.45	−.072	.076	.074	82.45
β_{v1}	−.006	.116	.114	95.40	−.095	.108	.107	86.60
β_{vx}	.002	.096	.095	95.30	.109	.084	.085	73.45
	SIMEX				Regression Calibration			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{u0}	.005	.080	.083	94.45	−.033	.083	.084	93.25
β_{u1}	.005	.120	.125	94.40	.015	.121	.121	94.55
β_{ux}	−.008	.102	.105	94.70	−.008	.102	.101	94.95
β_{v0}	−.007	.082	.085	93.35	−.007	.083	.083	94.55
β_{v1}	−.019	.120	.123	94.85	−.003	.121	.120	95.20
β_{vx}	.018	.102	.103	94.25	.018	.102	.101	94.65

The same setting as in Section 5.4.1 is used to generate the data, except the linear model (5.7) for X. Here, the error term e_{xi} in the linear model (5.7) is generated from the mixture normal distribution suggested by Li and Lin (2003)

$$\lambda \, \mathrm{N}\left\{-(1-\lambda)\,\mu, \sigma^2\right\} + (1-\lambda)\,\mathrm{N}\left(\lambda\mu, \sigma^2\right),$$

such that $E\left(e_{xi}\right) = 0$ and $\mathrm{var}\left(e_{xi}\right) = \lambda\left(1-\lambda\right)\mu^2 + \sigma^2$. We set $\lambda = 0.25$, $\mu = 1.5$, and $\sigma^2 = 0.5 - \lambda\left(1-\lambda\right)\mu^2 = 5/64$. This choice of λ, μ and σ^2 allows the distribution of e_{xi} to be bimodal (Li and Lin, 2003).

Table 5.2 summarizes simulation results based on the likelihood method using the true covariate, the naive method by replacing the true measurement with the average of surrogate measurements, the SIMEX method, and the RC method. The results show that the functional methods perform well in finite samples, when the true covariate is simulated from a mixture distribution. The SIMEX and RC method give comparable results, except that the bias of $\hat{\beta}_{u0}$ obtained using the RC method is larger than that obtained using the SIMEX method. The coverage rates in the SIMEX method are slightly lower than the nominal level due to the underestimated ASEs. In the RC method, the associated ASEs agree well with their empirical counterparts; the resulting coverage rates are close to the nominal level. The lower coverage rate of β_{u0} in the RC method is caused by the large bias in the estimate.

5.5 Maximum likelihood estimation via an Monte Carlo EM algorithm

In this section, we develop a likelihood method that yields consistent estimators under the correct model setup. Specifically, we propose an MCEM algorithm to obtain maximum likelihood

Table 5.2: Robustness investigation of the function methods for the two-state Markov model with a time-dependent covariate

	True Covariate				Naive Method			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{u0}	.007	.080	.079	95.50	.008	.076	.075	95.60
β_{u1}	.001	.111	.110	95.40	.064	.103	.102	89.45
β_{ux}	−.002	.091	.091	95.45	−.058	.079	.079	88.80
β_{v0}	.007	.080	.078	95.65	−.070	.077	.075	83.75
β_{v1}	−.001	.111	.112	95.50	−.089	.105	.107	87.65
β_{vx}	−.002	.091	.092	95.00	.104	.082	.083	74.15
	SIMEX				Regression Calibration			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{u0}	0.009	0.081	0.084	94.70	−0.040	0.084	0.083	91.25
β_{u1}	0.010	0.114	0.118	94.15	−0.001	0.118	0.116	95.55
β_{ux}	−0.013	0.096	0.098	94.35	0.004	0.099	0.098	95.05
β_{v0}	−0.002	0.083	0.085	94.85	−0.006	0.083	0.083	95.55
β_{v1}	−0.009	0.116	0.120	93.55	−0.004	0.118	0.117	95.55
β_{vx}	0.007	0.098	0.102	93.85	0.017	0.099	0.101	94.00

estimates. The true covariate X is postulated by the linear regression model

$$X = \gamma_0 + \boldsymbol{\gamma}_z^{\mathsf{T}} \mathbf{Z} + e_x, \tag{5.8}$$

where $\boldsymbol{\gamma} = \left(\gamma_0, \boldsymbol{\gamma}_z^{\mathsf{T}}\right)^{\mathsf{T}}$ is an unknown parameter vector, and e_x is independent of $U(t)$ in the measurement error model (5.4) and follow $\mathrm{N}\left(0, \sigma_x^2\right)$. We also assume that the measurement error is non-differential and the measurement error variance σ_u^2 is known. Let $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^{\mathsf{T}}, \boldsymbol{\gamma}^{\mathsf{T}}, \sigma_x^2\right)^{\mathsf{T}}$.

5.5.1 The MCEM algorithm

The complete data log-likelihood function contributed from subject i is

$$\begin{aligned}
\ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \mathbf{s}_i, \mathbf{z}_i, x_i, \mathbf{x}_i^*\right) &= \log\left\{ \Pr\left(\mathbf{s}_i, \mathbf{x}_i^*, x_i \mid \mathbf{z}_i; \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_x^2, \sigma_u^2\right) \right\} \\
&= \log\left\{ \Pr\left(\mathbf{s}_i \mid x_i, \mathbf{z}_i; \boldsymbol{\beta}\right) \right\} + \log\left\{ \Pr\left(\mathbf{x}_i^* \mid x_i; \sigma_u^2\right) \right\} + \log\left\{ \Pr\left(x_i \mid \mathbf{z}_i; \boldsymbol{\gamma}, \sigma_x^2\right) \right\} \\
&= \log\left\{ \Pr\left(s_{i0} \mid x_i, \mathbf{z}_i; \boldsymbol{\beta}\right) \right\} + \sum_{j=1}^{m_i} \log\left\{ \Pr\left(s_{ij} \mid s_{i,j-1}, x_i, \mathbf{z}_i; \boldsymbol{\beta}\right) \right\} \\
&\quad -\frac{m_i+1}{2} \log\left(2\pi\sigma_u^2\right) + \frac{1}{2\sigma_u^2} \sum_{j=0}^{m_i} \left(x_{ij}^* - x_i\right)^2 - \frac{1}{2} \log\left(2\pi\sigma_x^2\right) - \frac{\left(x_i - \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{z}_i\right)^2}{2\sigma_x^2},
\end{aligned} \tag{5.9}$$

where $\boldsymbol{z}_i = \left(1, \mathbf{z}_i^{\mathsf{T}}\right)^{\mathsf{T}}$, $\Pr\left(s_{i0} \mid x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)$ is defined to be the statinary probability $\pi_{s_{i0}}$, and $\Pr\left(s_{ij} \mid s_{i,j-1}, x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)$ is the transition probability $P_{s_{i,j-1}, s_{ij}}\left(t_{ij} - t_{i,j-1}\right)$ defined in Section 5.2.1.

The expected complete data log-likelihood at the $(k+1)$th iteration is

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right) = \sum_{i=1}^{n} E\left\{ \ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \mathbf{s}_i, \mathbf{z}_i, x_i, \mathbf{x}_i^*\right) \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2 \right\},$$

where $\boldsymbol{\theta}^{(k)}$ is the estimate of $\boldsymbol{\theta}$ at the kth iteration.

From (5.9), we can see the parameters $\boldsymbol{\beta}$ and $\left(\boldsymbol{\gamma}^{\mathsf{T}}, \sigma_x^2\right)^{\mathsf{T}}$ are disctinct from each other. The estimates of $\boldsymbol{\gamma}$ and σ_x^2 at the $(k+1)$th iteration take the form of

$$\boldsymbol{\gamma}^{(k+1)} = \left\{\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\right\}^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\mu}_k, \tag{5.10}$$

$$\begin{aligned}\left\{\sigma_x^2\right\}^{(k+1)} &= n^{-1}\sum_{i=1}^{n} E\left[\left[X_i - \left\{\boldsymbol{\gamma}^{(k+1)}\right\}^{\mathsf{T}}\boldsymbol{z}_i\right]^2 \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right] \\ &= n^{-1}\sum_{i=1}^{n}\int_{\mathcal{X}}\left[x - \left\{\boldsymbol{\gamma}^{(k+1)}\right\}^{\mathsf{T}}\boldsymbol{z}_i\right]^2 f\left\{x \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right\}\,\mathrm{d}x, \end{aligned} \tag{5.11}$$

where $\boldsymbol{Z} = \left(\boldsymbol{z}_1^{\mathsf{T}}, \ldots, \boldsymbol{z}_n^{\mathsf{T}}\right)^{\mathsf{T}}$, $\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{nk})^{\mathsf{T}}$,

$$\begin{aligned}\mu_{ik} &= E\left\{X_i \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right\} \\ &= \int_{\mathcal{X}} x \cdot f\left\{x \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right\}\,\mathrm{d}x,\end{aligned}$$

$f\left\{x \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right\}$ is the conditional probability density function of X_i given the observed data $(\mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*)$, and $\mathcal{X}$ denotes the sample space for the latent variable X_i.

The estimate of $\boldsymbol{\beta}$ at the $(k+1)$th iteration can be obtained by maximizing the function

$$\begin{aligned}Q\left(\boldsymbol{\beta}, \boldsymbol{\theta}^{(k)}\right) &= \sum_{i=1}^{n} E\left[\log\left\{\Pr\left(s_{i0} \mid x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)\right\} \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right] \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m_i} E\left[\log\left\{\Pr\left(s_{ij} \mid s_{i,j-1}, x_i, \mathbf{z}_i; \boldsymbol{\beta}\right)\right\} \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right] \\ &= \sum_{i=1}^{n}\int_{\mathcal{X}} \log\left\{\Pr\left(s_{i0} \mid x, \mathbf{z}_i; \boldsymbol{\beta}\right)\right\} f\left\{x \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right\}\,\mathrm{d}x \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\int_{\mathcal{X}} \log\left\{\Pr\left(s_{ij} \mid s_{i,j-1}, x, \mathbf{z}_i; \boldsymbol{\beta}\right)\right\} f\left\{x \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right\}\,\mathrm{d}x.\end{aligned} \tag{5.12}$$

The E step is to calculate the expected complete data log-likelihood, and the M step consists of updating $\boldsymbol{\gamma}^{(k+1)}$ and $\left\{\sigma_x^2\right\}^{(k+1)}$ as well as maximizing $Q\left(\boldsymbol{\beta}, \boldsymbol{\theta}^{(k)}\right)$ with respect to $\boldsymbol{\beta}$ to obtain the update $\boldsymbol{\beta}^{(k+1)}$. The EM algorithm iterates between the E and M steps until the convergence of the sequence $\{\boldsymbol{\theta}^{(k)}, k \geq 1\}$. Wu (1983) showed that, under regularity conditions, the sequence of values $\{\boldsymbol{\theta}^{(k)}, k \geq 1\}$ converges to maximum likelihood estimates $\hat{\boldsymbol{\theta}}$.

To perform the integration in (5.10), (5.11) and (5.12), we use the Monte Carlo method. In particular, we obtain a sample $x_{i1}^{(k)}, \ldots, x_{id}^{(k)}$ from the conditional distribution

$$f\left\{x \;\middle|\; \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right\},$$

and estimate μ_{ik} and (5.11) by the Monte Carlo sum

$$\mu_{ikd} = d^{-1} \sum_{t=1}^{d} x_{it}^{(k)} \quad \text{and} \quad \left\{\sigma_x^2\right\}_d^{(k+1)} = (nd)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{d} \left[x_{it}^{(k)} - \left\{\boldsymbol{\gamma}^{(k+1)}\right\}^{\mathsf{T}} \boldsymbol{z}_i\right]^2,$$

as well as the quantity in (5.12) by

$$\begin{aligned} Q_d\left(\boldsymbol{\beta}, \boldsymbol{\theta}^{(k)}\right) &= d^{-1} \sum_{i=1}^{n} \sum_{t=1}^{d} \log\left[\Pr\left\{s_{i0} \;\middle|\; x_{it}^{(k)}, \mathbf{z}_i; \boldsymbol{\beta}\right\}\right] \\ &\quad + d^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{t=1}^{d} \log\left[\Pr\left\{s_{ij} \;\middle|\; s_{i,j-1}, x_{it}^{(t)}, \mathbf{z}_i; \boldsymbol{\beta}\right\}\right], \end{aligned} \tag{5.13}$$

where the subscript d denotes the dependence of this estimator on the MC sample size.

By Law of Large Numbers, μ_{ikd} and both estimators in (5.11) and (5.13) converge in probability to their corresponding theoretical expectations. Then, the EM algorithm can be modified into an MCEM in which the expected complete data log-likelihood is estimated by the Monte Carlo method. In the M step, $\boldsymbol{\gamma}^{(k)}$ are obtained by substituting μ_{ikd} for μ_{ik}, $\left\{\sigma_x^2\right\}^{(k)}$ is estimated

by $\left\{\sigma_x^2\right\}_d^{(k+1)}$, and $\boldsymbol{\beta}^{(k)}$ is obtained by maximizing the Monte Carlo sum (5.13) using the Newton-Raphson method with respect to $\boldsymbol{\beta}$. More details on the convergence of an MCEM algorithm can be found in Chan and Ledolter (1995) and McCulloch (1997).

We now describe the independent Metropolis-Hastings algorithm (e.g., Robert and Casella, 2004, Section 7.4) to generate a random sample from the conditional density (target density) $f\left\{x \;\middle|\; \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}^{(k)}, \sigma_u^2\right\}$. The algorithm can be summarized as follows: given $x^{(t)}$,

1. Generate $y \sim h(y)$, where $h(y)$ is a proposal density.

2. Simulate $u \sim \text{Uniform}[0,1]$ and let

$$x^{(t+1)} = \begin{cases} y & \text{if } u \leq \min\left\{\dfrac{f\left(y \mid \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}, \sigma_u^2\right) h\left(x^{(t)}\right)}{f\left(x^{(t)} \mid \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \boldsymbol{\theta}, \sigma_u^2\right) h(y)}, 1\right\}, \\ x^{(t)} & \text{otherwise.} \end{cases}$$

To ensure the robust performance, it is recommended to chose the proposal density $h(\cdot)$ with a relatively long tail (Liu, 2001). Here, we use the Cauchy distribution with location parameter $t = \hat{\mu}_x$ and scale parameter $s = 1$ as the proposal density, where

$$\hat{\mu}_x = \left\{\sum_{i=1}^{n}(m_i+1)\right\}^{-1}\left\{\sum_{i=1}^{n}\sum_{j=0}^{m_i} x_{ij}^*\right\},$$

and the probability density function of the Cauchy distribution is

$$h(y; s, t) = \frac{1}{s\pi}\left\{1 + \left(\frac{x-t}{s}\right)^2\right\}^{-1}, \qquad -\infty < y < \infty.$$

5.5.2 Variance estimation in the MCEM algorithm

From the standard likelihood theory, under certain regularity conditions, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$ and asymptotically normally distributed:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \stackrel{d}{\rightarrow} \mathbf{N}\left\{\mathbf{0}, \mathcal{I}^{-1}\left(\boldsymbol{\theta}\right)\right\}, \qquad \text{as } n \rightarrow \infty,$$

where $\mathcal{I}\left(\boldsymbol{\theta}\right) = E\left\{-\partial^2 \ell\left(\boldsymbol{\theta}; \mathbf{s}, \mathbf{z}, \mathbf{x}^*\right) / \left(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}\right)\right\}$, and $\ell\left(\boldsymbol{\theta}; \mathbf{s}, \mathbf{z}, \mathbf{x}^*\right)$ is the log-likelihood based on the states $\mathbf{s}$, time-independent covariates $\mathbf{z}$, and time-dependent error-prone covariate $\mathbf{x}^*$ of one individual.

The variance estimation in the EM algorithm can be obtained from Louis' Formula (Louis, 1982):

$$\begin{aligned} E\left\{-\frac{\partial^2 \ell\left(\boldsymbol{\theta}; \mathbf{s}, \mathbf{z}, \mathbf{x}^*\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\right\} &= E\left\{-\frac{\partial^2 \ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \mathbf{s}, \mathbf{z}, x, \mathbf{x}^*\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}} \,\middle|\, \mathbf{s}, \mathbf{z}, \mathbf{x}^*; \boldsymbol{\theta}, \sigma_u^2\right\} \\ &\quad - \operatorname{var}\left\{\frac{\partial \ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \mathbf{s}, \mathbf{z}, x, \mathbf{x}^*\right)}{\partial\boldsymbol{\theta}} \,\middle|\, \mathbf{s}, \mathbf{z}, \mathbf{x}^*; \boldsymbol{\theta}, \sigma_u^2\right\}, \end{aligned}$$

where $\ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \mathbf{s}, \mathbf{z}, x, \mathbf{x}^*\right)$ is the complete data log-likelihood of one individual.

For the MCEM algorithm, the Monte Carlo evaluation of Louis' Formula can be divided into two parts:

$$\widehat{E}_d\left\{\frac{\partial^2 \ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \mathbf{s}, \mathbf{z}, x, \mathbf{x}^*\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}} \,\middle|\, \mathbf{s}, \mathbf{z}, \mathbf{x}^*; \boldsymbol{\theta}, \sigma_u^2\right\} = (nd)^{-1}\left\{\sum_{i=1}^{n}\sum_{t=1}^{d} \frac{\partial^2 \ell_{it}^{\mathrm{c}}\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right\},$$

and

$$\widehat{\operatorname{var}}_d\left\{\frac{\partial \ell^{\mathrm{c}}\left(\boldsymbol{\theta}; \mathbf{s}, \mathbf{z}, x, \mathbf{x}^*\right)}{\partial\boldsymbol{\theta}} \,\middle|\, \mathbf{s}, \mathbf{z}, \mathbf{x}^*; \boldsymbol{\theta}, \sigma_u^2\right\}$$

$$= d^{-1}\left[\sum_{t=1}^{d}\left[n^{-1}\left\{\sum_{i=1}^{n}\frac{\partial \ell_{it}^{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right\} - (nd)^{-1}\left\{\sum_{i=1}^{n}\sum_{t=1}^{d}\frac{\partial \ell_{it}^{c}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right\}\right]^{\otimes 2}\right],$$

where $\ell_{it}^{c}(\boldsymbol{\theta}) = \ell^{c}(\boldsymbol{\theta}; \mathbf{s}_i, \mathbf{z}_i, x_{it}, \mathbf{x}_i^*)$ is the complete data log-likelihoohd from the ith subject with x_{it}, generated from the conditional density $f\left\{x \,\middle|\, \mathbf{s}_i, \mathbf{z}_i, \mathbf{x}_i^*; \hat{\boldsymbol{\theta}}, \sigma_u^2\right\}$, $i = 1, \ldots, n$, $t = 1, \ldots, d$, and the subscript d denotes the dependence of the estimators on the Monte Carlo sample size.

5.5.3 Simulation results

Table 5.3 summarizes the averages of biases of point estimates and their asymptotic and empirical standard errors (ASEs and ESEs), along with coverage rates (CRs) of corresponding 95% confidence intervals for the likelihood method. The simulation setting is the same as Section 5.4. In the MCEM algorithm, the first 2000 samples are thrown away and the Monte Carlo sample size is set to be $d = 3000$.

The biases in the proposed MLEs of $\boldsymbol{\beta}$ are relatively small; the associated ASEs are slightly less than their empirical counterparts; the resulting coverage rates are slightly lower than the nominal level. However, the biases in the MLEs of $\boldsymbol{\gamma}$ and σ_x^2 are slightly larger; although the ASEs agree well with ESEs, the coverage rates are less than the nominal level.

5.6 Discussion

In this chapter, we develop estimation procedures for the analysis of panel data with time-dependent surrogate measurements under the two-state Markov model. The panel data of each subject consist of a binary outcome, time-dependent surrogate measurements, and time-independent covariates. The data are under intermittent observation and thus the exact transition times are interval censored. Therefore, the two-state continuous-time Markov model is

Table 5.3: Simulation results for the likelihood method under the two-state Markov model with a time-dependent covariate

	True Covariate				Naive Method			
	Bias	ASE	ESE	CR%	Bias	ASE	ESE	CR%
β_{u0}	.006	.079	.078	95.70	.010	.075	.074	96.05
β_{u1}	−.001	.115	.114	94.90	.073	.107	.107	88.20
β_{ux}	−.001	.095	.096	94.40	−.065	.084	.085	87.00
β_{v0}	.003	.079	.078	95.45	−.072	.076	.074	82.45
β_{v1}	−.006	.116	.114	95.40	−.095	.108	.107	86.60
β_{vx}	.002	.096	.095	95.30	.109	.084	.085	73.45

	MLE			
	Bias	ASE	ESE	CR%
β_{u0}	−.009	.082	.083	95.10
β_{u1}	.009	.115	.122	92.80
β_{ux}	.003	.100	.108	93.40
β_{v0}	.015	.082	.086	93.95
β_{v1}	−.029	.115	.125	93.25
β_{vx}	.002	.100	.108	93.45

used for the analysis. The time-dependent surrogate measurements are treated as surrogates of the unobserved true covariate. The additive measurement error model is used to describe the relationship between the surrogates and the unobserved true covariate.

First, we explore two functional methods, simulation extrapolation and regression calibration. They do not require the distribution assumption on the unobserved true covariate and provide convenient solutions to reduce the biases due to measurement error. The simulation studies show that both methods perform well in finite samples. Furthermore, we develop the maximum likelihood estimation via an Monte Carlo EM algorithm. The linear regression model is used to model the relationship between the unobserved true covariate and time-independent covariates. The independent Metropolis-Hastings algorithm is used in the Monte Carlo sample generation. The simulation studies show that the biases of parameters in the transition intensity model are relatively small, and their coverage rates are slightly less than the nominal level due to the underestimated standard errors. However, the biases of parameters in the linear model for the covariate are slightly larger and their coverage rates are poor.

In future, we will consider a two-stage method based on the likelihood to reduce the biases of parameters in the linear model and provide consistent estimates. In the first stage, we consider the following linear model

$$X^*(t) = \gamma_0 + \boldsymbol{\gamma}_z^{\mathsf{T}}\mathbf{Z} + e^*(t)$$

where

$$e^*(t) = e_x + U(t) \quad \text{and} \quad e^*(t) \sim \mathrm{N}\left(0, \sigma^2 = \sigma_x^2 + \sigma_u^2\right).$$

The standard linear regression model can be fitted to obtain $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}_x^2 = \hat{\sigma}^2 - \sigma_u^2$, where σ_u^2 is assumed to be known. In the second stage, we perform the maximum likelihood estimation procedure described in Section 5.5 to obtain $\hat{\boldsymbol{\beta}}$ by assuming the known $\boldsymbol{\gamma}$, σ_x^2 and σ_u^2.

In the MCEM algorithm, it is inefficient to start with a large number of Monte Carlo samples, when $\boldsymbol{\theta}^{(k)}$ at the first few iterations may be far from the true value. Therefore, it is recommended to increase the Monte Carlo sample size as the EM algorithm iterates, in order that the previous and current updates can be distinguished from the Monte Carlo error (Wei and Tanner, 1990). Booth and Hobert (1999) suggested a rule for automatically increasing the Monte Carlo sample size after iterations when random samples are directly generated from the target distribution or by the importance weighted sampling from a candidate distribution "close" to the target distribution. Levine and Casella (2001) presented another method based on central limit theorems for increasing the Monte Carlo sample size, when random samples are obtained via Markov chain Monte Carlo techniques. To improve the efficiency of our MCEM algorithm, it will be interesting to develop a method to update the Monte Carlo sample size at each iteration.

5.7 Technical details

5.7.1 Effects of parameters in simulation studies

The mean sojourn time in state i conditional on the time-independent perfectly measured covariate z is

$$\begin{aligned}
E(\tau_i \mid z) &= E\left\{\exp\left(-\beta_{i0} - \beta_{ix}X - \beta_{iz}Z\right) \mid z\right\} \\
&= E\left[\exp\left\{-\beta_{i0} - \beta_{ix}\left(\gamma_0 + \gamma_z Z + \varepsilon\right) - \beta_{iz}Z\right\} \mid z\right] \\
&= \exp\left\{-\beta_{i0} - \beta_{ix}\gamma_0 - \left(\beta_{ix}\gamma_z + \beta_{iz}\right)z\right\} E\left\{\exp\left(-\beta_{ix}\varepsilon\right)\right\} \\
&= \exp\left\{-\beta_{i0} - \beta_{ix}\gamma_0 - \left(\beta_{ix}\gamma_z + \beta_{iz}\right)z\right\} \exp\left(\beta_{ix}^2/2\right) \\
&= \exp\left\{-\beta_{i0} - \beta_{ix}\gamma_0 + \frac{\beta_{ix}^2}{2} - \left(\beta_{ix}\gamma_z + \beta_{iz}\right)Z\right\}.
\end{aligned}$$

By law of total expectation, the mean sojourn time in state i is

$$\begin{aligned} E(\tau_i) &= E\{E(\tau_i \mid Z)\} \\ &= E\left[\exp\left\{-\beta_{i0} - \beta_{ix}\gamma_0 + \frac{\beta_{ix}^2}{2} - (\beta_{ix}\gamma_z + \beta_{iz})Z\right\}\right] \\ &= \exp\left(-\beta_{i0} - \beta_{ix}\gamma_0 + \frac{\beta_{ix}^2}{2}\right) E\{-(\beta_{ix}\gamma_z + \beta_{iz})Z\} \\ &= \exp\left(-\beta_{i0} - \beta_{ix}\gamma_0 + \frac{\beta_{ix}^2}{2}\right) \exp\left\{\frac{(\beta_{ix}\gamma_z + \beta_{iz})^2}{2}\right\} \\ &= \exp\left\{-\beta_{i0} - \beta_{ix}\gamma_0 + \frac{\beta_{ix}^2}{2} + \frac{(\beta_{ix}\gamma_z + \beta_{iz})^2}{2}\right\}. \end{aligned}$$

5.7.2 Derivatives of transition probabilities in two-state Markov models

The first derivations of the logarithm of stationary probabilities are as follows

$$\frac{\partial \log \pi_1}{\partial \beta_{u0}} = -\frac{\partial \log \pi_1}{\partial \beta_{v0}} = -\frac{u}{u+v} \quad \text{and} \quad \frac{\partial \log \pi_2}{\partial \beta_{u0}} = -\frac{\partial \log \pi_2}{\partial \beta_{v0}} = \frac{v}{u+v}.$$

The second derivations of the logarithm of stationary probabilities are as follows

$$\frac{\partial^2 \log \pi_1}{\partial \beta_{u0}^2} = \frac{\partial^2 \log \pi_1}{\partial \beta_{v0}^2} = \frac{\partial^2 \log \pi_2}{\partial \beta_{u0}^2} = \frac{\partial^2 \log \pi_2}{\partial \beta_{v0}^2} = -\frac{uv}{(u+v)^2},$$

and

$$\frac{\partial^2 \log \pi_1}{\partial \beta_{u0} \partial \beta_{v0}} = \frac{\partial^2 \log \pi_2}{\partial \beta_{u0} \partial \beta_{v0}} = \frac{uv}{(u+v)^2}.$$

The first and second derivations of the logarithm of transition probabilities from state 1 to 1

are as follows

$$\frac{\partial \log P_{11}(t)}{\partial \beta_{u0}} = -u\left[\frac{1}{u+v}+\frac{ut-1}{u+v\exp\{(u+v)t\}}\right],$$

$$\frac{\partial \log P_{11}(t)}{\partial \beta_{v0}} = u\left[\frac{1}{u+v}-\frac{vt+1}{u+v\exp\{(u+v)t\}}\right],$$

$$\frac{\partial^2 \log P_{11}(t)}{\partial \beta_{u0}^2} = u\left[\frac{u(ut-1)\left[1+vt\exp\{(u+v)t\}\right]}{\left[u+v\exp\{(u+v)t\}\right]^2}-\frac{v}{(u+v)^2}-\frac{2ut-1}{u+v\exp\{(u+v)t\}}\right],$$

$$\frac{\partial^2 \log P_{11}(t)}{\partial \beta_{v0}^2} = uv\left[\frac{\exp\{(u+v)t\}(vt+1)^2}{\left[u+v\exp\{(u+v)t\}\right]^2}-\frac{1}{(u+v)^2}-\frac{t}{u+v\exp\{(u+v)t\}}\right],$$

and $$\frac{\partial^2 \log P_{11}(t)}{\partial \beta_{u0}\partial \beta_{v0}} = uv\left[\frac{1}{(u+v)^2}+\frac{\exp\{(u+v)t\}(ut-1)(vt+1)}{\left[u+v\exp\{(u+v)t\}\right]^2}\right].$$

The first and second derivations of the logarithm of transition probabilities from state 1 to 2 are as follows

$$\frac{\partial \log P_{12}(t)}{\partial \beta_{u0}} = \frac{ut}{\exp\{(u+v)t\}-1}+\frac{v}{u+v},$$

$$\frac{\partial \log P_{12}(t)}{\partial \beta_{v0}} = v\left[\frac{t}{\exp\{(u+v)t\}-1}-\frac{1}{u+v}\right],$$

$$\frac{\partial^2 \log P_{12}(t)}{\partial \beta_{u0}^2} = u\left[t\left[\frac{1}{\exp\{(u+v)t\}-1}+\frac{ut}{2-2\cosh\{(u+v)t\}}\right]-\frac{v}{(u+v)^2}\right],$$

$$\frac{\partial^2 \log P_{12}(t)}{\partial \beta_{v0}^2} = v\left[t\left[\frac{1}{\exp\{(u+v)t\}-1}+\frac{vt}{2-2\cosh\{(u+v)t\}}\right]-\frac{u}{(u+v)^2}\right],$$

$$\text{and} \qquad \frac{\partial^2 \log P_{12}(t)}{\partial \beta_{u0} \beta_{v0}} = uv\left[\frac{1}{(u+v)^2} + \frac{t^2}{2-2\cosh\{(u+v)t\}}\right].$$

The first and second derivations of the logarithm of transition probabilities from state 2 to 1 are as follows

$$\begin{aligned}
\frac{\partial \log P_{21}(t)}{\partial \beta_{u0}} &= u\left[\frac{t}{\exp\{(u+v)t\}-1} - \frac{1}{u+v}\right], \\
\frac{\partial \log P_{21}(t)}{\partial \beta_{v0}} &= \frac{vt}{\exp\{(u+v)t\}-1} + \frac{u}{u+v}, \\
\frac{\partial^2 \log P_{21}(t)}{\partial \beta_{u0}^2} &= \frac{\partial^2 \log P_{12}(t)}{\partial \beta_{u0}^2}, \\
\frac{\partial^2 \log P_{21}(t)}{\partial \beta_{v0}^2} &= \frac{\partial^2 \log P_{12}(t)}{\partial \beta_{v0}^2}, \\
\text{and} \qquad \frac{\partial^2 \log P_{21}(t)}{\partial \beta_{u0} \beta_{v0}} &= \frac{\partial^2 \log P_{12}(t)}{\partial \beta_{u0} \beta_{v0}}
\end{aligned}$$

The first and second derivations of the logarithm of transition probabilities from state 2 to 2 are as follows

$$\begin{aligned}
\frac{\partial \log P_{22}(t)}{\partial \beta_{u0}} &= v\left[\frac{1}{u+v} - \frac{ut+1}{v+u\exp\{(u+v)t\}}\right], \\
\frac{\partial \log P_{22}(t)}{\partial \beta_{v0}} &= -v\left[\frac{1}{u+v} + \frac{vt-1}{v+u\exp\{(u+v)t\}}\right], \\
\frac{\partial^2 \log P_{22}(t)}{\partial \beta_{u0}^2} &= uv\left[\frac{\exp\{(u+v)t\}(ut+1)^2}{\left[v+u\exp\{(u+v)t\}\right]^2} - \frac{1}{(u+v)^2} - \frac{t}{v+u\exp\{(u+v)t\}}\right],
\end{aligned}$$

$$\frac{\partial^2 \log P_{22}(t)}{\partial \beta_{v0}^2} = v\left[\frac{v(vt-1)\left[1+ut\exp\{(u+v)t\}\right]}{\left[v+u\exp\{(u+v)t\}\right]^2} - \frac{u}{(u+v)^2} - \frac{2vt-1}{v+u\exp\{(u+v)t\}}\right],$$

and
$$\frac{\partial^2 \log P_{22}(t)}{\partial \beta_{u0}\partial \beta_{v0}} = uv\left[\frac{1}{(u+v)^2} + \frac{\exp\{(u+v)t\}(ut+1)(vt-1)}{\left[v+u\exp\{(u+v)t\}\right]^2}\right].$$

Chapter 6

Summary

This thesis focuses on analyzing the longitudinal data under panel observation from disease progression studies. The observation times are irregularly spaced in such sampling scheme. Another feature is that the exact times of transitions are interval censored. The aim of the study is to estimate transition rates and understand risk factor influences on transitions.

There are several challenges in the analysis of panel data. The first one is state misclassification. It may arise from poor quality of a diagnostic test, the impossibility of the accurate assessment, or the reading error. The non-homogeneity of the data is another common issue. In addition, it is not necessarily realistic that transition intensities stay constant through time. Last but not least, the covariates are subject to measurement/classification error and may be time-dependent.

This thesis consists of four projects. In the first project, we consider disease states subject to misclassification and focus on progressive models. We derive three conditions for non-informative sampling scheme. In the likelihood method, the EM algorithm with unobserved true states treated as latent variables is used to obtain maximum likelihood estimates. However, the likelihood

method relies on the validity of models assumptions, and the output independence assumption may not hold in practice. To overcome the difficulty induced by the likelihood method, we propose the pairwise likelihood method. The conditions for the non-informative sampling scheme are derived in the pairwise likelihood formulation. The EM algorithm is straightforwardly extended to maximize the pairwise likelihood in progressive Markov models with misclassified states. The performance of estimation procedures is evaluated by simulation studies. The proposed progressive model is illustrated on coronary allograft vasculopathy data, in which the diagnosis based on the coronary angiography is subject to error.

The second project is analysis of mover-stayer models with misclassified states. The state misclassification is extended to a generic setting: discrete-valued surrogates are observed for true states. In this project, we consider one particular type of non-homogeneity when the population consists of two subpopulations. The stayer stays in the initial state, while the mover evolves according to a continuous-time Markov process. We propose hidden Markov models to facilitate heterogeneity for a population and to simultaneously account for state misclassification. The likelihood inference procedure based on the EM algorithm, which treats the mover-stayer indicator and underlying states as latent variables, is developed for the proposed model to make statistical inference. The performance of the likelihood method is investigated through simulation studies. The proposed method is applied to analyze the data arising from the Waterloo Smoking Prevention Project.

In the third project, we investigate the covariate misclassification in the analysis of piecewise-constant Markov models. In the piecewise constant framework, transition intensities are assumed to be constant within each pre-specified interval. We show that the joint model for the state process and reclassification process is not identifiable. To estimate parameters in the transition intensity model, we propose the likelihood methods for two practical situations: one is that

parameters in reclassification probabilities are known from empirical studies; the other is in the presence of an internal or external validation data. Simulation studies are carried out to evaluate the performance of the proposed MLEs. Our proposed methods are applied to analyze the data arising from the psoriatic arthritic (PsA) study.

The fourth project is statistical inference of two-state Markov models for panel data with time-dependent surrogate covariates. First, we introduce two functional modelling approaches, SIMEX method and regression calibration. In these approaches, no distributional assumption is made on the true covariate X. Although functional approaches enjoy the easy implementation and reduce the effects of measurement error, they are approximation methods and do not yield the consistent estimators. Therefore, we propose an Monte Carlo EM algorithm to obtain the MLEs which is consistent under the correct model setup. The performance of proposed methods is investigated based on simulation studies.

References

Aalen, O. O., Farewell, V. T., de Angelis, D., Day, N. E., and Nöel Gill, O. (1997). A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. *Statistics in Medicine*, 16(19):2191–2210.

Agresti, A. (2002). *Categorical Data Analysis*. New Jersey : Wiley-Interscience, 2nd edition.

Alioum, A., Commenges, D., Thiébaut, R., and Dabis, F. (2005). A multistate approach for estimating the incidence of human immunodeficiency virus by using data from a prevalent cohort study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(4):739–752.

Altman, R. M. and Petkau, A. J. (2005). Application of hidden Markov models to multiple sclerosis lesion count data. *Statistics in Medicine*, 24(15):2335–2344.

Andersen, P. K. (1988). Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Statistics in Medicine*, 7(6):661–670.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer series in statistics. New York : Springer-Verlag.

Andersen, P. K., Hansen, L. S., and Keiding, N. (1991). Assessing the influence of reversible disease indicators on survival. *Statistics in Medicine*, 10(7):1061–1067.

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

Bai, Y., Song, P. X.-K., and Raghunathan, T. E. (2012). Joint composite estimating functions in spatiotemporal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(5):799–824.

Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37(6):1554–1563.

Baum, L. E., 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.

Baum, L. E. and Sell, G. R. (1968). Growth transformations for functions on manifolds. *Pacific Journal of Mathematics*, 27(2):211–227.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, 45(250):164–180.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195.

Binquet, C., Teuff, G. L., Abrahamovicz, M., Mahboubi, A., Yazdanpanah, Y., D. Rey, C. R., Chirouze, C., Berger, J. L., Faller, J. P., Chavanet, P., Quantin, C., and Piroth, L. (2009). Markov modelling of HIV infection evolution in the HAART era. *Epidemiology and Infection*, 137(9):1272–1282.

Blackwell, D. (1985). Approximate normality of large products. Department of Statistics Technical Report 54, University of California Berkeley.

Blumen, I., Kogan, M., and McCarthy, P. J. (1955). *The Industrial Mobility of Labor as a Probability Proces*, volume 6 of *Cornell studies in industrial and labor relations*. Ithaca, Cornell University.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285.

Bross, I. (1954). Misclassification in 2 × 2 tables. *Biometrics*, 10(4):478–486.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Interdisciplinary statistics. Boca Raton : CRC Press.

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):441–462.

Cameron, R., Brown, K. S., Best, J. A., Pelkman, C. L., Madill, C. L., Manske, S. R., and Payne, M. E. (1999). effectiveness of a social influences smoking prevention program as a function of provider type, training method, and school risk. *American Journal of Public Health*, 89(12):1827–1831.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics. New York : Springer.

Caragea, P. C. and Smith, R. L. (2007). Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *Journal of Multivariate Analysis*, 98(7):1417–1440.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, Florida : Chapman & Hall/CRC, 2nd edition.

Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85(411):652–663.

Castro, R., Coates, M., Liang, G., Nowak, R., and Yu, B. (2004). Network tomography: recent developments. *Statistical Science*, 19(3):499–517.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429):242–252.

Chandler, R. E. and Bate, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika*, 94(1):167–183.

Chang, C.-M., Lin, W.-C., Kuo, H.-S., Yen, M.-F., and Chen, T. H.-H. (2007). Estimation and prediction system for multi-state disease process: application to analysis of organized screening regime. *Journal of Evaluation in Clinical Practice*, 13(6):867–881.

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):1175–1189.

Chen, B., Yi, G. Y., and Cook, R. J. (2011). Progressive multi-state models for informatively incomplete longitudinal data. *Journal of Statistical Planning and Inference*, 141(1):80–93.

Chen, H. H., Duffy, S. W., and Tabar, L. (1996). A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(3):307–317.

Chen, H. H., Duffy, S. W., and Tabar, L. (1997). A mover-stayer mixture of Markov chain models for the assessment of dedifferentiation and tumour progression in breast cancer. *Journal of Applied Statistics*, 24(3):265–278.

Chen, P.-L. and Sen, P. K. (1999). A piecewise transition model for analyzing multistate life history data. *Journal of Statistical Planning and Inference*, 78(1-2):385–400.

Chen, T. H.-H., Kuo, H.-S., Yen, M.-F., Lai, M. S., Tabar, L., and Duffy, S. W. (2000). Estimation of sojourn time in chronic disease screening without data on interval cases. *Biometrics*, 56(1):167–172.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics*, 25(3):573–578.

Chiang, C.-l. (1980). *An Introduction to Stochastic Processes and Their Application*. Huntington, N.Y. : Krieger.

Cochran, W. G. (1968). Errors of measurement in Statistics. *Technometrics*, 10(4):637–666.

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

Commenges, D. (2002). Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research*, 11(2):167–182.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.

Cook, R. J., Kalbfleisch, J. D., and Yi, G. Y. (2002). A generalized mover-stayer model for panel data. *Biostatistics*, 3(3):407–420.

Cook, R. J., Ng, E. T. M., Mukherjee, J., and Vaughan, D. (1999). Two-state mixed renewal processes for chronic disease. *Statistics in Medicine*, 18(2):175–188.

Cook, R. J., Yi, G. Y., Lee, K.-A., and Gladman, D. D. (2004). A conditional Markov model for clustered progressive multistate processes under incomplete observation. *Biometrics*, 60(2):436–443.

Cook, R. J., Zeng, L., and Lee, K.-A. (2008). A multistate model for bivariate interval-censored failure time data. *Biometrics*, 64(4):1100–1109.

Couto, E., Duffy, S. W., Ashton, H. A., Walker, N. M., Myles, J. P., Scott, R. A. P., and Thompson, S. G. (2002). Probabilities of progression of aortic aneurysms: estimates and implications for screening policy. *Journal of Medical Screening*, 9(1):40–42.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. London : Methuen.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737.

Curriero, F. C. and Lele, S. (1999). A composite likelihood approach to semivariogram estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, 4(1):9–28.

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

Dennis, Jr., J. E. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs : Prentice-Hall.

Devanarayan, V. and Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics & Probability Letters*, 59(3):219–225.

Duffy, S. W., Chen, H.-H., Tabar, L., and Day, N. E. (1995). Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Statistics in Medicine*, 14(14):1531–1543.

Espeland, M. A. and Hui, S. L. (1987). A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics*, 43(4):1001–1012.

Faddy, M. J. (1976). A note on the general time-dependent stochastic compartmental model. *Biometrics*, 32(2):443–448.

Farewell, V. T. and Su, L. (2011). A multistate model for events defined by prolonged observation. *Biostatistics*, 12(1):102–111.

Frank, E., Kupfer, D. J., Perel, J. M., Cornes, C., Jarrett, D. B., Mallinger, A. G., Thase, M. E., McEachran, A. B., and Grochocinski, V. J. (1990). Three-year outcomes for maintenance therapies in recurrent depression. *Archives of General Psychiatry*, 47(12):1093–1099.

Frydman, H. (1984). Maximum likelihood estimation in the mover-stayer model. *Journal of the American Statistical Association*, 79(387):632–638.

Frydman, H. and Kadam, A. (2004). Estimation in the continuous time mover-stayer model with an application to bond ratings migration. *Applied Stochastic Models in Business and Industry*, 20(2):155–170.

Fuchs, C. and Greenhouse, J. B. (1988). The EM algorithm for maximum likelihood estimation in the mover-stayer model. *Biometrics*, 44(2):605–613.

Fujii, Y. and Yanagimoto, T. (2005). Pairwise conditional score functions: a generalization of the Mantel-Haenszel estimator. *Journal of Statistical Planning and Inference*, 128(1):1–12.

Gao, X. and Song, P. X.-K. (2011). Composite likelihood EM algorithm with applications to multivariate hidden Markov model. *Statistica Sinica*, 21(1):165–185.

Gentleman, R. C., Lawless, J. F., Lindsey, J. C., and Yan, P. (1994). Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13(8):805–821.

Gladman, D. D., Farewell, V. T., and Nadeau, C. (1995). Clinical indicators of progression in psoriatic arthritis: multivariate relative risk model. *Journal of Rheumatology*, 22(4):675–679.

Gleser, L. J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In Brown, P. J. and Fuller, W. A., editors, *Statistical Analysis of Measurement Error Models and Applications*, volume 112. Providence, R.I.: American Mathematical Society.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31(4):1208–1211.

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

Guan, Y. (2006). A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association*, 101(476):1502–1512.

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

Hanfelt, J. J. (2004). Composite conditional likelihood for sparse clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):259–273.

He, W. and Lawless, J. F. (2003). Flexible maximum likelihood methods for bivariate proportional hazards models. *Biometrics*, 59(4):837–848.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443):1099–1111.

Henderson, R. and Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika*, 90(2):355–366.

Hjort, N. L. and Omre, H. (1994). Topics in spatial statistics [with discussion, comments and rejoinder]. *Scandinavian Journal of Statistics*, 21(4):289–357.

Hjort, N. L. and Varin, C. (2008). ML, PL, QL in markov chain models. *Scandinavian Journal of Statistics*, 35(1):64–82.

Hsieh, H.-J., Chen, T. H.-H., and Chang, S.-H. (2002). Assessing chronic disease progression using non-homogeneous exponential regression Markov models: an illustration using a selective breast cancer screening in Taiwan. *Statistics in Medicine*, 21(22):3369–3382.

Hubbard, R. A., Inoue, L. Y. T., and Fann, J. R. (2008). Modeling nonhomogeneous Markov processes via time transformation. *Biometrics*, 64(3):843–850.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. University of California Press.

Jackson, C. H. (2011). Multi-state models for panel data: the *msm* package for `R`. *Journal of Statistical Software*, 38(8):1–28.

Jackson, C. H. and Sharples, L. D. (2002). Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*, 21(1):113–128.

Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W., and Couto, E. (2003). Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209.

Jamshidian, M. and Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):257–270.

Jennrich, R. I. and Bright, P. B. (1976). Fitting systems of linear differential equations using computer generated exact derivatives. *Technometrics*, 18(4):385–392.

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

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.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Hoboken, N.J. : Wiley, 2nd edition.

Kang, M. and Lagakos, S. W. (2007). Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics*, 8(2):252–264.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 42(4):855–865.

Kirby, A. J. and Spiegelhalter, D. J. (1994). Modeling the precursors of cervical cance. In Lange, N. T., Ryan, L. M., Billard, L., Brillinger, D. R., Conquest, L. L., and Greenhouse, J. B., editors, *Case Studies in Biometry*, Wiley series in probability and mathematical statistics, chapter 18, pages 359–384. New York : Wiley.

Klein, J. P., Klotz, J. H., and Grever, M. R. (1984). A biological marker model for predicting disease transitions. *Biometrics*, 40(4):927–936.

Klotz, J. H. and Sharpless, L. D. (1994). Estimation for a Markov heart transplant model. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 43(3):431–438.

Kosorok, M. R. and Chao, W.-H. (1995). Further details on the analysis of longitudinal ordinal response data in continuous time. Department of Biostatistics Technical Report 92, University of Wisconsin-Madison.

Kosorok, M. R. and Chao, W.-H. (1996). The analysis of longitudinal ordinal response data in continuous time. *Journal of the American Statistical Association*, 91(434):807–817.

Küchenhoff, H., Lederer, W., and Lesaffre, E. (2007). Asymptotic variance estimation for the misclassification simex. *Computational Statistics & Data Analysis*, 51(12):6197–6211.

Küchenhoff, H., Mwalili, S. M., and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*, 62(1):85–96.

Kuha, J., Skinner, C., and Palmgren, J. (2005). Misclassification error. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. Chichester, West Sussex, England ; Hoboken, NJ : Wiley, 2nd edition.

Kuk, A. Y. and Nott, D. J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics & Probability Letters*, 47(4):329–335.

Larribe, F. and Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica*, 21(1):43–69.

Lele, S. and Taper, M. L. (2002). A composite likelihood approach to (co)variance components estimation. *Journal of Statistical Planning and Inference*, 103(1-2):117–135.

Leroux, B. G. and Puterman, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, 48(2):545–558.

Levine, R. A. and Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439.

Li, H. and Yi, G. Y. (2013a). Estimation methods for marginal and association parameters for longitudinal binary data with nonignorable missing observations. *Statistics in Medicine*, 32(5):833–848.

Li, H. and Yi, G. Y. (2013b). A pairwise likelihood approach for longitudinal data with missing observations in both response and covariates. *Computational Statistics & Data Analysis*, 68(10):66–81.

Li, Y. and Lin, X. (2003). Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. *Journal of the American Statistical Association*, 98(461):191–203.

Li, Y. and Lin, X. (2006). Semiparametric normal transformation models for spatially correlated survival data. *Journal of the American Statistical Association*, 101(474):591–603.

Liang, G. and Yu, B. (2003). Maximum pseudo likelihood estimation in network tomography. *IEEE Transactions on Signal Processing*, 51(8):2043–2053.

Liang, K.-Y. (1987). Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models. *Biometrics*, 43(2):289–299.

Lindsay, B. G. (1988). Composite likelihood methods. In Prabhu, N. U., editor, *Statistical Inference from Stochastic Processes* (Ithaca, NY, 1987), volume 80 of *Contemporary Mathematics*, pages 221–239. Providence, Rhode Island : American Mathematical Society.

Lindsay, B. G., Yi, G. Y., and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21(1):71–105.

Lindsey, J. C. and Ryan, L. M. (1993). A three-state multiplicative model for rodent tumorigenicity experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 42(2):283–300.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer series in statistics. New York : Springer.

Longini, I. M., Clark, W. S., Byers, R. H., Ward, J. W., Darrow, W. W., Lemp, G. F., and Hethcote, H. W. (1989). Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine*, 8(7):831–843.

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

Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H. (2008). A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics*, 36(1):99–109.

Marshall, G. and Jones, R. H. (1995). Multistate models and diabetic retinopathy. *Statistics in Medicine*, 14(18):1975–1983.

Marshall, R. J. (1990). Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*, 43(9):941–947.

Mathieu, E., Loup, P., Dellamonica, P., and Daures, J. P. (2005). Markov modelling of immunological and virological states in HIV-1 infected patients. *Biometrical Journal*, 47(6):834–846.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437):162–170.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Series in Statistics. Geert Verbeke. IMPRINT New York : Springer.

Moler, C. and van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49.

Morrissey, M. J. and Spiegelman, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics*, 55(2):338–344.

Nagelkerke, N. J. D., Chunge, R. N., and Kinoti, S. N. (1990). Estimation of parasitic infection dynamics when detectability is imperfect. *Statistics in Medicine*, 9(10):1211–1219.

Nott, D. J. and Rydén, T. (1999). Pairwise likelihood methods for inference in image models. *Biometrika*, 86(3):661–676.

Ocañ-Riola, R. (2005). Non-homogeneous Markov processes for biomedical data analysis. *Biometrical Journal*, 47(3):369–376.

O'Keeffe, A. G., Tom, B. D. M., and Farewell, V. T. (2011). A case-study in the clinical epidemiology of psoriatic arthritis: multistate models and causal arguments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(5):675–699.

O'Keeffe, A. G., Tom, B. D. M., and Farewell, V. T. (2013). Mixture distributions in multistate modelling: Some considerations in a study of psoriatic arthritis. *Statistics in Medicine*, 32(4):600–619.

Omar, R., Stallard, N., and Whitehead, J. (1995). A parametric multistate model for the analysis of carcinogenicity experiments. *Lifetime Data Analysis*, 1(4):327–346.

Parner, E. T. (2001). A composite likelihood approach to multivariate survival data. *Scandinavian Journal of Statistics*, 28(2):295–302.

Parzen, M., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Troxel, A., and Molenberghs, G. (2007). Pseudo-likelihood methods for the analysis of longitudinal binary data subject to nonignorable non-monotone missingness. *Journal of Data Science*, 5(1):1–21.

Pearson, K. (1902). On the mathematical theory of errors of judgment, with special reference to the personal equation. *Philosophical Transactions of the Royal Society of London. Series A*, 198:235–299.

Pérez-Ocón, R., Ruiz-Castro, J. E., and Gámiz-Pérez, M. L. (2001). Non-homogeneous Markov models in the analysis of survival after breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(1):111–124.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*. New York : Cambridge University Press, 3rd edition.

Pyke, R. (1961). Markov renewal processes with finitely many states. *Annals of Mathematical Statistics*, 32(4):1243–1259.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Renard, D., Molenberghs, G., and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis*, 44(4):649–667.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer texts in statistics. New York : Springer, 2nd edition.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8(9):1051–1069.

Rosychuk, R. J. and Islam, S. (2009). Parameter estimation in a model for misclassified Markov data – a Bayesian approach. *Computational Statistics & Data Analysis*, 53(11):3805–3816.

Rosychuk, R. J., Sheng, X., and Stuber, J. L. (2006). Comparison of variance estimation approaches in a two-state Markov model for longitudinal data with misclassification. *Statistics in Medicine*, 25(11):1906–1921.

Rosychuk, R. J. and Thompson, M. E. (2003). Bias correction of two-state latent Markov process parameter estimates under misclassification. *Statistics in Medicine*, 22(12):2035–2055.

Rosychuk, R. J. and Thompson, M. E. (2004). Parameter identifiability issues in a latent markov model for misclassified binary responses. *Journal of the Iranian Statistical Society*, 3(1):39–57.

Royall, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, 54(2):221–226.

Saint-Pierre, P., Combescure, C., Daurès, J. P., and Godard, P. (2003). The analysis of asthma control under a Markov assumption with use of covariates. *Statistics in Medicine*, 22(24):3755–3770.

Satten, G. A. (1999). Estimating the extent of tracking in interval-censored chain-of-events data. *Biometrics*, 55(4):1228–1231.

Satten, G. A. and Longini, Jr., I. M. (1996). Markov chains with measurement error: Estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45(3):275–295.

Sharples, L. D. (1993). Use of the gibbs sampler to estimate transition rates between grades of coronary disease following cardiac transplantation. *Statistics in Medicine*, 12(12):1155–1169.

Sharples, L. D., Jackson, C. H., Parameshwar, J., Wallwork, J., and Large, S. R. (2003). Diagnostic accuracy of coronary angiography and risk factors for post-heart-transplant cardiac allograft vasculopathy. *Transplantation*, 76(4):679–682.

Sharples, L. D., Taylor, G. J., and Faddy, M. (2001). A piecewise-homogeneous Markov chain process of lung transplantation. *Journal of Epidemiology and Biostatistics*, 6(4):349–355.

Spiegelman, D., Carroll, R. J., and Kipnis, V. (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statistics in Medicine*, 20(1):139–160.

Spiegelman, D., Rosner, B., and Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, 95(449):51–61.

Spreij, P. (2001). On the Markov property of a finite hidden Markov chain. *Statistics & Probability Letters*, 52(3):279–288.

Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: the measurement error Jackknife. *Journal of the American Statistical Association*, 90(432):1247–1256.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296.

Sutradhar, R. and Cook, R. J. (2008). Analysis of interval-censored data from clustered multistate processes: Application to joint damage in psoriatic arthritis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(5):553–566.

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):1161–1174.

Tfelt-Hansen, P. and Olesen, J. (1985). Methodological aspects of drug trials in migraine. *Neuroepidemiology*, 4(4):204–226.

Titman, A. C. (2007). *Model diagnostics in multi-state models of biological systems*. PhD thesis, University of Cambridge.

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

Tolusso, D. and Cook, R. J. (2009). Robust estimation of state occupancy probabilities for interval-censored multistate data: an application involving spondylitis in psoriatic arthritis. *Communications in Statistics - Theory and Methods*, 38(18):3307–3325.

Tom, B. D. M. and Farewell, V. T. (2011). Intermittent observation of time-dependent explanatory variables: a multistate modelling approach. *Statistics in Medicine*, 30(30):3520–3531.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22.

Tuma, N. B., Hannan, M. T., and Groeneveld, L. P. (1979). Dynamic analysis of event histories. *American Journal of Sociology*, 84(4):820–854.

van den Hout, A., Jagger, C., and Matthews, F. E. (2009). Estimating life expectancy in health and ill health by using a hidden Markov model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(4):449–465.

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

van den Hout, A. and Matthews, F. E. (2009). A piecewise-constant Markov model and the effects of study design on the estimation of life expectancies in health and ill health. *Statistical Methods in Medical Research*, 18(2):145–162.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, 92(1):1–28.

Varin, C., Høst, G., and Skare, Ø. (2005). Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics & Data Analysis*, 49(4):1173–1191.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528.

Varin, C. and Vidoni, P. (2008). Pairwise likelihood inference for general state space models. *Econometric Reviews*, 28(1-3):170–185.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312.

Wang, M. and Williamson, J. M. (2005). Generalization of the Mantel-Haenszel estimating function for sparse clustered binary data. *Biometrics*, 61(4):973–981.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

White, B. J. G. (2007). *Measurement error and misclassification in interval-censored life history data*. PhD thesis, University of Waterloo.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

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

Yan, Y. and Yi, G. Y. (2015). A class of functional methods for error-contaminated survival data under additive hazards models with replicate measurements. *Journal of the American Statistical Association*. To appear.

Yi, G. Y. (2008). A simulation-based marginal method for longitudinal data with drop-out and mismeasured covariates. *Biostatistics*, 9(3):501–512.

Yi, G. Y. (2009). Covariate measurement error in life histories. *International Journal of Statistical Sciences*, 9:177–197.

Yi, G. Y. and He, W. (2006). Methods for bivariate survival data with mismeasured covariates under an accelerated failure time model. *Communications in Statistics - Theory and Methods*, 35(8):1539–1554.

Yi, G. Y. and He, W. (2012). Bias analysis and the simulation-extrapolation method for survival data with covariate measurement error under parametric proportional odds models. *Biometrical Journal*, 54(3):343–360.

Yi, G. Y. and Lawless, J. F. (2007). A corrected likelihood method for the proportional hazards model with covariates subject to measurement error. *Journal of Statistical Planning and Inference*, 135(6):1816–1828.

Yi, G. Y. and Lawless, J. F. (2012). Likelihood-based and marginal inference methods for recurrent event data with covariate measurement error. *Canadian Journal of Statistics*, 40(3):530–549.

Yi, G. Y., Liu, W., and Wu, L. (2011a). Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing responses. *Biometrics*, 67(1):67–75.

Yi, G. Y., Ma, Y., and Carroll, R. J. (2012). A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*, 99(1):151–165.

Yi, G. Y., Ma, Y., Spiegelman, D., and Carroll, R. J. (2015). Functional and structural methods with mixed measurement error and mislassification in covariates. *Journal of the American Statistical Association*. To appear.

Yi, G. Y., Zeng, L., and Cook, R. J. (2011b). A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. *Canadian Journal of Statistics*, 39(1):34–51.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: an Introduction Using R*, volume 110 of *Monographs on Statistics and Applied Probability*. Boca Raton : CRC Press.