Mixture Models for Coarsened Multivariate Failure Time Data

by

Shu Jiang

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

Waterloo, Ontario, Canada, 2018

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Abdulkadir Hussein
Professor, Dept. of Math and Stats,
University of Windsor

Supervisor(s): Richard J. Cook
Professor, Dept. of Stats and ActSc,
University of Waterloo

Internal Member: Audrey Beliveau
Assistant Professor, Dept. of Stats and ActSc,
University of Waterloo
Leilei Zeng
Associate Professor, Dept. of Stats and ActSc,
University of Waterloo

Internal-External Member: Martin Cooke
Associate Professor, School of Public Health
University of Waterloo

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

The aim of this thesis is to develop statistical methodology for the analysis of life history data under incomplete observation schemes and with latent features which must be accommodated to ensure models provide a reasonable representation of the processes of interest and advance scientific understanding.

Life history data frequently arise in health studies of disease processes in which individuals pass through a series of stages of disease. Multistate models offer an appealing approach to modelling processes in settings where the stages can be meaningfully characterized into a finite number of disjoint stages and we adopt such models for much of the research in this thesis. In many instances, because processes are only observed intermittently, the precise number, types and times of transitions between assessments are not available. For failure time processes at most a single transition can occur between assessments and the resulting data are called interval-censored failure time data. For more general multistate processes it is more generally called a panel data observation scheme. We investigate problems related to interval-censored data throughout this thesis, and consider a more extreme form of incomplete data due to aggregation. The term coarsened data is used to unify these settings.

Despite careful attempts to collect and exploit available information to characterize the dynamic features of life history processes, substantial unexplained variability often exists between individuals or groups of individuals. Heterogeneity can be accommodated in various ways. Finite mixture models can be specified to accommodates distinct classes, or sub-populations, in which different disease processes govern progression in the different classes; latent class models are often used when class membership is fixed. When there are two classes and no disease progression occurs in one class, so-called cure rate models are often used. Classical mixture models with continuous random effect models are also often used to account for heterogeneity which can be characterized by a more finely distinguished nature of unexplained variation. This approach is often used in frailty models for survival data or more generally accommodating between cluster variation in clustered data.

In this thesis, the focus is on methods for statistical modeling and inference for multivariate failure time and multistate processes subject to intermittent observation; the resulting data are interval-censored multivariate failure time data and panel data respectively.

Finite mixture models offer a powerful approach for accommodating heterogeneity when there are distinct types of processes present in a population with latent sub-populations following one of such processes. Methods for fitting finite mixture models and conducting score tests for genetic markers are developed in Chapter 2 for a problem involving heterogeneous multistate processes under intermittent observation.

When there are multiple marginal processes of interest, the correlation between such processes must be taken into account. In Chapter 3 we develop multivariate models for the joint analysis of marginal processes. Copula models are popular for modeling the correlation between marginal failure time processes, while odds ratios are commonly used to capture the association between binary variables. Through the use of multivariate mixture models the dependence structure can be decomposed into one for susceptibility and one for the failure times given joint susceptibility.

Mixed multistate processes involving aggregate data are developed in Chapter 4 and 5. The computational challenges are addressed through the use of composite likelihood. We deal with between-cluster variation/within-cluster correlation in both chapters and propose two approaches to deal with such data. Specifically, we propose a marginal approach where we introduce dependence modeling via copulas, propose a composite likelihood and derive procedure for inference. A random effect model is also formulated in which a cluster-level latent variable accommodates heterogeneity between clusters. An optimal cost-effective design is also proposed which gives insights regarding the efficiency of studies involving aggregation and tracking. In Chapter 5, sample size criteria are developed to meet design objectives and cost-effective optimal allocations of clusters to the tracking and aggregate observation schemes are developed.

Acknowledgements

I would to take this opportunity to express my deep gratitude to my supervisor Dr. Richard J. Cook. I very much appreciate the support, guidance and mentorship he has shown me throughout the course of my Ph.D. He is very inspiring and is a role model that motivates me in many ways as a researcher. I would also like to thank my committee members Dr. Audrey Beliveau, Dr. Leilei Zeng, Dr. Abdulkadir Hussein and Dr. Martin Cooke for sharing valuable comments and feedback with regard to my thesis. I also want to thank Ker-Ai Lee for her advice on statistical computing. Last but not least I want to thank my friends and family that supported me throughout my Ph.D. journey.

Dedication

This is dedicated to my parents M. Huang and A. S. Jiang.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General Introduction to Research Topics

In studies of chronic disease processes, interest often lies in the time at which a certain event occurs. Usually events represent the occurrence of an undesirable change in a disease process, such as the development of a complication, the onset of organ damage, or death. The time from some origin until the event occurs is referred to as a failure time. When several facets of a disease process are under study, multiple failure times are often of interest. In such cases scientific interest may lie in modeling the relationship between the various events, but even when this is not the case it is usually important to deal with the fact that the failure times may be correlated within individuals. When the events represent the occurrence of events with a very different meaning and nature, the data are often referred to as multivariate failure time data (Hougaard, 2012). When the multiple times represent the occurrence of the same type of event, the resulting data are called recurrent event data (Cook and Lawless, 2007). In particular settings, multistate models can offer a structured and appealing way of modeling event occurrence and the relation between events of different types. The associated state space diagrams (Klein and Moeschberger, 2003; Cook and Lawless, 2018) can be useful for representing the possible sequences of events in a disease process.

Failure times are often not observed due to incomplete observation. Right-censored data arise when processes are not observed for a sufficiently long period of time to record the failure times (Kalbfleisch and Prentice, 2002; Lawless, 2003). In other settings failure times are not known exactly because they cannot be observed directly, but the status of an individual can be assessed by periodic intensive examination through imaging, blood tests or other means. In such cases event times are only known to have fallen in a particular interval, perhaps between the last negative assessment and the first positive assessment; such data are referred to as interval-censored data (Sun, 2006). When multistate processes are under intermittent observation the resulting data are referred to as panel data (Kalbfleisch and Lawless, 1985) and when such data are reported in aggregated form it may be more broadly characterized as coarsened data (Heitjan and Rubin, 1991). An additional theme involves dependence modeling and accounting for heterogeneity in life history processes under incomplete observation. The particular topics to be considered are described briefly in Section 1.3 but we first briefly discuss some motivating settings.

They are described in more detail prior to the development and application of the methodological advances.

1.2 Some Motivating Settings

1.2.1 University of Toronto Psoriatic Arthritis Cohort

The Centre for Prognosis Studies in the Rheumatic Diseases treats patients with various rheumatic diseases. A registry was created in 1976 called the University of Toronto Psoriatic Arthritis Cohort which recruits and follows patients with psoriatic arthritis (PsA), an immunological disease in which persons experience pain, inflammation and ultimately destruction of joints in the body (Gladman and Chandran, 2010). Upon entry to the clinic, patients undergo a detailed clinical and radiological examination and provide serum samples for genetic testing. They are then assessed annually or biannually to record joint damage scores (Rahman et al., 1998), and other factors such as biomarker levels.

Inflammation of the sacroiliac joints, the spine and neck, and reduced lateral range of motion of the back are all conditions that characterize spondylitis. Spondylitis is one of the musculoskeletal manifestations of psoriatic arthritis and it can have a severe detrimental effect on functional ability and quality of life of patients. Hanly et al. (1988) identified 52 of 220 (23.6%) patients recruited to a cohort of patients with psoriatic arthritis as having this disease. The degree of damage of the SI joints was scored according to the New York Criteria (Bennett and Wood, 1968) with the following categories: 0 for a normal joint, 1 if the presence of damage is equivocal, 2 if the joint is abnormal due to erosions of the bone surface or sclerosis, 3 if the joint is unequivocally abormal, and 4 if there is evidence of ankylosis (abnormal stiffening and immobility due to bone fusion) of the joint.

This motivates our work in multiple aspects. If we focus only on the intensities of the diseases progression, it can be explored considering a multistate process or through correlated marginal processes. However, the data suggest that a big fraction of the population do not develop damage on either the left or right side of the body. Hence a naive analysis ignoring the nonsusceptible fractions in the population may lead to poor inferences. In Chapters 2 and 3, the statistical methods that we consider will focus on the intensities of the disease process while accounting for the nonsusceptible fraction at the same time.

Data in the U of T PsA Cohort are recorded on a total of 64 joints in the body, 28 of which are in the two hands, 12 of which are in the two feet. The degree of joint damage in each of the 64 joints of each individual is recorded upon each visit (Rahman et al., 1998).

Note that joint damage is a strictly progressive process, meaning that once the joint is damaged it can no longer be repaired. Models used to represent progression in the severity of joint damage can be based on the state-space diagram in Figure 1.1. Sutradhar and Cook (2008) considered four states of increasing severity of damage based on the modified Steinbrocker score (Rahman et al., 1998). Each of the 28 hand joints were then classified into one of these states at each assessment time and the data were analyzed to model the development of joint damage. In Chapter 3 we consider three methods that deals with the co-occurrence of damage in the hands and feet which also accommodates the presence of the nonsusceptible population.

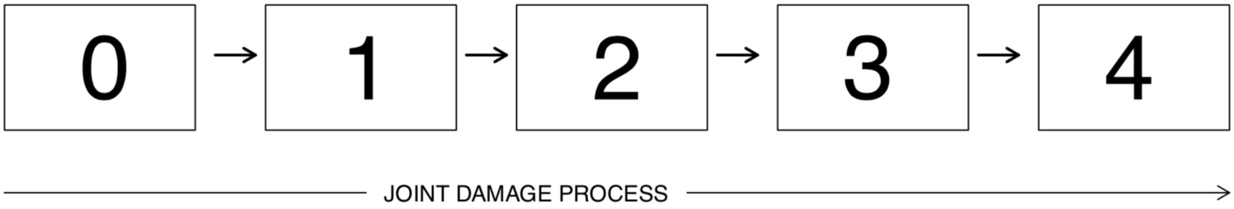


Figure 1.1: Grading joint damage based on radiologic assessments with state 0 being normal, state 1 being swelling of the soft tissues, erosion in stage 2, erosion along with joint space narrowing in state 3 and joint destruction in state 4 .

1.2.2 Growth and Development of *Lepidopsetta polyxystra*

Lepidopsetta polyxystra is also known as the Northern rock sole. The study took place in 2011 in the Hatfield Marine Science Center in Newport, Oregon and the larvae were collected in Chiniak Bay, Kodiak, Alaska (Laurel et al., 2014). The larvae developments are known to be sensitive to change in temperatures and the study was aiming to understand the effect of temperature on growth rate of the larvae. In order to study the temperature effect, larvae were maintained in 15 incubation containers under four different temperature set-ups. The larvae metamorphosis through different stages as illustrated in Figure 4.1.

Larval development was scored by the degree of observed tail flexion using the criteria established by Hawkyard et al. (2014). Stage 1 is characterized by a straight notochord (no flexion); stage 2 represents straight notochord with the appearance of caudal peduncle node near the tail where caudal peduncle is the narrow part of a fish's body to which the tail fin is attached; larvae in stage 3 have a bent notochord with caudal peduncle formation near the tail; stage 4 larvae have bent notochord and initial envelopment of the notochord by the caudal peduncle; and stage 5 is characterized by the full envelopment of the notochord by the caudal peduncle with only a remnant of bent notochord still visible (Laurel et al., 2014). Two observation schedules were employed: for tanks 1-7 the developmental stages were to be assessed on days 24, 28, 46, 66 and 90 days, while for tanks 8-15 the classifications were to occur on days 10, 18, 30 and 47. There were incomplete data in many tanks due to missing assessment on the scheduled days. Larvae are mobile small animals that are difficult to identify, The process of identification of individual larvae can take time, effort and cost and may still be subject to error. Thus, it has motivated us to develop methods that do not require tracking of individual organisms but rather relies on the frequency counts for the different stages at each assessment time in Chapter 4. Moreover, this identifiability issue of small organisms has also motivated us to develop optimal cost-effective design for prospective studies in Chapter 5. Note that in the original dataset sampled fish randomly from the 15 tanks due to identifiability problems. For the purpose of demonstrating our method, we treat them as the same individuals.

1.3 Outline of Thesis

1.3.1 A Finite Mixture Model for Multistate Panel Data

Transition times reflecting disease progression are often interval-censored when disease status is only known at a series of assessment times. When the precise state of a multistate process is only available at period assessment times, the resulting data are often referred to as panel data (Kalbfleisch and Lawless, 1985). In many settings there is considerable variability in the nature and rate of disease progression and this can be more than expected based on a simple model. For example, some individuals may tend to progress quickly from one state to another, while others may not experience certain types of disease complications; this motivates us to formulate a model which includes both aspects of the disease, i.e., the progression of the disease and the damage status. In Chapter 2, a finite mixture model is described which accommodates different Markov processes followed in different latent classes as well as a nonsusceptible sub-populations under intermittent observations. Under this framework, a score test is developed which enables one to identify covariates of interest for further investigation. Simulation studies examining the performances of the proposed model and the type I error of the score tests show good attainment. An application involving progression in joint damage in psoriatic arthritis provides illustration.

1.3.2 Mixture Models for Multivariate Interval-censored Data

The finite mixture model that we have considered in Section 1.3.1 accounts for heterogeneity among individuals by accommodating underlying classes of individuals defined by latent variables. Here we consider multiple processes within an individual, each with a latent binary susceptibility variable. By adopting the marginal specifications we can directly intepret the covariate effects at the population level. In Chapter 3, we consider a bivariate cure rate model for interval-censored failure time. We introduce two types of dependence structures between the two processes within an individual. Specifically, we introduce one association model capturing the dependence between susceptibility to the disease in each disease process and the other characterizing the association between failure times of the two processes given joint susceptibility. We introduce three approaches to estimation: maximum likelihood, two-stage pseudo-likelihood and weighted estimating equations. Simulations show good performances of the three proposed methods. An application involving the onset of damage in the hands and feet joints in psoriatic arthritis provides further illustration.

1.3.3 Analysis of Aggregate Multistate Data

Markov processes offer a useful basis for modeling the passage of organisms through developmental stages. When organisms are under intermittent observation, likelihoods based on panel data can naturally be constructured using the transition probability functions. In Chapter 4 we consider the problem in which organisms are not tracked individually due to the difficulty of identifying them, but rather aggregate counts of the number of organisms in different stages of development are recorded at successive time points in each of a number of tanks. Methods are developed to accommodate clustering of transition rates within tanks through use of marginal models with robust variance estimates, and using random effect models. Composite likelihood is used as a basis of inference. The methods are shown to perform well in empirical studies and are applied to a dataset on the developmental stages of the Northern rock sole.

1.3.4 Cost-effective Design with Aggregation and Tracking

Studies of the development and growth of organisms are often conducted in laboratories where organisms maintained in tanks are examined repeatedly over time. Collection and recording of cross-sectional aggregate count data on stage occupancy is both less expensive and administratively more convenient than tracking the stages of each organism over time. In such settings tank-to-tank variation must also be taken into account as growth rates may be more similar among organisms within the same tank than for those in different tanks. In Chapter 5 we consider the cost-effect design of a prospective developmental study of organisms based on a marginal Markov model which deals with between tank variation and within tank dependence. We develop a flexible design in which some tanks provide repeated cross-sectional aggregate data, and other tanks provide serial responses through tracking individuals. We assess the relative efficiency of aggregate and individual-level longitudinal data. The optimal cost-effective design is shown to depend on whether primary interest lies in transition intensities or associated cluster-level covariate effects. We also give an illustrative example on the growth and development of the Northern rock sole.

Chapter 2

A Finite Mixture Model for Multistate Panel Data

2.1 Introduction

2.1.1 Literature Review

It is often of interest to model the rate at which chronic diseases progress for scientific understanding, making prognoses, and health policy decision making. Multistate models offer an appealing and powerful framework for modeling disease processes in settings where the degree of damage can be meaningfully characterized into a finite number of disjoint states. Among individuals with hepatitis C infection for example, the extent of liver damage is quantified using a five point scale with state 1 representing no fibrosis, states 2 to 4 representing increasing degrees of fibrosis and state 5 representing cirrhosis (Sweeting et al., 2006). In diabetic retinopathy the extent of damage is measured on an eleven point scale with state 1 representing no damage and state 11 severe damage (The Diabetes Control and Complications Trial Research Group, 1993). Multistate models have also proven useful in characterizing decline in cognitive function in dementia (Tyas et al., 2007), loss of functional ability in arthritic conditions (Husted et al., 2007), and progression of immunological disease (Gentleman et al., 1994), and the development of asymptotic vertebral fractures in patients with osteoporosis (Riggs et al., 1981).

Despite careful attempts to use available information to characterize such processes, substantial unexplained variability in disease processes between individuals is often evident. While Markov models often provide a natural and convenient starting point for modeling such processes, generalizations are warranted in such settings. Satten (1999) considered a conditionally Markov model for a progressive multistate process where a single non-negative random effect was specified to act multiplicatively on each transition intensity to account for between-subject heterogeneity. Cook et al. (2002) described a conditional Markov model for generalized mover-stayer model for panel data. Cook et al. (2004) and Sutradhar and Cook (2008) developed an extension for clustered progressive processes with correlated random effects which were unique to each possible transition.

Discrete random effect models are also useful, with the most popular being the so-called mover-stayer model in which some fraction of the population of interest may not be at risk for disease progression; in this case individuals are considered "stayers" if they are not at risk, while those who are at risk of progress are thought of as "movers". Frydman (1984) developed maximum likelihood methods for this setting and Fuchs and Greenhouse (1988) outlined an expectation-maximization (EM) algorithm (Dempster et al., 1977) which accommodates censoring. O'Keeffe et al. (2013) consider random effects which accommodate a point mass at zero and a continuous random effect for susceptible individuals; specifically in their spatial analyses on the location of joint damage in psoriatic arthritis they explore random effect models with a mover-stayer inverse Gaussian and a compound Poisson distribution. Finite mixture models offer significant generalization of mover-stayer models and far less has been done for this setting. Here the target population is envisioned as being comprised of several distinct sub-populations and the disease processes are allowed to differ in some ways between these sub-populations. In general it will not be known which sub-population an individual is in so the membership is considered a latent variable; in this case the mixing distribution and the parameters governing the process dynamics in each sub-population are estimated. The EM algorithm can again be useful in this setting (Dempster et al., 1977).

In many instances it is not apparent when a disease process has progressed and so the precise times of transitions between states are not available. This will be the case in most of the examples given in the opening paragraph. When the precise state of a multistate process is only available at periodic assessment times the resulting data are often referred to as panel data. In recent years, much statistical research has taken place on the analysis of such data which are referred to as panel data, or alternatively multistate data with interval-censored transition times. Kalbfleisch and Lawless (1985) developed an efficient algorithm for maximum likelihood estimation under a Markov assumption which is implemented in the *msm* package by Jackson (2011). Grüger et al. (1991) described the conditions that need to be satisfied for the observation process to be ignorable and such analyses valid, which are in effect the sequentially missing at random assumption given by Hogan et al. (2004).

The remainder of this contribution is organized as follows. In the next subsection, we introduce the University of Toronto Psoriatic Arthritis Registry and describe the data that motivates this work. In Section 2.2, we define notation and describe a model for a finite mixture of Markov processes. Specifically, we construct the likelihood for the setting where individuals are under intermittent observation and describe how to estimate the asymptotic covariance matrix for the estimates. Score tests are developed in Section 2.3 where their finite sample properties are also studied by simulation. An application involving joint damage in patients with psoriatic arthritis is given in Section 2.4 and concluding remarks and topics for further research are given in Section 2.5.

2.1.2 Sacroiliac Joint Involvement in Psoriatic Arthritis

The University of Toronto Psoriatic Arthritis Clinic is a tertiary referral center for individuals with psoriatic arthritis (PsA), an immunological condition which features both skin and joint involvement (Gladman and Chandran, 2010). A registry of patients was created in 1976, which has been recruiting and following patients continuously since its inception. Patients undergo a detailed clinical and radiological examination upon entry to the clinic, and provide serum samples for genetic testing. Follow-up clinical and radiological assessments (Rahman et al., 1998) are scheduled annually and every two years respectively in order to track changes in joint damage. Spondylitis is one of the musculoskeletal manifestations of psoriatic arthritis and is characterized by inflammation of the sacroiliac joints, the spine and neck, and reduced lateral range of motion of the back. Hanly et al. (1988) identified 52 of 220 (23.6%) patients recruited to a cohort of patients with psoriatic arthritis as having this disease. There is particular interest in involvement of the sacroiliac joints since when these become damaged it can have a severe detrimental effect on functional ability and quality of life of patients.

A recent study by Harron et al. (2016) investigated the association between human leukocyte antigen (HLA) B and C loci and sacroiliac joint involvement in a cohort of patients with psoriatic arthritis. These authors used radiographic evidence of sacroilitis (SI) to define axial disease. The criterion for identifying SI presence was of at least grade 2 radiographic damage (unilateral or bilateral) on a five point grading scheme. This analysis was cross-sectional, however, and did not fully account for the variable times individuals have been at risk for developing damage in the sacroiliac joints. The proposed analysis is based on a multistate model in which individuals make transitions between disease states as damage occurs.

The particular formulation of our model is motivated by the possible sub-types of patients with sacroiliac involvement and the relation between so-called psoriatic spondylitis, a condition in which psoriatic arthritis patients develop back involvement, and ankylosing spondylitis, an arthritic condition primarily affecting the spine. The former is often, but not always unilateral, while the latter is more commonly bilateral. We therefore formulate a model which accommodates a different course (unilateral or bilateral) of the disease in psoriatic arthritis and aim to detect HLA alleles associated with these courses. We give the details of this in the next section.

Figure 2.1 shows the time course of damage for a sample of six individuals. For each individual, the duration of follow-up since time of disease onset is represented by the length of the horizontal line; vertical hatch marks reflect the times that clinic visits are made and joints are assessed. Four different types of line segments are used to convey the damage status of each individual at a given time with a solid line representing no sacroiliac joint involvement, a dashed line representing left side involvement, a dashed-dotted line representing right side involvement, and a dotted line representing bilateral involvement. The periods of time when no line segment is drawn are intervals in which the status is unknown because there was a different damage status for the visit at the left endpoint than at the right endpoint. Since damage is assessed radiologically, the exact times at which damage occurs is unknown so the times are interval-censored. We note from Figure 2.1 that some individuals develop the damage shortly after diagnosis (e.g. individual 1) and some were not observed to develop damage despite long follow-up (e.g. individual 3). Moreover, it is apparent that some individuals progress quickly from unilateral to bilateral damage once they enter the unilateral stage, as is the case of individuals 2 and 5, while some continue to have only unilateral involvement until the end of follow-up. This motivates the formulation of a model which accommodates a mover-stayer component along with a component which allows for changes in damage status.

We let state 0 = no SI damage, state 1 = unilateral (left), state 2 = unilateral (right), and state 3 = bilateral. Patients are classified into four possible classes as in Figure 2.2. We denote $\lambda_{k\ell}$ as the intensity of transitioning from state k to state ℓ. We constrain λ_{01} and λ_{02} to be the same for the two unilateral classes as they are for the bilateral class.

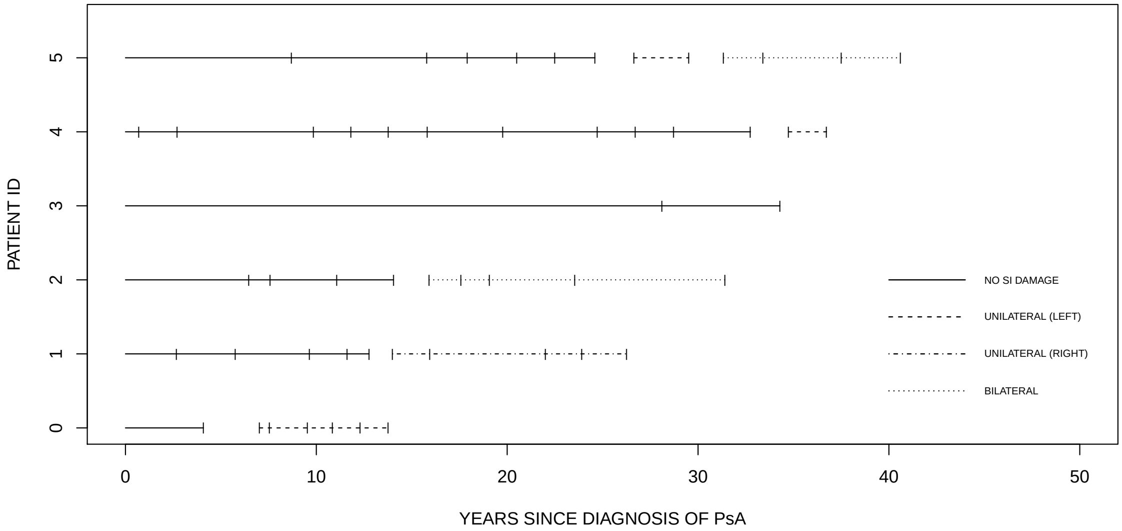


Figure 2.1: Plot of assessment times (hatch marks) and the type of joints damage (four types of line segments) between assessments from onset of PsA for a selected sample of patients from the University of Toronto Psoriatic Arthritis Clinic.

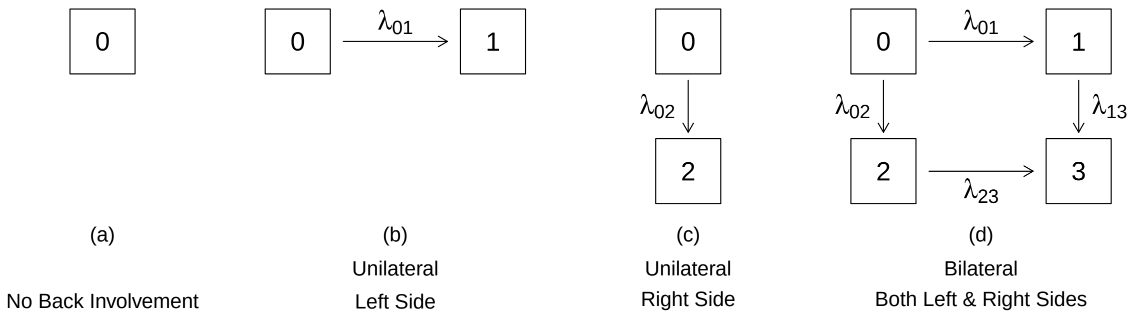


Figure 2.2: Multistate diagram for the processes of the four classes of individuals in the finite mixture model.

2.2 Model Formulation

2.2.1 Notation

We restrict our attention to progressive processes in which the states represent, for example, the severity of irreparable damages. Assume there are K states labelled $k = 1, ..., K$. Let $Z(t)$ represent the state occupied by the disease progress at time t since disease onset and $\{Z(s), 0 < s\}$ denote the associated stochastic process. Let X denote a $p \times 1$ covariate vector. When there exist heterogeneity across subjects, a traditional Markov model may be insufficient. We therefore consider the setting where the population arises from distinct sub-populations via a discrete mixture of different processes.

In finite mixture models with discrete mixtures, individuals in the same class are governed by a common process. Let C be a latent random variable indicating the class to which a particular individual belongs. We let $P(C = j|X;\beta) = \pi_j(X;\beta)$ where $\sum_{j=0}^{J} \pi_j(X;\beta) = 1$ denotes the probability of belonging to class j given a set of fixed covariates X. The term $P(C = j|X;\beta)$ is typically modeled by a multinomial logistic regression (McCullagh and Nelder, 1989). Using the first class as the reference category, we let

$$P(C = j|X;\beta) = \frac{\exp(X'\beta_j)}{1 + \sum_{j=0}^{J}\exp(X'\beta_j)}, \tag{2.1}$$

where $X = (1, X_1, ..., X_p)'$.

We let $\mathscr{H}(t) = \{Z(s), 0 < s < t; X\}$ denote the history at time t. For class j ($j = 0, 1, ..., J$), we denote the transition intensities as

$$\lim_{\triangle t \downarrow 0} \frac{P(Z(t + \triangle t^-) = l|Z(t^-) = k, \mathscr{H}(t), C = j)}{\triangle t} = \lambda_{kl}(t|\mathscr{H}(t), C = j) = \lambda_{jkl}(t|\mathscr{H}(t))$$

for $k, l = 1, ..., K$ and $j = 0, ..., J$. We further restrict attention to Markov processes for $\lambda_{jkl}(t|\mathscr{H}(t)) = \lambda_{jkl}(t)$. Notice that we do not consider the effects of covariate on the transition intensities. Doing so can raise estimability challenges and the primary goal is to examine covariate effects on class membership. We also let $\lambda_j = \lambda_{jkl}$ for $\forall k \neq l$ and $\theta = (\alpha', \beta')'$ where $\alpha = (\alpha'_0, ..., \alpha'_J)'$ and $\alpha_j = \log \lambda_j$.

We now consider a sample of size m comprised of independent individuals labelled $i = 1, ..., m$. The inspection times for individual i are denoted by a_{ir}, $r = 0, ..., R_i$ and if X_i is a fixed covariate, the resulting data for individual i are denoted by $\mathscr{D}_i = \{(Z_i(a_{ir}), a_{ir}), r = 0, 1, ..., R_i; X_i\}$. We define the likelihood contribution for a particular individual i on a finite mixture model of Markov processes as

$$L_i(\theta) = \sum_{j=0}^{J}\left\{\prod_{r=1}^{R_i} P(Z_i(a_r)|Z_i(a_{r-1}), C_i = j, X_i; \alpha)\right\} P(C_i = j|X_i; \beta). \tag{2.2}$$

To simplify the notation, we let

$$L_{ij}(\alpha) = \prod_{r=1}^{R_i} P(Z_i(a_r)|Z(a_{r-1}), C_i = j, X_i; \alpha).$$

The observed likelihood for individual i is then

$$L_i(\theta) = \sum_{j=0}^{J} L_{ij}(\alpha)\pi_j(X_i; \beta). \tag{2.3}$$

The model defined in (2.2) is based on the assumption that the latent classes are mutual exclusive and exhaustive; that is, each individual is a member of one and only one of the latent classes.

We can obtain the estimate of θ by maximizing $L(\theta) = \prod_{i=1}^{m} L_i(\theta)$, or equivalently solving the corresponding observed data score equation $U(\theta) = 0$ where $U(\theta) = (U_1'(\theta), U_2'(\theta))$ with $U_1(\theta) = \sum_{i=1}^{m} U_{i1}(\theta)$, $U_2(\theta) = \sum_{i=1}^{m} U_{i2}(\theta)$ where $U_{i1}(\theta) = \partial \log L_{ij}(\alpha)/\partial\alpha$, $U_{i2}(\theta) = \partial \log P(C_i|X_i; \beta)/\partial\beta$.

2.2.2 Estimation via the EM algorithm and Louis' Observed Information

Suppressing the subscript i for individuals and considering the contribution from a generic individual, the complete data likelihood is

$$\mathscr{L}(\theta) \propto \prod_{j=0}^{J} \{L_j(\alpha)\pi_j(X; \beta)\}^{I(C=j)}. \tag{2.4}$$

If we let θ^r denote the estimate of θ at the rth iteration and $\mathscr{D}$ be the observed data, then

$$w_j^r = P(C = j|\mathscr{D}; \theta^r) = \frac{L_j(\alpha^r)\pi_j(X; \beta^r)}{\sum_{j=0}^{J} L_j(\alpha^r)\pi_j(X; \beta^r)}.$$

At the rth iteration of the EM algorithm, the E-step takes the conditional expectation $Q(\theta;\theta^r) = E\{\log\ \mathscr{L}(\theta) \mid \mathscr{D};\theta^r\}$ where $Q(\theta;\theta^r) = Q_1(\alpha;\theta^r) + Q_2(\beta;\theta^r)$ and

$$\begin{aligned}
Q_1(\alpha;\theta^r) &= \sum_{j=0}^{J} w_j^r \log L_j(\alpha),\\
Q_2(\beta;\theta^r) &= \sum_{j=0}^{J} w_j^r \log \pi_g(X;\beta).
\end{aligned}$$

The M-step involves maximizing $Q(\theta;\theta^r)$ with respect to θ and obtaining the updated estimate $\theta^{(r+1)}$. Note that we can maximize Q_1 using a Fisher-scoring algorithm (Kalbfleisch and Lawless, 1985) class by class if α_j $(j = 0, ..., J)$ are distinct. The function $Q_2(\beta;\theta^r)$ can be maximized using functions for estimation based on quasi-likelihood. We then iterate between the E-step and M-step until the convergence criterion $\max|\theta^r - \theta^{r-1}| < \epsilon$ is achieved, where ϵ is the user-specified tolerance.

Note that $U(\theta) = E\{S(\theta)|\mathscr{D}\}$ where $S(\theta) = \partial\mathscr{L}(\theta)/\partial\theta$, so the EM algorithm is simply one approach to maximize the observed data log-likelihood function. To avoid computation of the second derivative matrix of the observed log-likelihood, we compute the observed information matrix $I(\theta) = -\partial U(\theta)/\partial\theta'$ based on the approach of Louis (Louis, 1982) who showed that

$$I(\theta) = E\{\mathscr{J}(\theta)|\mathscr{D}\} - E\{S(\theta)S'(\theta)|\mathscr{D}\} + U(\theta)U'(\theta) \tag{2.5}$$

where

$$\mathscr{J}(\theta) = \begin{pmatrix} \partial^2 \log\mathscr{L}(\theta)/\partial\alpha'\partial\alpha & 0 \\ 0 & \partial^2 \log\mathscr{L}(\theta)/\partial\beta'\partial\beta \end{pmatrix} = \begin{pmatrix} \partial S_1(\alpha)/\partial\alpha' & 0 \\ 0 & \partial S_2(\beta)/\partial\beta' \end{pmatrix}$$

is the complete data information matrix. To compute $E\{S(\theta)S'(\theta)|\mathscr{D}\}$, we note it is equal to

$$E\{S(\theta)S'(\theta)|\mathscr{D}\} = \mathrm{var}(S(\theta)|\mathscr{D}) + E(S(\theta)|\mathscr{D})E(S'(\theta)|\mathscr{D}) = \mathrm{var}(S(\theta)|\mathscr{D}) + U(\theta)U'(\theta),$$

since $U(\theta) = E\{S(\theta)|\mathscr{D}\}$. Note that if we find $\mathrm{var}(S(\theta)|\mathscr{D})$, the third term of the Louis' formula (Louis, 1982) cancels, and we no longer require $U(\theta)$. We let $S(\theta) = (S(\alpha)', S(\beta)')'$. In matrix notation,

$$S(\theta) = (A'_\alpha Y, A'_\beta Y),$$

where

$$A'_\alpha = \begin{pmatrix} \partial \log L_0(\alpha)/\partial\alpha_0 & \partial \log L_1(\alpha)/\partial\alpha_0 & ... & \partial \log L_J(\alpha)/\partial\alpha_0 \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ \partial \log L_0(\alpha)/\partial\alpha_J & \partial \log L_1(\alpha)/\partial\alpha_J & ... & \partial \log L_J(\alpha)/\partial\alpha_J \end{pmatrix},$$

$$A'_\beta = \begin{pmatrix} \partial \log \pi_0(\beta)/\partial\beta_1 & \partial \log \pi_1(\beta)/\partial\beta_1 & ... & \partial \log \pi_G(\beta)/\partial\beta_1 \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ \partial \log \pi_0(\beta)/\partial\beta_p & \partial \log \pi_1(\beta)/\partial\beta_p & ... & \partial \log \pi_G(\beta)/\partial\beta_p \end{pmatrix}$$

and $Y = (Y_0, Y_1, ..., Y_J)'$, where $Y_j = I(C = j)$, $j = 0, 1, ..., J$. Hence,

$$\mathrm{var}\{S(\theta)\} = \begin{pmatrix} A'_\alpha \mathrm{cov}(Y|X)A_\alpha & A'_\alpha \mathrm{cov}(Y|X)A_\beta \\ A'_\beta \mathrm{cov}(Y|X)A_\beta & A'_\alpha \mathrm{cov}(Y|X)A_\beta \end{pmatrix}$$

where

$$\mathrm{cov}(Y|X) = \begin{pmatrix} \mathrm{var}(Y_0|X) & \mathrm{cov}(Y_0, Y_1|X) & ... & \mathrm{cov}(Y_0, Y_J|X) \\ ... & \mathrm{var}(Y_1|X) & ... & ... \\ ... & ... & ... & ... \\ \mathrm{cov}(Y_J, Y_0|X) & \mathrm{cov}(Y_J, Y_1|X) & ... & \mathrm{var}(Y_J|X) \end{pmatrix}.$$

Since C represents the class membership, we note from (2.1) that $\mathrm{var}(Y_j|X) = \pi_j(X;\beta)(1-\pi_j(X;\beta))$ for $j = 0, ..., J$ and $\mathrm{cov}(Y_{j_1}, Y_{j_2}) = -\pi_{j_1}(X;\beta)\pi_{j_2}(X;\beta)$ for $j_1 \neq j_2$, $j_1, j_2 = 0, ..., J$. Then we can obtain the variance estimate by summing over all individuals i and calculate the inverse of the observed information matrix.

2.3 Score Tests for Genetic Effects

2.3.1 Construction of the Test Statistic

Since we are interested in the covariates on genetic effects, we let Z denote the covariates which include both the nuisance covariates X and the genetic covariates G. We then have $Z = (X', G')$. We further partition $\eta_j = (\beta_{j0}, \beta_{j1}, ..., \beta_{j(p-q)}, \gamma_{j1}, ..., \gamma_{jq})$ into nuisance parameters, i.e., β_j where $\beta_j = (\beta_{j0}, ..., \beta_{j(p-q)})$ and parameters of interest, i.e., γ_j where $\gamma_j = (\gamma_{j1}, .., \gamma_{jq})$ for each of the latent classes j. The resulting multinomial logistic regression (McCullagh and Nelder, 1989) is then

$$P(C = j|Z;\eta) = \frac{\exp(Z'\eta_j)}{1 + \sum_{j=0}^{J} \exp(Z'\eta_j)}, \tag{2.6}$$

where $Z'\eta = X'\beta + G'\gamma$.

Due to the time consuming and computationally demanding problem of simultaneously estimating all p parameters, we consider a score test of the null hypothesis of no genetic effect. The null hypothesis is

$$H_0 : \gamma = \gamma_0,$$

which specifies that $\gamma' = (\gamma_1, .., \gamma_j, ..., \gamma_C)$ are simultaneously equal to γ_0. The null hypothesis is tested against the alternative hypothesis:

$$H_1 : \gamma \neq \gamma_0.$$

Following Boos (1992), the score test statistic is

$$T = [U(\gamma_0, \hat{\psi}(\gamma_0))]' I^{\gamma\gamma}(\gamma_0, \hat{\psi}(\gamma_0)) U(\gamma_0, \hat{\psi}(\gamma_0)), \tag{2.7}$$

where $\psi = (\theta', \gamma')'$ and $U(\cdot)$ is a $q \times 1$ score function for γ. $\hat{\psi}(\gamma_0)$ is the maximum likelihood estimates of ψ under the constrained null model, and $I^{\gamma\gamma}(\cdot)$ is the $q \times q$ covariance matrix of γ. Asymptotically the score test statistic satisfies $T \sim \chi^2_q$ under the null hypothesis.

2.3.2 Simulation Studies

The purpose of the simulation studies are to (1) demonstrate the performance of a proposed finite mixture model and (2) investigate the rejection percentage of the score test. We consider 4 classes ($j = 0, 1, 2, 3$) and constrain α as in Figure 2.2. To model class membership for individual i we generate a Bernoulli covariate X_{i1} with $P(X_{i1} = 1) = 0.5$. Let $X_i = (1, X_{i1})'$, generate G_i as Bernoulli with probability of success 0.05 or 0.20, and let $Z_i = (X_i', G_i)'$, $i = 1, \ldots, m$. We set the coefficients for X_i to be $\beta_{11} = \beta_{21} = \log 1.1$ and $\beta_{31} = \log 1.2$; the coefficients for the genetic variable in the multinomial regression model is set such that $\gamma_1 = \gamma_2 = 0$ and $\gamma_3 = 0$. We determine the intercepts β_{10}, β_{20} and β_{30} so that $P(C = 0) = 0.30$, $P(C = 1) = 0.25$, $P(C = 2) = 0.25$, and $P(C = 3) = 0.20$.

The transition intensities in the multistate framework are set so that $P(Z(E) = 1|C =$

$1) = 0.80$, $P(Z(E) = 2|C = 2) = 0.80$, and $P(Z(E) = 3|C = 3) = 0.60$. The number of inspection times R_i for individual i is generated by a time homogeneous Poisson process with rate ρ giving $E\{N_i(E)\} = \mu$, where $\mu = 15$ and 30, $E = 1$ without loss of generality. To assess the performance of estimators we fit the correct model under the constraint $\gamma = 0$ and examine the empirical performance of the other parameter estimates. We display these in a table reporting the empirical bias, the empirical standard error, and the empirical coverage probability where the sample standard deviation is computed based on Louis' method (Louis, 1982). We also use these estimates as a basis for the score test of the null hypothesis of $\gamma = 0$. For each setting we consider the empirical type I error rate of $\gamma = 0$. We also consider the setting when $\gamma \neq 0$ and for each case we examine the empirical power of the score test. For this we consider several different values of the components of $\gamma = (\gamma_1, \gamma_2)'$ where $\gamma = (\log 1.25, \log 1.5)$ and $\gamma = (\log 1.5, \log 2)$; for each combination the values of β_{jc}, $c = 1, 2, 3$ to give the same marginal probabilities of class membership.

Table 2.1 contains simulation results with 2000 individuals per simulation under 500 simulations. From the simulation results in Table 2.1 we see good performances under the correct model and the type I error of the score test under H_0 is well within the nominal level. Figure 2.3 gives a 2×2 layout of the Q-Q plots for the 2 d.f. test statistic under the null hypothesis (represented in the third to last row of the Table 2.1). The top row displays the plots when the expected number of assessments is 15 and the bottom row when it is 30. The plots suggest good agreement between the empirical and asymptotic distributions for the case of 30 assessments on average; the slight increase in the empirical type I error rate when the expected number of assessments is low as 15 is explained by the fatter right tail of the empirical distribution of the test statistic than would be expected under the 2 d.f. chi-squared statistic. The increase in the empirical power with increasing effect size under H_A is also apparent from the bottom of Table 2.1 suggesting this test statistic can be useful in detecting genetic effects.

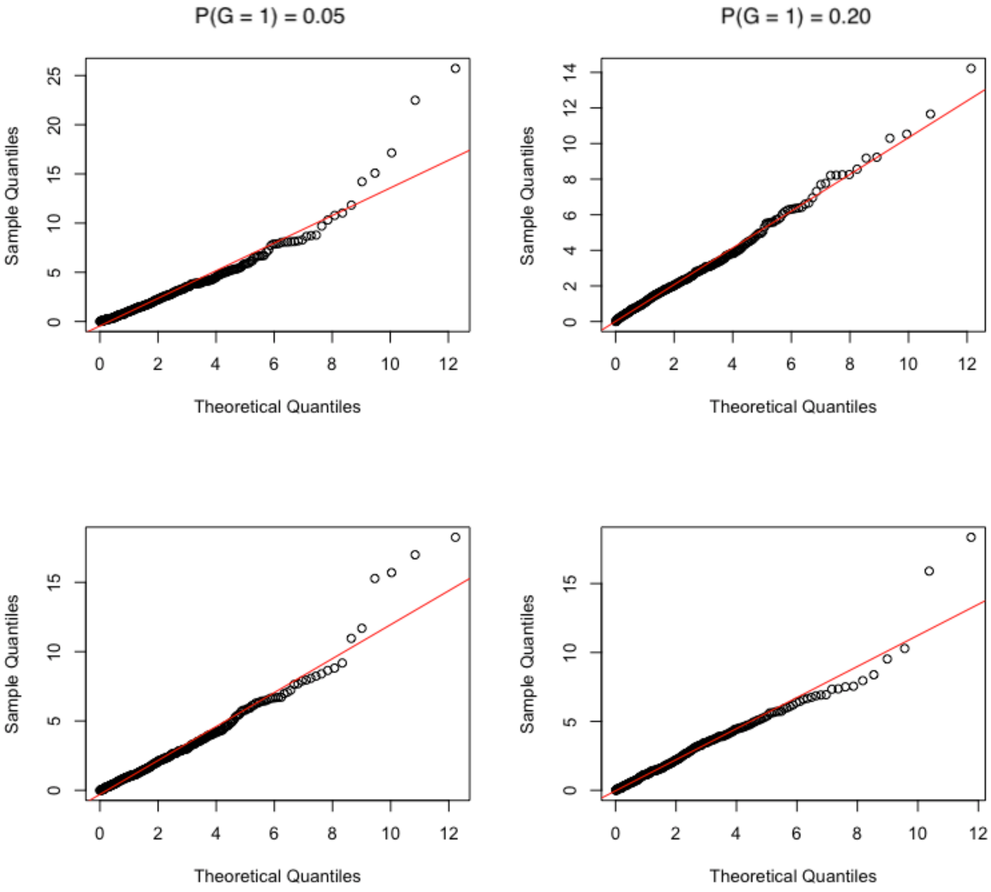


Figure 2.3: Q-Q plots of χ^2_2 statistics with the top row displaying results when the expected number of assessments is 15 and bottom row when is 30 .

2.4 Sacroiliac Damage in Psoriatic Arthritis

The methods developed in the previous sections were applied to data on joint damage in patients with PsA from the University of Toronto Psoriatic Arthritis Clinic. Specific interest lies in examining the effects of human leukocyte antigen (HLA) markers on the types of back involvement. We examine the effects of HLA markers on the SI joint damage while controlling for gender and patient age. We let β_{j0} = Intercept, β_{j1} = Gender (baseline = male), β_{j2} = Age (centered) for class j. We formulate our null hypothesis of $\gamma = 0$ where $\gamma = (\gamma_1, \gamma_3)$ with subscript 1 indicating patients who belong to the unilateral classes and 3 indicating the bilateral class.

Table 2.2 includes estimates, standard errors and 95% confidence intervals for all parameters under the null model. From this fitted model the odds of males experiencing bilateral disease (compared to no SI joint involvement) is lower than it is for females. The results of applying the score tests of Section 2.3.1 are given in Table 2.3. For the HLA analysis we excluded the HLA markers with a frequency of less than 1% due to sparsity. Among the HLA markers, HLA-A3, HLA-A29, HLA-B27, HLA-B35, HLA-C2, HLA-C4 and HLA-C12 have significant association with the types of back involvement.

We explore the effects of the identified HLA markers further by fitting a model including each of the HLA markers found to be significant at the 5% level. From Table 2.4, we see that HLA-A3 and HLA-C2 are risk factors, and HLA-B35 and HLA-C4 are protective for unilateral damage compared to no damage. HLA-B27 is a risk factor and HLA-A29 and HLA-B35 are protective for bilateral damage compared to no SI joint damage. A generalization of interest would be to assess whether the effects are different from unilateral and bilateral damage. We have reparametrized the model to assess the effects between unilateral and bilateral damage and the results are presented in Table 2.4. Among patients with SI damage, HLA-B35 and HLA-C4 are risk factors for, and HLA-A3 is protective for, bilateral SI involvement.

2.5 Discussion

We have formulated the finite mixture model of Markov processes under intermittent observations to accommodate the different transition intensities and regression coefficients between latent classes. Under this framework, the score test can be adopted to assess the effect of the markers. This approach is especially convenient when markers are large in number – only one model needs to be fitted instead of many. We study the empirical performances of the proposed model and show that the coverage probabilities are all compatible with the nominal 95% level. Then we link the score test to study its type I error under the null hypothesis of no markers effect, and show that the type I error is within the nominal 5% level. In line with the previous study (Harron et al., 2016), our method accommodated potential nonsusceptibles while considering the longitudinal responses of individuals instead of a single point in time. In addition, more significant markers showed up in our method compared to Harron et al. (Harron et al., 2016).

We consider the setting in which the transition intensities for the onset of damage in the left SI joint in class 1 is the same as the onset time of damage in the left side in class 3, where in the latter both joints are at risk of damage. This constraint can be relaxed in practice to allow for more general, functionally independent models in the different classes, but more data is required to enable estimation. It is also worth noting that the intensity in class 1 is for the only joint damage that can occur in this class, whereas in class 3 the λ_{01} term is the intensity for the first joint damage being in the left side in the competing risk setting where the first damage can be in the left or right sides. While justified on the basis of the assumption of robustness of intensities to removal of competing causes of failure, larger data sets involving more frequent inspection for joint damage would be of interest to relax this assumption. Alternative approaches to dependence modeling in susceptibility and in failure time given susceptibility which don't feature this complication are discussed in the next chapter.

		$E\{N_i(E)\} = 15$				$E\{N_i(E)\} = 30$			
	Value	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
Class model									
β_{10}	-0.182	-0.005	0.102	0.101	94.3	-0.014	0.098	0.097	95.3
β_{11}	$\log 1.1$	-0.012	0.100	0.102	95.4	-0.016	0.094	0.096	95.4
β_{20}	-0.182	-0.027	0.274	0.284	94.1	-0.033	0.243	0.248	94.1
β_{30}	-0.405	-0.009	0.283	0.283	93.7	-0.036	0.258	0.259	94.7
β_{31}	$\log 1.2$	-0.007	0.171	0.173	96.3	-0.008	0.157	0.161	97.1
Multistate Model									
α_{01}	0.476	-0.005	0.166	0.172	97.0	-0.004	0.159	0.162	95.8
α_{02}	0.476	0.032	0.213	0.221	95.6	0.036	0.205	0.205	96.0
α_{13}	0.354	-0.013	0.154	0.162	95.6	-0.001	0.148	0.158	97.5
α_{23}	0.354	-0.001	0.163	0.164	93.0	0.004	0.163	0.162	92.8
% Rejection									
		$P(G=1) = 0.05$		$P(G=1) = 0.20$		$P(G=1) = 0.05$		$P(G=1) = 0.20$	
Under H_0:		7.10		5.80		6.41		5.60	
Under ${H_A}^1$:		19.6		42.0		19.1		42.3	
Under ${H_A}^2$:		36.6		85.0		39.9		85.7	

[1] $\gamma_1 = \log 1.25$, $\gamma_2 = \log 1.5$
[2] $\gamma_1 = \log 1.5$, $\gamma_2 = \log 2$

Table 2.1: Empirical performance of estimators for β and α under the null model (upper part of table) and empirical rejection rate for 2 d.f. score tests based on 500 simulations with 2000 individuals per simulation.

	EST.	S.E.	95% C.I.
Class model			
β_{10}	-1.547	0.265	(-2.066, -1.028)
β_{11}	-0.668	0.397	(-1.446, 0.111)
β_{12}	0.013	0.015	(-0.017, 0.042)
β_{20}	-2.134	0.667	(-3.442, -0.827)
β_{21}	0.838	0.655	(-0.447, 2.123)
β_{22}	0.002	0.017	(-0.031, 0.036)
β_{30}	0.823	0.139	(0.551, 1.096)
β_{31}	-0.884	0.180	(-1.237, -0.531)
β_{32}	-0.013	0.007	(-0.027, 0.001)
Multistate Model			
α_{01}	-2.261	0.106	(-2.469, -2.053)
α_{02}	-2.684	0.133	(-2.945, -2.423)
α_{13}	0.283	0.161	(-0.033, 0.599)
α_{23}	-1.748	0.194	(-2.128, -1.368)

Table 2.2: Results of fitting the finite mixture model under the null hypothesis (omitting HLA markers) for the occurence of sacroiliac joint damage.

	$\chi^2_{(2)}$	p		$\chi^2_{(2)}$	p		$\chi^2_{(2)}$	p
HLA-A								
A1	1.762	0.414	A2	3.482	0.175	A3†	7.152	0.028
A11	1.603	0.449	A24	0.823	0.663	A25	1.973	0.373
A26	5.881	0.053	A29†	11.99	0.003	A30	1.784	0.410
A31	0.839	0.657	A32	3.199	0.202	A33	1.621	0.445
A68	1.278	0.528	A23*	2.870	0.238	A28*	0.848	0.654
A34*	0.550	0.760	A66*	2.810	0.245	A69*	0.368	0.832
HLA-B								
B7	0.523	0.770	B8	1.015	0.602	B13	3.494	0.174
B14	3.182	0.204	B15	2.886	0.236	B62	1.916	0.384
B18	0.481	0.786	B27†	8.086	0.018	B35†	10.01	0.007
B37	0.801	0.670	B38	2.936	0.230	B39	3.314	0.191
B40	0.007	0.997	B44	0.669	0.716	B50	0.987	0.610
B51	0.221	0.895	B52	0.345	0.841	B55	0.668	0.716
B57	0.466	0.792	B58	0.291	0.865	B60	0.774	0.679
B61	0.319	0.853	B70*	2.935	0.231	B41*†	23.53	<0.001
B45*	2.741	0.254	B46*	4.585	0.101	B47*†	22.53	<0.001
B48*	2.328	0.312	B49*	3.893	0.143	B53*	0.048	0.976
B56*	1.768	0.413	B63*	1.092	0.579	B67*	0.442	0.802
HLA-C								
C1	1.189	0.552	C2†	9.095	0.011	C3	0.656	0.720
C4†	7.316	0.026	C5	0.428	0.807	C6	2.562	0.278
C7	1.792	0.408	C8	1.754	0.416	C12*	6.572	0.037
C14	0.235	0.889	C15	2.608	0.271	C16	0.822	0.663
C17*†	23.88	<0.001	C18*	1.609	0.558			

* markers of <1% in frequency

† siginificant marker.

Table 2.3: Results of applying the 2 d.f. score test for each of the HLA-A, HLA-B and HLA-C markers to the University of Toronto Psoriatic Arthritis Cohort

	$\hat{\gamma}$	S.E.($\hat{\gamma}$)	O.R.	C.I.	$p*$
Unilateral vs. None					
A3	0.905	0.398	2.473	(1.133, 5.398)	0.023
A29	-1.169	0.772	0.311	(0.068 1.411)	0.130
B27	0.680	0.450	1.974	(0.816 4.774)	0.131
B35	-1.468	0.459	0.231	(0.094, 0.566)	0.001
C2	1.309	0.598	3.701	(1.146, 11.965)	0.029
C4	-1.210	0.620	0.298	(0.088, 1.004)	0.051
C12	-0.255	0.459	1.582	(0.316, 1.904)	0.578
Bilateral vs. None					
A3	0.179	0.303	1.196	(0.660, 2.168)	0.554
A29	-0.976	0.335	0.377	(0.195, 0.727)	0.004
B27	0.692	0.305	1.999	(1.100, 3.629)	0.023
B35	-0.503	0.238	0.605	(0.379, 0.964)	0.034
C2	0.831	0.466	2.297	(0.922, 5.720)	0.074
C4	-0.380	0.226	0.684	(0.440, 1.065)	0.093
C12	0.446	0.238	1.562	(0.980, 2.489)	0.061
Bilateral vs. Unilateral					
A3	-0.726	0.326	0.484	(0.256, 0.916)	0.026
A29	0.193	0.779	1.213	(0.264, 5.579)	0.804
B27	0.012	0.361	1.012	(0.499, 2.054)	0.973
B35	0.964	0.420	2.623	(1.153, 5.972)	0.022
C2	-0.477	0.435	0.621	(0.264, 1.456)	0.273
C4	0.831	0.399	2.295	(1.049, 5.013)	0.037
C12	0.701	0.369	2.012	(0.978, 4.154)	0.058

$*$ p-values are based on Wald tests.

Table 2.4: Results of fitting a finite mixture model for sacroiliac joint damage including all significant HLA markers with a frequency of greater than 1% found from screening score tests.

Chapter 3

Marginal Mixture Models for Multivariate Interval-censored Times

3.1 Introduction

3.1.1 Background

Life history data are frequently collected for use in investigations within disciplines such as medicine. Such data arise when individuals are observed over time and information on occurrence of one or more events for these individuals are recorded on each visit. When individuals are observed at prespecified assessment times, their information on the occurrence of events are determined only at these times. When the precise time of event occurrences are unknown due to intermittent examination of a sub-clinical feature the transition times are interval-censored (Sun, 2006). Chronic diseases for example, frequently affect multiple organ systems or multiple locations of the body, which often renders multivariate censored failure times. Developments in recent years have been directed at analysis of bivariate right-censored (Shih and Louis, 1995), bivariate interval-censored data (Betensky and Finkelstein, 1999; Kim and Xue, 2002; Sun, 2006; Cook et al., 2008b), and bivariate current status (Wang and Ding, 2000). The need to accommodate dependencies between associated failure times may be to advance scientific understanding, to facilitate joint statements about two or more disease features, or simply to ensure valid inference.

Despite careful attempts to use available information to characterize multivariate disease processes, substantial unexplained variability often remains between individuals. In the analysis of lifetime data, individuals who do not experience the event of interest by the end of the study are often treated as having right-censored event time. However, if a subgroup of the individuals are not susceptible to an event of interest, it may be important to accommodate this in analyses. In order to address the potential nonsusceptible subpopulation in the context of lifetime data, mixture models or "cure-rate" models are often used. Farewell (Farewell, 1977; Farewell, 1982) pioneered the work for right-censored data, and there have been considerable further developments on mixture models during the past 20 years (Sy and Taylor, 2000; Peng et al., 2001; Peng, 2003). Chatterjee and Shih (2001) considered a bivariate model for modeling familial association in disease onset which accommodates a nonsusceptible fraction and deals with data subject to right censoring. Kim (2016) considered inference for a cure rate model with bivariate interval-censored data via an approximate likelihood.

We consider a bivariate cure rate model for interval-censored failure times. Like Chatterjee and Shih (2001) we use a pairwise odds ratio to model the association between latent susceptibility indicators and use a copula model for the dependence structure between the failure times for which individuals are susceptible. This formulation enables one to consider two aspects to the dependence structure between processes within an individual. Moreover we consider three approaches of estimation including maximum likelihood, two-stage estimation, and the use of weighted second-order estimating functions.

The remainder of this contribution is as follows. In the following subsection, we describe the data from the University of Toronto Psoriatic Arthritis Cohort that motivates this work. In Section 3.2 we define notation, formulate the model, and give the association structure. In Section 3.3, the likelihood is constructed and variance estimation is described for simultaneous and two-stage estimation. An alternative approach is also developed based on weighted second-order estimating functions. Empirical studies and an application involving the incidence of hand and foot joint damage in individuals with psoriatic arthritis are given in Section 3.4. General remarks are given in Section 3.5.

3.1.2 A Study of Joint Damage in the Hands and Feet

The Center for Prognosis Studies in Rheumatic Disease maintains a registry of patients with psoriatic arthritis who are followed according to a standard protocol with annual visits and biannual radiographic assessments. Patients with arthritis experience inflammation in and around the joints in several areas including the wrists, hands, knees, ankles, lower back, and neck. Individuals are assessed at each visit to determine the level of damage in each of 64 joints, by both radiological and clinical examination. The level of the damage is measured according to a validated modified Steinbrocker scoring system (Rahman et al., 1998) where a score of 0 represents no damage, and scores from 1 to 5 represent progressively more advanced stages of damage with states 4 and 5 representing severe damage. Individuals with 5 or more severely damaged joints are considered to have an aggressive form of the disease called arthritis mutilans.

Of clinical interest is in the relationship between the development of damage in the hand and foot joints. To study this we restrict attention to the 28 joints of the hands and 12 joints in the feet. Individuals are considered damage-free in the hands when all hand joints are in state 0, but when one or more hand joint enters state 1, they are considered to have developed damage in the hands; the time to the onset of foot damage is defined similarly. We let T_1 denote the time to the onset of damage in the hands and T_2 the time to the onset of damage in the feet. Because individuals are only seen periodically for the assessment of joint damage, T_1 and T_2 are interval-censored ; that is, we only know them to lie between the last negative assessment and the first positive assessment. For those joints not recorded to become damaged the onset times of damage are right-censored at the last assessment. Let $\mathscr{B}_{ij} = [l_{ij}, r_{ij}]$ denote the censoring interval for T_{ij} and $\mathscr{B}_i = \mathscr{B}_{i1} \times \mathscr{B}_{i2}$ be the censoring region for $T_i = (T_{i1}, T_{i2})'$; r_{ij} or $l_{ij} = \infty$ is a right censoring time for T_{ij} in individual i if they are not observed to develop damage. Figure 3.1 contains a plot of the censoring regions for eight sample individuals in the cohort. We return to this example in Section 3.4.2 when we analyze this data.

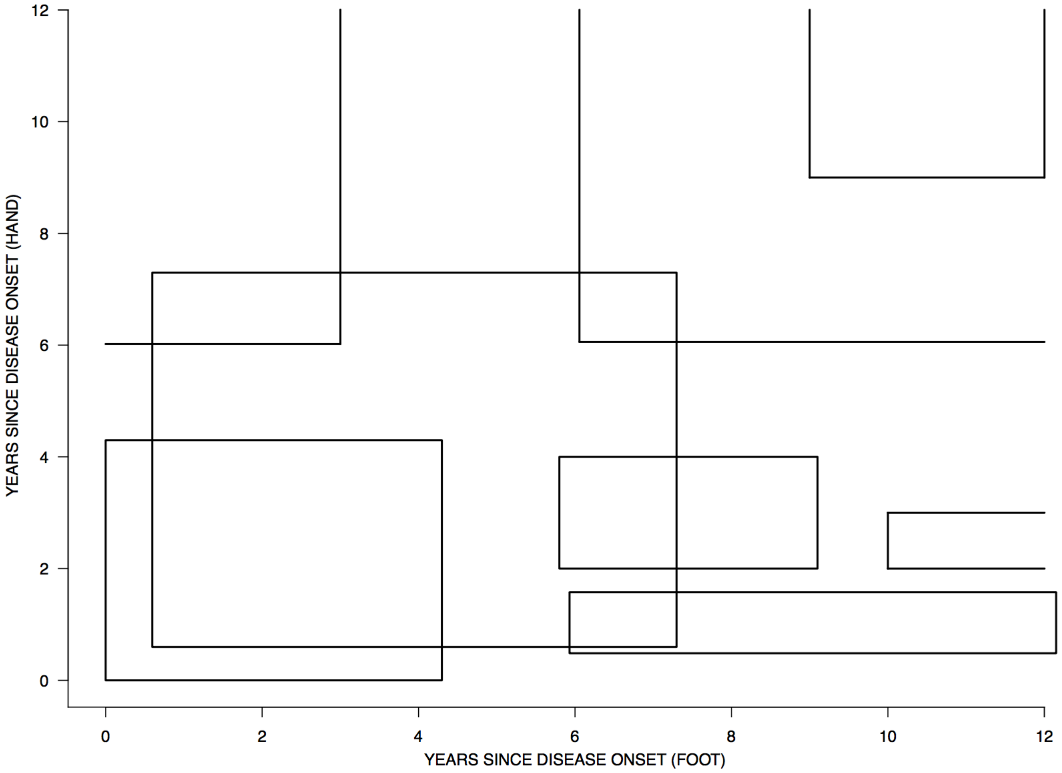


Figure 3.1: Censoring regions for the times to joint damage in the hands and feet in a sample of eight individuals from the University of Toronto Psoriatic Arthritis Cohort.

3.2 Model Formulation

Let $Z_{ij} = 1$ if individual i is susceptible to the event of interest for process j, $j = 1, 2$, and $X_i = (1, X_{i1}, ..., X_{i,p-1})'$ be a $p \times 1$ vector of fixed covariates. Let $\mu_{ij} = E(Z_{ij}|X_i; \beta_j)$ with $g_1(\mu_{ij}) = X_i' \beta_j$ where $g_1()$ is a monotonic differentiable link function mapping the interval $[0, 1]$ onto the real line and β_j is a $p \times 1$ vector of regression coefficients. To accommodate an association between Z_{i1} and Z_{i2} given X_i we construct a joint model and express the association in term of the odds ratio,

$$\psi_i = \frac{P(Z_{i1} = 1, Z_{i2} = 1|X_i)P(Z_{i1} = 0, Z_{i2} = 0|X_i)}{P(Z_{i1} = 1, Z_{i2} = 0|X_i)P(Z_{i1} = 0, Z_{i2} = 1|X_i)}$$

(Lipsitz and Laird, 1991). A second-order dependence model can be specified via $g_2(\psi_i) = v_i'\gamma$ where $g_2()$ maps the non-negative real line onto the real line, $v_i = (1, v_{i1}, ..., v_{i,q-1})'$ is a $q \times 1$ vector of covariates, and γ is a $q \times 1$ vector of regression coefficients. The resulting joint distribution of $Z_i = (Z_{i1}, Z_{i2})'$ is $P(Z_i|X_i; \beta, \gamma)$ where $\beta = (\beta_1', \beta_2')'$.

If $Z_{ij} = 1$ then individual i is at risk of event j and we let T_{ij} denote the time of the type j event and let $T_i = (T_{i1}, T_{i2})'$. We let $\mathscr{F}_j(t|X_i; \lambda_j) = P(T_{ij} > t|Z_{ij} = 1, X_i)$ denote the marginal distribution of T_{ij} given $(Z_{ij} = 1, X_i)$ indexed by λ_j, $j = 1, 2$, and $\mathscr{F}_{12}(t_1, t_2|X_i; \lambda, \phi) = P(T_{i1} > t_1, T_{i2} > t_2|Z_i = (1, 1)', X_i)$ where $\lambda = (\lambda_1', \lambda_2')'$ and ϕ is an association parameter. In keeping with the marginal specification of the model for $Z_i|X_i$, we use a copula model to model the bivariate distribution for $T_i|Z_i = (1, 1)', X_i$. Copula functions are bivariate cumulative distribution functions with uniform $[0, 1]$ margins. Specifically we use an Archimedian copula $C(u_1, u_2; \phi)$ which can be written in the form

$$C(u_1, u_2; \phi) = G^{-1}(G(u_1; \phi) + G(u_2; \phi); \phi)$$

where $G : [0, 1] \to [0, \infty]$ is a continuous, strictly decreasing and convex generator function with $G(1; \phi) = 0$. The Clayton copula has generator $G(u; \phi) = \phi^{-1}(u^{-\phi} - 1)$, while the Frank and Gumbel copulas have generators $G(u; \phi) = -\text{log}((\exp(-\phi u) - 1)/(\exp(-\phi) - 1))$ and $G(u; \phi) = (-\text{log}u)^{-\phi}$ respectively. Kendall's τ, a natural dependence measure within the Archimedian family is expressed as

$$\tau = 1 + 4\int_0^1 \frac{G(u; \phi)}{G'(u; \phi)} du$$

(Nelsen, 2006).

Consider a Clayton copula for example which has the form

$$C(u_1, u_2) = (u_1^{-\phi} + u_2^{-\phi} - 1)^{-\phi^{-1}}.$$

The joint survivor function for $T_i|Z_i = (1,1), X_i$ is obtained by letting

$$\mathscr{F}_{12}(t_1, t_2|Z_i = (1,1)', X_i) = C(\mathscr{F}_1(t_1|Z_{i1} = 1, X_i; \lambda_1), \mathscr{F}_2(t_2|Z_{i2} = 1, X_i; \lambda_2); \phi).$$

Because of the monotonic form of the survivor function, the association between u_1 and u_2 is the same as the association between $T_{i1}, T_{i2}|Z_i = (1,1)', X_i$. Note that this formulation involves a common distribution for $T_{ij}|Z_{ij}$ for $Z_{ij} = 0$ or 1. One may alternatively adopt a bivariate model with different margins so that

$$\mathscr{F}_{12}^{(2)}(t_1, t_2|Z_i = (1,1)', X_i) = C(\mathscr{F}_1^{(2)}(t_1|Z_{i1} = 1, X_i; \lambda_1^{(2)}), \mathscr{F}_2^{(2)}(t_2|Z_{i2} = 1, X_i; \lambda_2^{(2)}); \phi)$$

where $\mathscr{F}_j^{(2)}(t_j|Z_{ij} = 1, X_i; \lambda_j^{(2)})$ denotes the marginal distribution in the bivariate model which is possibly different than the respective marginal distributions in the settings where individuals are only at risk of one event.

3.3 Methods for Estimation and Inference

3.3.1 Maximum Likelihood

We now consider the analysis of the data discussed in Section 1.2. We let $D_i = (\mathscr{B}_i, X_i)$ denote the data from individual i, $i = 1, ..., n$. The full vector of parameters is $\theta = (\beta', \gamma', \lambda', \phi)'$. We partition θ here in terms of $\theta_1 = (\beta', \gamma')'$ which reflects the parameters governing the distribution of Z_i and $\theta_2 = (\lambda', \phi)'$ which governs the distribution of $T_i|Z_i$.

The observed likelihood is

$$L_i(\theta) \propto \sum_{z_i \in \mathscr{Z}_i} P(Z_i = z_i | X_i; \theta_1) P(T_i \in \mathscr{B}_i | Z_i = z_i; \theta_2). \tag{3.1}$$

and the complete likelihood is

$$\mathscr{L}_i(\theta) \propto \prod_{z_i \in \mathscr{Z}_i} \left[P(Z_i = z_i | X_i; \theta_1) P(T_i \in \mathscr{B}_i | Z_i = z_i; \theta_2) \right]^{I(Z_i = z_i)} \tag{3.2}$$

and the score functions from (3.1) can be written as

$$\begin{aligned} S_{i1}(\theta) &= E\{\mathscr{S}_1(Z_i | X_i; \theta_1) | D_i; \theta\} \\ S_{i2}(\theta) &= E\{\mathscr{S}_2(T_i \in \mathscr{B}_i | Z_i, X_i; \theta_2) | D_i; \theta\} \end{aligned} \tag{3.3}$$

for θ_1 and θ_2 respectively where $\mathscr{S}_1(Z_i|X_i;\theta_1) = \partial \text{log} \mathscr{L}_i(\theta)/\partial\theta_1$ and $\mathscr{S}_2(T_i \in \mathscr{B}_i | Z_i, X_i; \theta_2) = \partial \text{log} \mathscr{L}_i(\theta)/\partial\theta_2$. We obtain $P(Z_i = z_i | D_i; \theta)$ using Baye's rule as

$$P(Z_i = z_i | D_i; \theta) = \frac{P(T_i \in \mathscr{B}_i | Z_i; \theta_2) P(Z_i | X_i; \theta_1)}{\sum_{z_i \in \mathscr{Z}_i} P(T_i \in \mathscr{B}_i | Z_i; \theta_2) P(Z_i | X_i; \theta_1)}$$

Equations (3.3) can be solved directly, or iteratively using an EM algorithm (Dempster et al., 1977). If θ^r denotes the estimate at the r^{th} iteration, θ^{r+1} is obtained by solving

$$\begin{aligned} S_1(\theta_1; \theta^r) &= \sum_{i=1}^{n} E\{\mathscr{S}_1(Z_i | X_i; \theta_1) | D_i; \theta^r\} = 0 \\ S_2(\theta_2; \theta^r) &= \sum_{i=1}^{n} E\{\mathscr{S}_2(T_i \in \mathscr{B}_i | Z_i, X_i; \theta_2) | D_i; \theta^r\} = 0 \end{aligned}$$

Louis' method (Louis, 1982) can be used to obtain the observed information matrix based on the identity,

$$I_i(\theta) = E\left[\mathscr{J}_i(\theta) | D_i\right] - E\left[\mathscr{S}_i(\theta)\mathscr{S}_i'(\theta) | D_i\right] + S_i(\theta) S_i'(\theta),$$

where we write $\mathscr{S}_i(\theta)$ for $(\mathscr{S}_1'(Z_i|X_i;\theta_1), \mathscr{S}_2'(T_i \in \mathscr{B}_i|Z_i;\theta_2))'$ when it is not necessary to write the variables in explicitly, and $\mathscr{I}_i(\theta) = -\partial \mathscr{S}_i(\theta)/\partial\theta$ is the block-diagonal observed information for the complete data. We then sum up over all individuals in order to obtain $I(\theta)$.

3.3.2 Two-stage Estimation

Instead of simultaneously estimating all the parameters as in the full likelihood function (3.1), a two-stage procedure can be adopted. Shih and Louis (1995) suggested such an approach where we may estimate the marginal parameters and the association parameters separately. To this end, we partition the parameter vector θ as $(\alpha_1', \alpha_2')'$ where $\alpha_1 = (\beta', \lambda')'$ and $\alpha_2 = (\gamma', \phi)'$. At stage 1 we estimate $\alpha_1 = (\beta', \lambda')'$ by assuming independence between the two processes. The association parameters $\alpha_2 = (\gamma', \phi)'$ are then estimated at stage 2 with the parameters governing the marginal distribution set at the values obtained at stage 1. We describe this in more detail in what follows.

At stage 1, let $\delta_{ij} = 1$ if individual i is known to have experienced the event for process j. We can then construct the following observed likelihood,

$$L_{i1}(\alpha_1) \propto \prod_{j=1}^{2} (\mu_{ij}\left[\mathscr{F}_{ij}(l_{ij}) - \mathscr{F}_{ij}(r_{ij})\right])^{\delta_{ij}} \left[\mu_{ij}\mathscr{F}_{ij}(r_{ij}) + (1-\mu_{ij})\right]^{(1-\delta_{ij})} \tag{3.4}$$

where $[l_{ij}, r_{ij})$ denotes the censored interval for individual i for process j. Let $\widetilde{\alpha}_1 = (\widetilde{\beta}', \widetilde{\lambda}')'$ denote the value that maximizes (3.4).

At stage 2, α_2 can be estimated by inputing the estimate $\tilde{\alpha_1}$ from stage 1 into the observed likelihood

$$L_{i2}(\widetilde{\alpha}_1, \alpha_2) \propto \sum_{z_i \in \mathscr{Z}_i} P(Z_i = z_i|X_i; \tilde{\beta}, \gamma) P(T_i \in \mathscr{B}_i | Z_i = z_i; \tilde{\lambda}, \phi). \tag{3.5}$$

Let $\bar{S}_{i1}(\alpha_1) = \partial \text{log} L_{i1}(\alpha_1)/\partial\alpha_1$ and $\bar{S}_{i2}(\alpha) = \partial \text{log} L_{i2}(\widetilde{\alpha}_1, \alpha_2)/\partial\alpha_2$ be the observed data score functions from stage 1 and stage 2 respectively. Standard estimating function theory gives

$$\sqrt{n}\begin{pmatrix}\widetilde{\alpha}_1 - \alpha_1 \\ \widetilde{\alpha}_2 - \alpha_2\end{pmatrix} \to \text{MVN}(0, \bar{\mathscr{A}}^{-1}(\theta)\bar{\mathscr{B}}(\theta)\left[\bar{\mathscr{A}}^{-1}(\theta)\right]'),$$

with $\bar{\mathscr{A}}(\theta) = E\{-\partial\bar{S}_i(\theta)/\partial\theta\}$ and $\bar{\mathscr{B}}(\theta) = E\{\bar{S}_i(\theta)\bar{S}_i'(\theta)\}$, where

$$\widehat{A}(\tilde{\theta}) = -n^{-1}\sum_{i=1}^{n}\frac{\partial\bar{S}_i(\theta)}{\partial\theta'}|_{\theta=\tilde{\theta}}$$
$$\widehat{B}(\tilde{\theta}) = n^{-1}\sum_{i=1}^{n}\bar{S}_i(\theta)\bar{S}_i'(\theta)|_{\theta=\tilde{\theta}}$$

3.3.3 Weighted Second-Order Estimating Equations

In this section we describe weighted second-order estimating equations which have the form of expected complete data estimating equations which would be employed if the susceptibility indicators were known. Given their use of second-order moments they are expected to yield more efficient estimates and the opportunity to consider different structures means this framework enables one to investigate robustness and efficiency trade-offs.

The estimating equation we propose for θ_1 is of the form

$$U_1(\theta) = \sum_{i=1}^{n}\sum_{z_i\in\mathscr{Z}_i} P(Z_i = z_i|D_i;\theta) \;\cdot\; H_{i1}'(\theta_1)\,\Sigma_{i1}^{-1}(\theta_1)\begin{pmatrix}Z_i - \mu_{i1} \\ W_{i1} - \omega_{i1}\end{pmatrix} \tag{3.6}$$

where $Z_i = (Z_{i1}, Z_{i2})'$, $\mu_{i1} = E(Z_i|X_i;\beta)$, $W_{i1} = Z_{i1}Z_{i2}$, $\omega_{i1} = E(W_{i1}|X_i;\theta_1)$,

$$H_{i1}(\theta_1) = \begin{pmatrix}\partial\mu_{i1}/\partial\beta' & \partial\mu_{i1}/\partial\gamma' \\ \partial\omega_{i1}/\partial\beta' & \partial\omega_{i1}/\partial\gamma'\end{pmatrix}\;, \quad \text{and} \quad \Sigma_{i1}(\theta_1) = \begin{pmatrix}\text{cov}(Z_i|X_i) & \text{cov}(Z_i, W_{i1}|X_i) \\ \text{cov}(W_{i1}, Z_i|X_i) & \text{var}(W_{i1}|X_i)\end{pmatrix}.$$

The second set of estimating equations for θ_2 has the form

$$U_2(\theta) = \sum_{i=1}^{n} \sum_{z_i \in \mathscr{Z}_i} P(Z_i = z_i | D_i; \theta) \ \cdot \ H'_{i2}(\theta_2) \bigtriangleup_i \Sigma_{i2}^{-1} \ (\theta_2) \begin{pmatrix} Y_i - \mu_{i2} \\ W_{i2} - \omega_{i2} \end{pmatrix} \tag{3.7}$$

where $Y_{ik} = I(T_{ik} \in \mathscr{B}_{ik})$ and $Y_i = (Y_{i1}, Y_{i2})'$, $\mu_{i2} = E(Y_i|Z_i; \lambda)$, $W_{i2} = Y_{i1}Y_{i2}$, $\omega_{i2} = E(W_{i2}|Z_i; \theta_2)$,

$$H_{i2}(\theta_2) = \begin{pmatrix} \partial\mu_{i2}/\partial\lambda' & \partial\mu_{i2}/\partial\phi \\ \partial\omega_{i2}/\partial\lambda' & \partial\omega_{i2}/\partial\phi \end{pmatrix}, \qquad \Sigma_{i2}(\theta_2) = \begin{pmatrix} \text{cov}(Y_i|Z_i) & \text{cov}(Y_i, W_{i2}|Z_i) \\ \text{cov}(W_{i2}, Y_i|Z_i) & \text{var}(W_{i2}|Z_i) \end{pmatrix},$$

and

$$\bigtriangleup_i = \begin{pmatrix} z_{i1} & 0 & 0 \\ 0 & z_{i2} & 0 \\ 0 & 0 & z_{i1}z_{i2} \end{pmatrix}.$$

The derivative matrix $H_{i2}(\theta_2)$ and covariance matrix $\Sigma_{i2}(\theta_2)$ are analogous to those of (3.6) but are given by the underlying model for the failure times. If Z_i were known, the matrix $\bigtriangleup_i$ would ensure that the appropriate elements of this estimating equation contributed information about the marginal and association parameters of the multivariate failure time distribution; as Z_i is unknown we take the conditional expectation over the possible values of Z_i.

We let $U_{i1}(\theta)$ and $U_{i2}(\theta)$ denote the contribution from the ith individual to (3.6) and (3.7) respectively and $U_i(\theta) = (U'_{i1}(\theta), U'_{i2}(\theta))'$. Then $\sum_{i=1}^{n} U_i(\theta) = 0$ can be solved for θ simultaneously, or one can exploit the weighted structure of each set of equations (3.6) and (3.7) and solve them iteratively. In this case at the kth step we insert an estimate $\hat{\theta}^{k-1}$ from the $(k-1)$st step into $P(Z_i|D_i; \theta)$ in (3.6) and (3.7) and solve them for $\hat{\theta}_1^k$ and $\hat{\theta}_2^k$ respectively. Algorithms for GEE2 can be exploited for estimation of θ_1 at each step by the creation of a pseudo-dataframe containing multiple lines per individual corresponding to each realization of Z_i with weights based on $P(Z_i|D_i;\hat{\theta}^{k-1})$. The same approach can be employed in principle for (3.7) but the derivative and covariance matrices are unique in this setting (e.g. the moments are determined by the multivariate failure time model and the observation times) and would require specialized coding; a similar situation is described by Tolusso and Cook (2009).

Note that when solving the set the estimating equations, robustness comes in when we use a diagonal $H_i(\theta)$ matrix, because we are not trying to draw any information about the marginals from the association structures. To simplify calculations, we have used a diagonal $H_i(\theta)$ and $\Sigma_i(\theta)$ matrix in our simulations and what follows. Subject to correct specification of the conditional moments, (3.6) and (3.7) are unbiased estimating functions, so the estimator $\breve{\theta}_1$ and $\breve{\theta}_2$ solving the two estimating equations by setting them to zero is consistent with an asymptotic normal distribution

$$\sqrt{n}(\breve{\theta}-\theta) \rightarrow \mathrm{N}(0, \bar{\bar{\bar{\mathscr{A}}}}^{-1}(\theta)\bar{\bar{\bar{\mathscr{B}}}}(\theta)\left[\bar{\bar{\bar{\mathscr{A}}}}^{-1}(\theta)\right]'),$$

with $\bar{\bar{\bar{\mathscr{A}}}}(\theta) = E\left[-\partial U_i(\theta)/\partial\theta'\right]$ and $\bar{\bar{\bar{\mathscr{B}}}}(\theta) = E\left[U_i(\theta)U_i'(\theta)\right]$, where

$$\widehat{A}(\breve{\theta}) = -n^{-1}\sum_{i=1}^{n}\frac{\partial U_i(\theta)}{\partial\theta'}|_{\theta=\breve{\theta}}$$
$$\widehat{B}(\breve{\theta}) = n^{-1}\sum_{i=1}^{n}U_i(\theta)U_i'(\theta)|_{\theta=\breve{\theta}}$$

3.4 Empirical Studies and Application

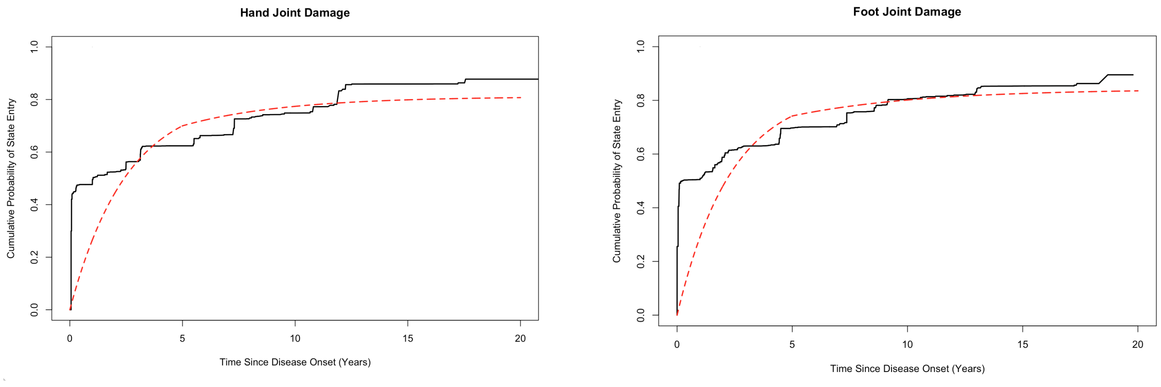
3.4.1 Simulation Studies

To gain some insights on the empirical performances of the three methods in Section 3, simulation studies are performed. We set the covariate of interest $X_i \sim$ Bern(0.5) with the corresponding link function $\text{logit}(P(Z_{ij} = 1|X_i; \beta_j)) = \beta_{j0} + \beta_{j1} X_i$. We let $P(Z_{ij} = 1|X_i = 1; \beta_j) = 0.66$, $\beta_j = (\beta_{j0}, \beta_{j1})'$, $j = 1, 2$, where $\beta_{j1} = \text{log}1.5$. The association between the susceptibility indicators (i.e. the log odds ratio) is then set as $\log \psi = \log 1.5$ and $\log 3$. For generating the failure times, we suppose $T_{ij}|Z_{ij} = 1, X_i \sim \text{Weibull}(\lambda_j)$, $\lambda_j = (\lambda_{j1}, \lambda_{j2})'$, and hence $P(T_{ij} > t|Z_{ij} = 1, X_i; \lambda_j) = \exp(-\lambda_{j1} t)^{\lambda_{j2}}$. The copula for the bivariate failure time is then $\mathscr{F}_{12}(t_1, t_2|Z_i = (1, 1), X_i) = C(\mathscr{F}_1(t_1; \lambda_1), \mathscr{F}_2(t_2; \lambda_2); \phi)$, with Kendall's $\tau = 0.3$ and 0.6. For simplicity we consider an exponential margin with $\lambda_{j2} = 1, j = 1, 2$. We specify λ_{j1} such that $P(T_{ij} < A|Z_{ij} = 1) = 0.9$, where A is the administrative censoring time, and set $A = 1$ without loss of generality. To generate interval censored data, we let $\{N_i(s), 0 < s\}$ denote the counting process for the assessments and let $\{N_i(s), 0 < s\} \sim \text{Poisson}(\rho)$. We set $\rho = 10$ and 20 to correspond to an average of 10 and 20 visits over $(0, 1]$ respectively. In the analyses a piecewise-constant baseline hazard was adopted for each component failure time model with cut-points 0.25 and 0.50.

The results presented in Tables 3.1 and 3.2 are from the analysis of 500 simulated samples of $n = 1000$ individuals each with and average of 10 and 20 visits respectively; we comment here on the results of Table 3.1 . We see that empirical biases of all methods are negligible, there is excellent agreement between the empirical and average robust standard errors, and the empirical coverage probability are close to the nominal 95% level. The estimators from the two-stage procedure are less efficient than the maximum likelihood estimates, but the efficiency of estimators from the weighted estimating equation approach is remarkably good. Findings from Table 3.2 are similar with slightly smaller standard errors.

3.4.2 Co-Occurrence of Damage in the Extremities in Psoriatic Arthritis

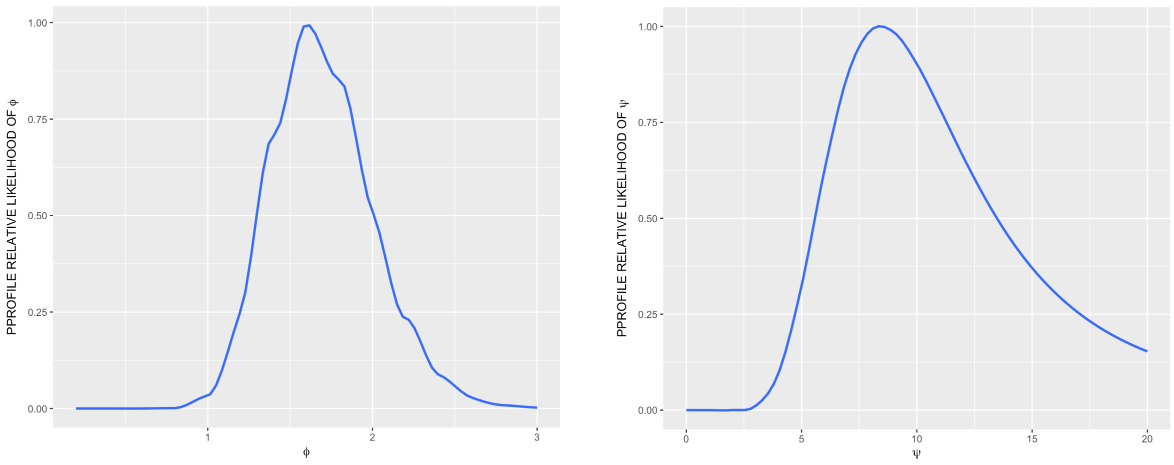
Our data consist of n=657 patients, each assessed at multiple visits to determine the level of radiological damage in the two disease processes, i.e., hands and feet. Every joint is considered to be at state 0 at the time of disease onset. In the regression model we include sex as a covariate with male as the reference level. After inspecting the quantiles of the distribution of the assessment times, we consider a two-piece piecewise-constant hazards model with a cut-point at 5 years post-onset so the hazard is constant over the interval $[0, 5)$ and $[5, \infty)$. We apply maximum likelihood, two-stage estimation, and the estimating equations approach of Section 3.3; the results in Table 3.3 reveal that the three methods yield estimates and standard errors with similar results. Figure 3.2 shows the fitted piecewise model compared to the nonparametric estimates, we see a good agreement between the two. Females are shown to have lower risk of developing damage in the hands and feet joints compared to males. Moreover, based on maximum likelihood estimates, the odds ratio for the association in the susceptibility is 8.440 with a 95% CI (3.229, 13.651). Among individuals who are susceptible to damage in both locations, the association in the onset times is estimated based on Kendall's τ to be 0.449 with a 95% CI (0.212, 0.686). One dimensional profile relative likelihood plots of ϕ and ψ are shown in Figure 3.3. Two dimensional contour profile relative likelihood plot showing the relationship between the two association parameters is also shown in Figure 3.4.



(a)

(b)

Figure 3.2: Estimates from the fitted model (red) via piecewise maximum likelihood vs nonparametric estimates (black) of $\mathscr{F}(t) = P(T \leq t)$ for hand and foot damage.



(a) Profile relative likelihood plot of ϕ

(b) Profile relative likelihood plot of ψ

Figure 3.3: One dimensional profile relative likelihood plots of ϕ and ψ for onset of damage in the hands and feet.

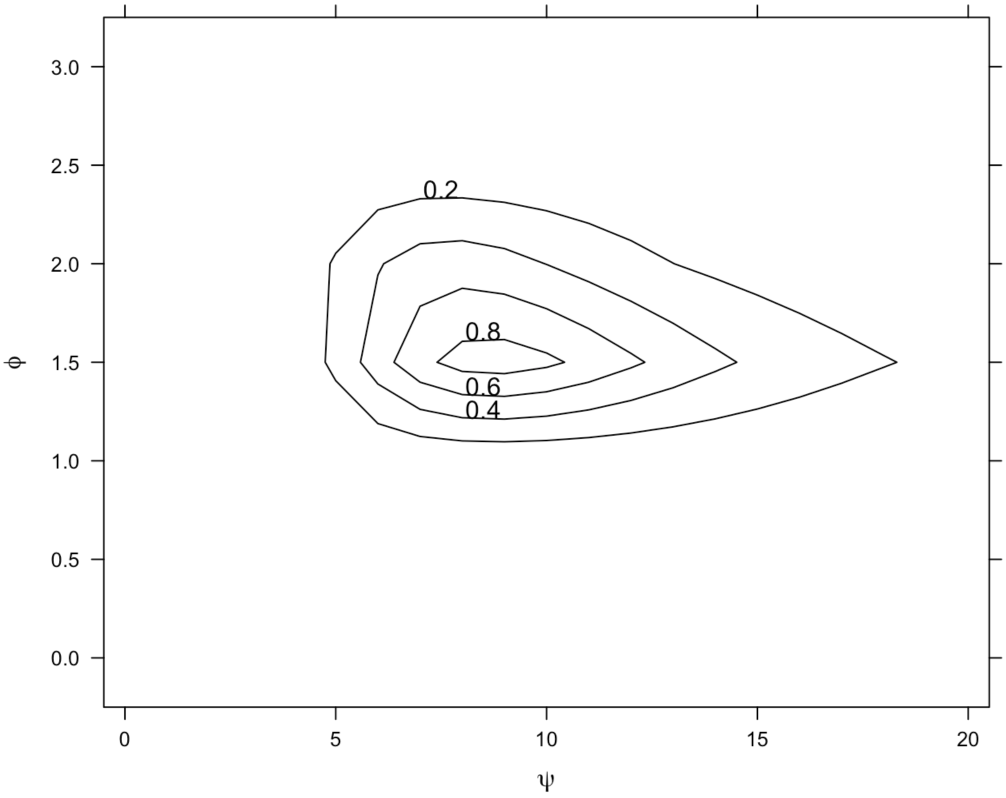


Figure 3.4: Two dimensional countour profile relative likelihood plot of ϕ and ψ for onset of damage in the hands and feet.

3.5 Discussion

In this paper we have developed flexible methods for modeling multivariate interval-censored data which accommodates the possibility that some individuals will be nonsusceptible to one or more of the conditions of interest. This framework enables one to decompose dependence measures into one component for susceptibility and one component for the dependence in the failure times given joint susceptibility. There are numerous applications where this framework can yield useful insight into disease processes. In diabetes some individuals do not experience significant complications, some develop nephropathy, some develop retinopathy, and some develop both. In settings where genes play a role in susceptibility for the different complications the mover-stayer formulation seems appropriate, particularly when variable follow-up or assessment times make classification of individuals difficult. The model may also be used in the analysis of family data when interest lies in modeling within-family dependence to gain insight into the genetic basis for disease. It may be quite natural to examine the effect of genetic markers on the susceptibility indicators rather than the failure times themselves.

	$\psi = 1.5$								$\psi = 3$							
	$\tau = 0.3$				$\tau = 0.6$				$\tau = 0.3$				$\tau = 0.6$			
	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
								Maximum Likelihood								
β_{10}	0.005	0.141	0.143	94.9	<0.001	0.126	0.124	95.4	0.003	0.140	0.137	94.3	-0.002	0.128	0.128	94.5
β_{11}	0.017	0.168	0.167	94.9	0.014	0.164	0.162	94.1	0.011	0.158	0.160	96.4	0.010	0.156	0.156	95.3
β_{20}	0.010	0.139	0.142	96.8	0.009	0.128	0.135	97.0	0.004	0.138	0.141	96.6	-0.004	0.126	0.126	96.2
β_{21}	0.015	0.158	0.159	95.3	0.012	0.153	0.157	95.9	0.021	0.156	0.157	95.3	0.020	0.153	0.156	95.0
$\log\psi$	0.002	0.205	0.206	95.0	-0.020	0.224	0.219	95.7	0.016	0.204	0.205	96.4	-0.008	0.205	0.206	94.8
$\log\lambda_1^{(1)}$	-0.013	0.116	0.114	93.1	-0.011	0.102	0.105	95.0	-0.011	0.115	0.112	93.4	-0.010	0.104	0.105	94.6
$\log\lambda_1^{(2)}$	-0.014	0.117	0.115	94.1	-0.011	0.108	0.104	93.7	-0.021	0.117	0.114	93.5	-0.016	0.106	0.105	94.5
$\log\lambda_1^{(3)}$	-0.002	0.186	0.188	96.3	<0.001	0.156	0.153	94.8	0.001	0.172	0.174	95.1	0.004	0.145	0.142	93.2
$\log\lambda_2^{(1)}$	-0.009	0.109	0.111	95.2	-0.011	0.099	0.100	95.3	-0.007	0.113	0.112	93.9	-0.010	0.102	0.100	93.9
$\log\lambda_2^{(2)}$	-0.015	0.119	0.118	94.7	-0.007	0.106	0.104	94.9	-0.021	0.115	0.116	96.4	-0.011	0.099	0.102	96.7
$\log\lambda_2^{(3)}$	-0.007	0.175	0.173	93.9	-0.012	0.153	0.151	93.9	-0.010	0.169	0.167	93.2	-0.006	0.143	0.146	95.0
$\log\phi$	<0.001	0.183	0.185	96.1	0.005	0.114	0.118	95.7	0.002	0.182	0.182	94.0	0.012	0.113	0.115	95.3
								Weighted Estimating Equations								
β_{10}	0.006	0.145	0.144	95.0	< 0.001	0.128	0.130	95.4	0.003	0.140	0.140	96.2	-0.002	0.129	0.128	94.6
β_{11}	0.017	0.168	0.163	94.4	0.014	0.165	0.157	93.6	0.011	0.159	0.159	95.6	0.010	0.156	0.156	95.2
β_{20}	0.008	0.137	0.136	96.0	0.007	0.129	0.132	96.6	0.003	0.138	0.142	96.6	0.004	0.126	0.128	96.2
β_{21}	0.014	0.157	0.162	97.0	0.012	0.153	0.157	96.2	0.022	0.156	0.159	95.4	0.020	0.153	0.157	95.4
$\log\psi$	0.002	0.206	0.200	94.8	-0.017	0.226	0.215	95.8	0.015	0.205	0.201	94.4	-0.007	0.207	0.202	95.6
$\log\lambda_1^{(1)}$	-0.014	0.121	0.123	93.4	-0.012	0.112	0.107	94.4	-0.013	0.120	0.120	93.2	-0.011	0.115	0.117	93.4
$\log\lambda_1^{(2)}$	-0.016	0.118	0.120	94.6	-0.012	0.112	0.113	95.0	-0.021	0.118	0.119	95.8	-0.019	0.113	0.114	95.6
$\log\lambda_1^{(3)}$	-0.004	0.191	0.188	94.2	0.001	0.164	0.164	94.2	0.003	0.172	0.174	94.2	0.006	0.148	0.150	95.8
$\log\lambda_2^{(1)}$	-0.010	0.114	0.114	93.8	-0.014	0.111	0.108	94.8	-0.009	0.117	0.114	93.8	-0.010	0.111	0.107	93.0
$\log\lambda_2^{(2)}$	-0.014	0.120	0.121	95.0	-0.006	0.111	0.113	96.4	-0.021	0.117	0.119	96.6	-0.012	0.104	0.103	96.7
$\log\lambda_2^{(3)}$	-0.003	0.173	0.174	95.0	-0.007	0.159	0.157	94.2	-0.008	0.171	0.176	94.4	-0.005	0.147	0.151	95.2
$\log\phi$	-0.011	0.184	0.185	96.2	-0.009	0.115	0.115	95.4	-0.006	0.183	0.179	95.8	0.002	0.112	0.111	95.4
								Two-stage								
β_{10}	0.023	0.185	0.183	95.2	0.023	0.185	0.184	95.0	0.017	0.169	0.170	96.7	0.018	0.170	0.170	96.4
β_{11}	0.026	0.184	0.184	94.4	0.026	0.184	0.184	94.6	0.018	0.161	0.167	96.1	0.017	0.162	0.167	96.0
β_{20}	0.014	0.155	0.157	96.6	0.022	0.170	0.176	96.0	0.013	0.157	0.159	96.5	0.015	0.172	0.175	96.6
β_{21}	0.018	0.160	0.157	94.6	0.019	0.159	0.162	96.6	0.027	0.165	0.169	95.3	0.027	0.165	0.171	94.4
$\log\psi$	0.001	0.205	0.205	95.8	-0.040	0.270	0.261	96.6	0.023	0.213	0.216	96.7	-0.011	0.221	0.221	96.0
$\log\lambda_1^{(1)}$	-0.021	0.131	0.127	94.6	-0.021	0.131	0.125	93.6	-0.020	0.131	0.132	95.3	-0.019	0.128	0.123	94.4
$\log\lambda_1^{(2)}$	-0.024	0.128	0.131	96.0	-0.024	0.128	0.132	95.8	-0.029	0.127	0.123	93.7	-0.030	0.128	0.132	95.6
$\log\lambda_1^{(3)}$	-0.020	0.223	0.215	93.8	-0.020	0.223	0.217	93.8	-0.012	0.208	0.205	94.9	-0.013	0.209	0.216	94.6
$\log\lambda_2^{(1)}$	-0.013	0.123	0.123	95.4	-0.019	0.124	0.126	95.0	-0.013	0.124	0.123	95.1	-0.018	0.125	0.125	95.0
$\log\lambda_2^{(2)}$	-0.017	0.125	0.132	95.8	-0.013	0.127	0.132	96.4	-0.026	0.126	0.123	93.6	-0.022	0.117	0.132	96.8
$\log\lambda_2^{(3)}$	-0.008	0.194	0.205	95.8	-0.017	0.214	0.219	96.4	-0.016	0.202	0.206	96.7	-0.020	0.216	0.217	95.2
$\log\phi$	-0.029	0.190	0.203	96.6	-0.042	0.136	0.151	96.0	-0.026	0.185	0.195	96.9	-0.034	0.120	0.138	96.6

Table 3.1: Empirical performance of estimators for 500 simulations with 1000 individuals per simulation under a three-piece piecewise constant hazards bivariate mixture model with $E\{N_i(E)\} = 10$.

	$\psi = 1.5$								$\psi = 3$							
	$\tau = 0.3$				$\tau = 0.6$				$\tau = 0.3$				$\tau = 0.6$			
	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
								Maximum Likelihood								
β_{10}	0.012	0.133	0.137	95.9	0.007	0.121	0.124	95.2	0.014	0.127	0.128	95.3	0.011	0.118	0.120	94.3
β_{11}	0.007	0.155	0.157	95.2	0.006	0.152	0.153	95.3	0.005	0.148	0.147	94.4	0.004	0.144	0.142	94.0
β_{20}	0.013	0.130	0.134	96.2	0.013	0.123	0.123	93.9	0.008	0.127	0.125	92.6	0.003	0.118	0.119	95.3
β_{21}	0.008	0.153	0.153	95.4	0.009	0.153	0.153	95.9	0.016	0.152	0.156	97.0	0.016	0.152	0.153	96.0
$\log \psi$	0.003	0.180	0.181	96.6	-0.013	0.192	0.194	95.0	0.024	0.185	0.187	97.0	0.007	0.188	0.189	96.7
$\log \lambda_1^{(1)}$	-0.008	0.098	0.102	94.8	-0.007	0.090	0.089	93.2	-0.005	0.094	0.096	96.7	-0.003	0.088	0.090	95.9
$\log \lambda_1^{(2)}$	-0.013	0.106	0.109	96.2	-0.010	0.093	0.095	94.1	-0.010	0.099	0.097	93.3	-0.008	0.091	0.093	95.3
$\log \lambda_1^{(3)}$	-0.011	0.162	0.164	95.3	-0.008	0.139	0.136	93.7	-0.010	0.152	0.151	94.2	-0.009	0.130	0.129	95.0
$\log \lambda_2^{(1)}$	-0.008	0.097	0.099	95.0	-0.006	0.087	0.087	93.8	-0.012	0.094	0.092	94.0	-0.006	0.088	0.090	96.1
$\log \lambda_2^{(2)}$	-0.005	0.104	0.102	94.1	-0.004	0.092	0.092	94.7	-0.006	0.105	0.104	95.1	-0.006	0.092	0.092	95.3
$\log \lambda_2^{(3)}$	-0.008	0.153	0.154	94.7	-0.012	0.136	0.136	95.5	-0.018	0.150	0.152	95.2	-0.012	0.127	0.130	95.8
$\log \phi$	-0.003	0.150	0.151	96.7	0.006	0.096	0.098	95.3	<0.001	0.145	0.150	96.7	0.005	0.092	0.092	95.3
								Weighted Estimating Equations								
β_{10}	0.009	0.133	0.132	95.8	0.006	0.123	0.122	95.4	0.012	0.128	0.128	96.4	0.008	0.119	0.119	95.7
β_{11}	0.007	0.154	0.155	95.6	0.005	0.152	0.153	95.2	0.006	0.147	0.154	95.8	0.004	0.145	0.152	96.1
β_{20}	0.013	0.131	0.130	95.6	0.014	0.127	0.122	94.2	0.007	0.128	0.129	96.0	0.002	0.119	0.119	95.9
β_{21}	0.009	0.155	0.155	96.8	0.010	0.154	0.153	96.6	0.016	0.151	0.155	96.8	0.015	0.152	0.152	95.9
$\log \psi$	0.004	0.181	0.183	95.2	-0.012	0.194	0.196	96.4	0.022	0.186	0.190	97.0	0.008	0.188	0.189	94.9
$\log \lambda_1^{(1)}$	-0.006	0.101	0.098	94.6	-0.005	0.097	0.093	95.0	-0.005	0.097	0.097	94.5	-0.004	0.094	0.092	94.3
$\log \lambda_1^{(2)}$	-0.011	0.107	0.105	94.4	-0.009	0.101	0.098	94.0	-0.009	0.102	0.103	94.6	-0.007	0.097	0.097	94.5
$\log \lambda_1^{(3)}$	-0.008	0.161	0.157	93.2	-0.006	0.142	0.137	94.2	-0.007	0.152	0.151	94.0	-0.005	0.135	0.131	94.7
$\log \lambda_2^{(1)}$	-0.008	0.102	0.097	93.6	-0.008	0.096	0.093	93.4	-0.012	0.097	0.097	94.6	-0.007	0.095	0.092	95.3
$\log \lambda_2^{(2)}$	-0.005	0.106	0.104	95.8	-0.007	0.099	0.098	94.2	-0.002	0.107	0.103	93.8	-0.005	0.099	0.097	95.5
$\log \lambda_2^{(3)}$	-0.008	0.157	0.155	93.6	-0.013	0.145	0.136	93.4	-0.017	0.152	0.152	93.2	-0.009	0.130	0.132	95.5
$\log \phi$	-0.010	0.150	0.152	95.4	-0.005	0.097	0.096	95.6	-0.007	0.146	0.149	96.4	-0.003	0.093	0.093	95.7
								Two-stage								
β_{10}	0.011	0.145	0.143	96.4	0.011	0.146	0.144	96.0	0.021	0.147	0.146	96.8	0.021	0.147	0.144	96.4
β_{11}	0.008	0.156	0.158	96.4	0.008	0.156	0.158	96.2	0.008	0.153	0.158	96.8	0.008	0.153	0.159	96.8
β_{20}	0.021	0.147	0.146	95.6	0.032	0.167	0.163	95.0	0.016	0.147	0.147	96.0	0.010	0.141	0.144	96.2
β_{21}	0.011	0.158	0.160	96.8	0.016	0.161	0.163	96.8	0.018	0.153	0.156	97.0	0.016	0.154	0.158	96.4
$\log \psi$	-0.002	0.181	0.190	95.6	-0.029	0.218	0.226	96.8	0.029	0.190	0.204	96.8	0.002	0.196	0.210	95.6
$\log \lambda_1^{(1)}$	-0.007	0.105	0.103	95.4	-0.007	0.105	0.104	95.6	-0.009	0.101	0.104	95.0	-0.009	0.101	0.104	94.4
$\log \lambda_1^{(2)}$	-0.011	0.113	0.111	95.0	-0.011	0.113	0.112	95.4	-0.013	0.111	0.111	95.4	-0.013	0.111	0.112	94.4
$\log \lambda_1^{(3)}$	-0.007	0.175	0.175	94.6	-0.007	0.176	0.178	95.2	-0.014	0.177	0.177	94.4	-0.014	0.178	0.177	94.6
$\log \lambda_2^{(1)}$	-0.011	0.110	0.105	94.0	-0.015	0.108	0.106	93.6	-0.016	0.105	0.104	95.2	-0.010	0.104	0.104	95.2
$\log \lambda_2^{(2)}$	-0.009	0.113	0.113	95.4	-0.016	0.116	0.115	95.4	-0.006	0.117	0.112	94.0	-0.009	0.109	0.112	95.4
$\log \lambda_2^{(3)}$	-0.014	0.180	0.178	95.4	-0.029	0.197	0.191	94.0	-0.024	0.180	0.179	96.4	-0.011	0.172	0.178	96.0
$\log \phi$	-0.021	0.150	0.153	97.4	-0.023	0.189	0.182	95.6	-0.019	0.148	0.155	96.7	-0.020	0.157	0.162	95.8

Table 3.2: Empirical performance of estimators for 500 simulations with 1000 individuals per simulation under a three-piece piecewise constant hazards bivariate mixture model with $E\{N_i(E)\} = 20$.

	EST.	SE	EST.	SE	EST.	SE
	ML		Two-Stage		WGEE	
β_{10}	1.622	0.191	1.656	0.203	1.518	0.184
β_{11}	-0.361	0.252	-0.367	0.258	-0.355	0.237
β_{20}	1.938	0.245	1.690	0.212	1.681	0.199
β_{21}	-0.559	0.281	-0.487	0.240	-0.492	0.241
$\log\psi$	2.133	0.351	2.051	0.314	2.023	0.294
$\log\lambda_1^{(1)}$	-0.924	0.199	-0.851	0.211	-0.778	0.215
$\log\lambda_2^{(1)}$	-1.520	0.080	-1.566	0.209	-1.512	0.117
$\log\lambda_1^{(2)}$	-0.855	0.184	-0.623	0.131	-0.618	0.124
$\log\lambda_2^{(2)}$	-1.719	0.088	-1.403	0.231	-1.503	0.126
$\log\phi$	0.487	0.186	0.444	0.237	0.495	0.163

Table 3.3: Estimates and standard errors from fitting a two-piece piecewise constant hazards bivariate mixture model for the onset of damage in the hands and feet.

Chapter 4

Analysis of Aggregate Data from Clustered Multistate Processes

4.1 Introduction

4.1.1 Literature Review

Progressing processes arise in studying the development of organisms through different stages of maturation. The life cycle of insects, for example, is characterized by a series of instars stages until maturation (Borror and White, 1970). Multistate models offer an appealing and powerful framework for modeling these and other progressive processes when the stages can be meaningfully characterized into a finite number of disjoint states. Among individuals with hepatitis C infection for example, the extent of liver damage is quantified using a five point scale with state 1 representing no fibrosis, states 2 to 4 representing increasing degrees of fibrosis and state 5 representing cirrhosis (Sweeting et al., 2006). Joint damage (Gladman and Chandran, 2010) can also be viewed in this way as patients with arthritis progress through a sequence of damage stages until joint destruction. In many such instances it is not apparent when a change of state has happened and so the precise times of transitions between states are unavailable. Such data are referred to as panel data or multistate interval-censored data (Kalbfleisch and Lawless, 1985). Kalbfleisch and Lawless (1985) developed a maximum likelihood approach under Markov assumption and is implemented in the *msm* package in R by Jackson (2011).

In many instances it is difficult to identify individual organisms under study in which case the data available are often aggregated in some way. This happens, for example, when individual are not indistinguishable as in some insect studies, and hence they cannot be tracked over time. A study of the lifecycle of *Chorthippus parallelus* (Munholland and Kalbfleisch, 1991) involved the recording of aggregate data on insects in different stages due to such an identifiability problem. Studies on the development of plants through a series of stages (e.g. *Arabidopsis thaliana* (Gouno et al., 2011)) may only record the aggregate data. In this article we consider the metamorphosis of *Lepidopsetta polyxystra* where the identification of larvae and tracking is difficult.

The focus in this article will be in clustered multistate aggregate data. Frequently, individuals within the same cluster tend to be similar comparing to the other clusters. As a result, the heterogeneity between clusters should be taken into consideration when modeling the multistate processes. Several articles in the literature have dealt with clustered but individual-level multistate data.

The focus here is on modeling clustered multistate aggregate data with the focus on both marginal models and random effect models. Diao and Cook (2014) considered composite likelihood for joint analysis of multiple multistate progressive processes via copula-based marginal models. Satten (1999) considered a conditionally time-homogeneous Markov models for progressive disease for panel data with random effects incorporated. Random effects on clustered progressive disease processes under intermittent observation schemes have also been developed recently by Cook et al. (2004) and Sutradhar and Cook (2008).

The remainder of this contribution is organized as follows. In the next subsection, we describe a study on the growth and development of *Lepidopsetta polyxystra*. In Section 4.2, we define notation and describe the likelihood construction with clustered multistate aggregate data under a Markov assumption. In Section 4.3 we introduce dependence modeling via copulas, propose a composite likelihood, derive a procedure for robust variance estimation, and evaluate the methods by simulation. A random effect model is formulated in Section 4.4 in which a cluster-level random effect accommodates heterogeneity in the growth rates between clusters; simulation studies are also carried out. Both methods are applied to data on the growth and development of *Lepidopsetta polyxystra* in Section 4.5 and concluding remarks are given in Section 4.6.

4.1.2 Data on the Growth and Development of *Lepidopsetta polyxystra*

We consider modeling the development of larvae from a laboratory cohort of Northern rock sole (*Lepidopsetta polyxystra*) where the study took place in the Hatfield Marine Science Center in Newport, Oregon (Laurel et al., 2014). Larvae after hatching were distributed evenly across 15 tanks to monitor their development under four different temperatures. The larvae pass through distinct stages as depicted in Figure 1. Larval development was scored by the degree of observed tail flexion using the criteria established by Hawkyard et al. (2014). According to Hawkyard et al. (2014), stage 1 of the larvae development is characterized by a straight notochord (no flexion); while stage 2 larvae have straight notochord with the appearance of caudal peduncle 'node' near the posterior end; larvae in stage 3 have a bent notochord with caudal peduncle formation near the posterior end; stage 4 larvae have bent notochord and initial envelopment of the notochord by the caudal peduncle; and stage 5 is characterized by the full envelopment of the notochord by the caudal peduncle with only a remnant of bent notochord still visible (Laurel et al., 2014; Hawkyard et al., 2014).

The tanks were maintained at either 2, 4, 7, or 10°C with a view to studying the effect of temperature on the rate of transitions through the developmental stages. Two observation schedules were employed: for tanks 1-7 the developmental stages were to be assessed on days 24, 28, 46, 66 and 90 days, while for tanks 8-15 the classifications were to occur on days 10, 18, 30 and 47. In many tanks assessments did not occur on the scheduled days leading to incomplete data. Figure 4.2 gives a graphical plot of the data from tank 13.

In general , accurate identification of individual larvae is difficult, as is often the case in developmental studies of small mobile organisms. The process of identification of individual larvae can take time, effort and cost and may still be subject to error. Laurel et al. (2014) sampled the 10 fishes from 15 tanks due to identifiability problems. This is yet another motivation for our work, because aggregating data does not require identifying the fishes on an individual level. Here we treat them as the same ones and proceed with the analysis in the paper. We therefore develop methods that do not require tracking of individual organisms but rather relies on the frequency counts for the different stages at each assessment time.

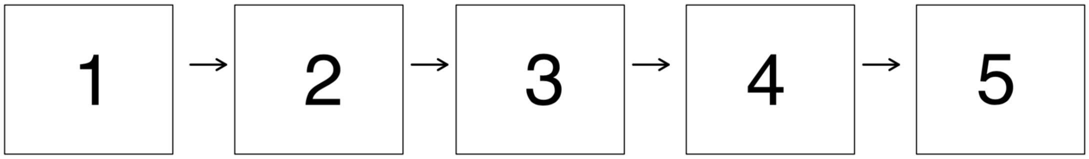


Figure 4.1: A 5-state progressive model for the development of Northern rock sole; state 1 is the stage with a straight notochord, state 2 corresponds to the development of a caudal peduncle node, state 3 involves a bent notochord, state 4 involves a bent notochord plus envelopment and state 5 involves full envelopment.

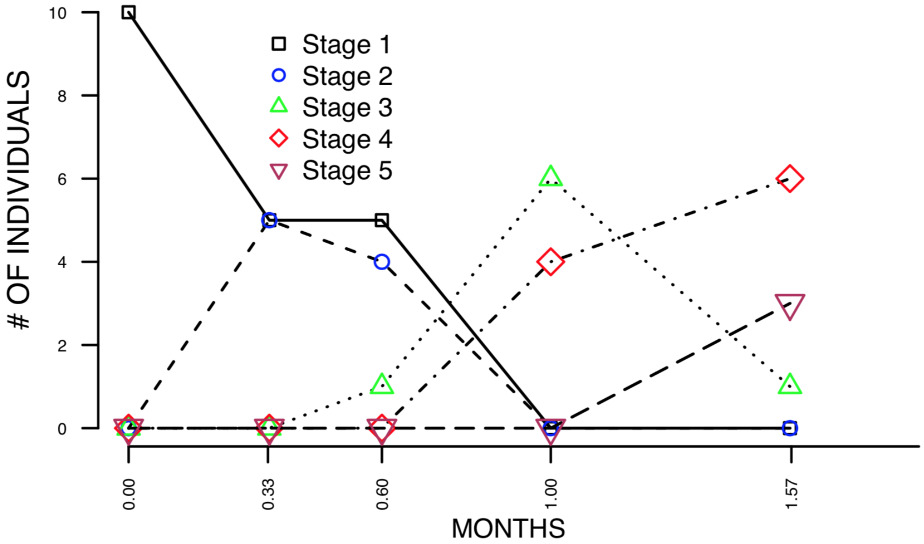


Figure 4.2: Plot of the frequency of the different stages over time for tank 13 containing 10 fish examined under schedule 1.

4.2 Aggregate Data with Independent Units

4.2.1 Notation and Model Formulation

We consider a strictly progressive model in this paper. Suppose that observations are made on a group of individuals who act independently of one another, with each individual passing through states according to a multistate process with state space $\{1, 2, ..., K\}$. We let $Z_i(t)$ denote the state occupied by ith individual at time t and $\{Z_i(s), 0 < s\}$ the multistate process. Let $\mathscr{H}_i(t) = \{Z_i(s), 0 \leq s < t\}$ denote the history of the process for individual i at time t and let

$$\lambda_k(t|\mathscr{H}_i(t)) = \lim_{\triangle t \downarrow 0} \frac{P(Z_i(t + \triangle t^-) = k + 1 | Z_i(t^-) = k, \mathscr{H}_i(t))}{\triangle t} \qquad (4.1)$$

denote the $k \to k+1$ transition intensity, $k = 1, \ldots, K-1$. For Markov processes the intensity in (5.1) does not depend on the history $\mathscr{H}_i(t)$ in which case we write it as $\lambda_k(t)$. Given a $K \times K$ transition intensity matrix $\Lambda(t)$ with $(k, k+1)$ entry $\lambda_k(t)$, diagonal entries $-\lambda_k(t)$, $k = 1, \ldots, K-1$, and zeros elsewhere, the $K \times K$ transition probability matrix $P(s,t)$ is obtained by product integration as

$$P(s,t) = \prod_{(s,t]} \{1 + \Lambda(u)du\}$$

with (k,l) entry $P(Z_i(t) = l | Z_i(s) = k)$ (Cook and Lawless, 2018).

Suppose that observations for individual i are made at times $0 = a_{i0} < a_{i1} < ... < a_{iR_i}$ yielding panel data $\{(Z_i(a_{ir}), a_{ir}), r = 0, 1, 2, \ldots, R_i\}$ for individual i. If $\lambda_k(t) = \lambda_k$, $k = 1, \ldots, K-1$, are time homogeneous then we can write $P(Z_i(a_{ir}) = l | Z_i(a_{i,r-1}) = k) = p_{kl}(a_{i,r-1}, a_{ir})$ with a particular form as

$$p_{kl}(a_{i,r-1}, a_{ir}) = \begin{cases} \displaystyle\sum_{h=k}^{l} B(k,h,l) e^{-\lambda_h \triangle a_{ir}}, & k \leq l \\ 0 & k > l \end{cases} \tag{4.2}$$

where $\triangle a_{ir} = a_{ir} - a_{i,r-1}$, $B(k,h,l)$ is given by,

$$B(k,h,l) = \prod_{u=k}^{l-1} \lambda_u / \prod_{\substack{u=k \\ u \neq h}}^{l} (\lambda_u - \lambda_h), \qquad k \leq h \leq l, \tag{4.3}$$

and $B(h,h,h) = 1$ provided $\lambda_k \neq \lambda_l$ for all $k \neq l = 1, ..., K-1$ (Satten, 1999). This can lead to simplifications of the likelihood which can be useful in certain settings as we discuss in Sections 4.3 and 4.4.

4.2.2 Likelihood Construction with Data Aggregated over a Single Tank

We now consider the setting in which observations are made at a common set of times $0 = a_0 < a_1 < ... < a_R$ for all organisms in a single tank where data are aggregated within the tank. We let $N_{kl}(\triangle a_r) = \sum_{i \in \mathscr{S}} I(Z_i(a_{r-1}) = k, Z_i(a_r) = l)$ for $k \leq l$ where $\mathscr{S}$ is the set of labels for individuals in the tank. Note that $N_{kl}(a_r)$ is unobserved in the aggregate data setting where we only know the total number of individuals in state l at each time a_r denoted by $M_l(a_r) = \sum_{k=1}^{l} N_{kl}(\triangle a_r)$, $l = 1, ..., K$, $r = 1, ..., R$; we also let $M(a_r) = (M_1(a_r), ..., M_K(a_r))'$ denote the vector of frequencies for the different stages of development (i.e., states).

Let $\mathscr{H}^0(a_r) = \{M(a_s), s = 1, \ldots, r-1\}$ denote history of observed marginal frequencies. Under the Markov property, to construct the likelihood with aggregate data the states occupied at time a_r depend only on the occupancy at a_{r-1} (i.e., $M(a_{r-1})$). We therefore only need consider consecutive assessment times, and the joint distribution is built up as a product of the conditional probabilities. Table 4.1 displays the data for a progressive K state Markov process for assessment times a_{r-1} and a_r, with the missing information (the transition counts) represented by the entries inside the table. Take row k for example, noting that $\sum_{l=k}^{K} N_{kl}(\triangle a_r) = M_k(a_{r-1})$ corresponds to the number of individuals occupying state k at time a_{r-1}; the lth column sum then corresponds to the number of individuals occupying state l at time a_r. If $N_k(\triangle a_r) = (N_{kk}(\triangle a_r), ..., N_{kK}(\triangle a_r))'$ denotes the potential non-zero elements in the kth row, the distribution of these latent counts is multinomial with

$$N_k(\triangle a_r) | M_k(a_{r-1}) \sim \text{Multinom}(M_k(a_{r-1}); p_{kk}(\triangle a_r), ..., p_{kK}(\triangle a_r))$$

for $k = 1, .., K-1$; note $P(N_{KK}(\triangle a_r) = M_K(a_{r-1}) | M_K(a_{r-1})) = 1$ since K is an absorbing state. Let $N(\triangle a_r) = (N'_1(\triangle a_r), ..., N'_K(\triangle a_r))'$ denote the full vector of latent counts in Table 4.1 and $M(a_{r-1}) = (M_1(a_{r-1}), ..., M_K(a_{r-1}))'$ denote the marginal counts at a_{r-1}. Then if we let $\theta_k = \log \lambda_k$ and $\theta = (\theta_1, ..., \theta_{K-1})'$, the observed data likelihood can be constructed as

$$L(\theta) \propto \prod_{r=1}^{R} \sum_{N(\triangle a_r) \in \mathscr{N}_r} P(N(\triangle a_r)|\mathscr{H}^0(a_r), M(a_{r-1}); \theta) = \prod_{r=1}^{R} \prod_{k=1}^{K} P(M_k(a_r)|M_k(a_{r-1}); \theta) \tag{4.4}$$

where $\mathscr{N}_r = \{N(\triangle a_r) : N_{k.}(\triangle a_r) = M_k(a_{r-1}), N_{.l}(\triangle a_r) = M_l(a_r), \forall\ (k,l)\}$ is the set of latent transition counts that are compatible with the margins of the table. In particular, $N_{k.}(\triangle a_r) = \sum_{l=k}^{K} N_{kl}(\triangle a_r)$ and $N_{.l}(\triangle a_r) = \sum_{k=k}^{K} N_{kl}(\triangle a_r)$.

$N_{11}(\triangle a_r)$	$N_{12}(\triangle a_r)$	...	$N_{1K}(\triangle a_r)$	$M_1(a_{r-1})$
0	$N_{22}(\triangle a_r)$	...	$N_{2K}(\triangle a_r)$	$M_2(a_{r-1})$
...	...	...		...
0	...	...		$M_{K-1}(a_{r-1})$
0	0	0	$N_{KK}(\triangle a_r)$	$M_K(a_{r-1})$
$M_1(a_r)$	$M_2(a_r)$	...	$M_K(a_r)$	

Table 4.1: Complete data on transitions and marginal counts over (a_{r-1}, a_r).

4.3 Marginal Models for Clustered Aggregate Data

4.3.1 Composite Likelihood for a Marginal Model

Consider J tanks (clusters) $j = 1, ..., J$. We assume covariates are only at the cluster level, and denote the vector by x_j for cluster j, $j = 1, ..., J$. We let n_j denote the number of individuals per cluster. Diao and Cook (2014) describe how models can be formulated for correlated Markov processes which accommodate dependence between processes and retain the marginal Markov property for processes. For progressive processes, the dependence is accommodated by selecting a sojourn or entry time of interest and using a copula function to induce a dependence between the corresponding times within a cluster. We consider here the class of Archimedian copulas of the form $C(u_1, u_2, \ldots, u_{n_j}; \eta) = G^{-1}(G(u_1; \eta) + \cdots + G(u_{n_j}; \eta), \eta)$, where $G : [0, 1] \to [0, \infty)$ is a continuous, strictly decreasing and convex generator function with dependence parameter η and $G(1; \eta) = 0$. We consider a progressive process and select the first transition time (i.e. the entry time to state 2) as the time on which the dependence is based. Let T_{ij2} denote the entry time to state 2 for individual i in tank j and $T_{j2} = (T_{1j2}, \ldots, T_{n_j j2})'$ denote the vector of all state 2 entry times in tank j, $j = 1, \ldots, J$. We use the Clayton copula (Nelsen, 2006) in this setting with generator $G(u; \eta) = \eta^{-1}(u^{-\eta} - 1)$ to model the dependence. Kendall's τ, a common dependence measure for copula models in the Archimedian family is obtained as

$$\tau = 1 + 4 \int_0^1 \frac{G(u; \eta)}{G'(u; \eta)} du.$$

The joint survivor function for T_{j2} is obtained via the probability integral transform and linking all marginal survivor functions $\mathscr{F}_{ij}(t_{ij2}) = \exp(-\lambda_1 t_{ij2})$ via the Clayton copula as

$$\mathscr{F}(t_{ij}; \theta_1, \eta) = \left(\mathscr{F}(t_{1j2}; \theta_1)^{-\eta} + \cdots + \mathscr{F}(t_{n_j j2}; \theta_1)^{-\eta} - (n_j - 1)\right)^{-1/\eta}.$$

Since the process is progressive the association in the entry times to state 2 will induce an association in the entry times to subsequent states within clusters. An alternative approach would be to remodel the association in the absorption times as considered in Diao and Cook (2014).

Here we adopt a composite likelihood by adopting a working independence assumption and considering contribution from marginal frequency data observed at each time point as arising independently from the data at different time points for the same cluster. This gives

$$CL_{jr}(\theta) \propto P(M_j(a_{jr})|M_{j1}(a_{j0}) = n_j, x_j; \theta) \tag{4.5}$$

where we let $\lambda_{jk} = \lambda_k \exp(x_j'\beta)$ and let $\log \lambda_{jk}$ be indexed by θ_{jk}, $j = 1, ..., J$. The component likelihood for each cluster j is given by (5.7) and a overall composite likelihood is simply the product of the component likelihoods,

$$CL(\theta) = \prod_{j=1}^{J} \prod_{r=1}^{R_j} CL_{jr}(\theta)$$

if we assume independence within clusters. Robust sandwich variance estimates are used to (see later) ensure correct inferences. The estimating equations corresponding to the composite likelihood is

$$S(\theta) = \sum_{j=1}^{J} \sum_{r=1}^{R_j} S_{jr}(\theta)$$

where $S_{jr}(\theta) = \partial \log CL_{jr}(\theta)/\partial\theta$. Since the contributions $CL_{jr}(\theta)$ in (5.7) are valid likelihood contributions, then

$$E(S(\theta)) = 0.$$

Under standard regularity conditions (White, 1982), we can then construct the robust sandwich variance

$$\sqrt{J}(\widehat{\theta} - \theta) \to N(0, A^{-1}(\theta)B(\theta)\left[A^{-1}(\theta)\right]') \tag{4.6}$$

where $A(\theta) = -E(\partial \sum_{r=1}^{R_j} S_{jr}(\theta)/\partial\theta')$ and $B(\theta) = E(S_j(\theta)S_j'(\theta))$. Here $S_j(\theta) = (S_{j1}(\theta), \ldots, S_{jR_j}(\theta))'$. The matrices $A(\theta)$ and $B(\theta)$ can be estimated empirically by

$$\widehat{A}(\theta) = -J^{-1} \sum_{j=1}^{J} \frac{\partial \sum_{r=1}^{R_j} S_{jr}(\theta)}{\partial\theta'}|_{\theta=\widehat{\theta}}$$

and

$$\widehat{B}(\theta) = J^{-1} \sum_{j=1}^{J} S_j(\theta) S_j'(\theta)|_{\theta=\widehat{\theta}}.$$

4.3.2 A Simulation Study

Here we consider a strictly progressive process with all individuals starting at state 1. We set λ_{12} such that $P(Z_i(1) = 1|Z_i(0) = 1) = 0.135$, We set $\lambda_{23} = 1.1\lambda_{12}$, $\lambda_{23} = 1.1^2\lambda_{12}$ and $\lambda_{45} = 1.1^3\lambda_{12}$. The data is generated such that the failure time to state 2 within each cluster is correlated under a copula model, we then generate sojourn times in state 2 by an exponential distribution with intensity λ_{23}. The entrance time to state 3 is then the sum of the failure time of state 2 and the sojourn time. We adopt the Clayton's copula with Kendall's τ of 0.1 and 0.2. To assess the effect of cluster size, we have used 25, 50 and 100 clusters with 10 and 30 individuals per cluster in the simulation for a total of 500 simulations under 4 common assessment times. Results are displayed in Table 4.2. From the result, we see that the empirical biases (EBIAS) are all close to zero, the empirical standard errors (ESE) agree well with the average robust standard errors (ASE), and the empirical coverage (ECP) is well within the nominal level. Note that the standard error decreases as the number of clusters increase, as well as when the number of organisms per per cluster (n_j/cluster) increases for a given correlation. We also observe the anticipated increase in variation with increasing association as reflected by Kendall's τ.

4.4 Accommodating Heterogeneity via Random Effects

Another approach for accommodating within-cluster variation dependence is to model between-cluster variation. We consider this here through the use of random effect models.

Let U_j be a scalar cluster-level random effect for cluster j with $E(U_j) = 1$, $\text{var}(U_j) = \phi$, and distribution function $G(U_j; \phi)$. We assume $U_j \perp\!\!\!\perp U_{j'}$ for $j \neq j' = 1, ..., J$. The time-homogeneous transition intensities given U_j for an individual in cluster j is then

$$\lim_{\triangle t \downarrow 0} \frac{P(Z_{ij}(t + \triangle t^-) = k + 1 | Z_{ij}(t^-) = k, U_j = u_j, x_j, \mathscr{H}_{ij}(t))}{\triangle t} = u_j \lambda_{jk}$$

where $\lambda_{jk} = \lambda_k \exp(x_j' \beta)$.

The marginal likelihood is obtained by integrating the joint likelihood for the aggregate data over the random effect. Maximum likelihood estimates are obtained by maximizing with respect to both the parameters λ in the conditional Markov model and the parameter ϕ in the random effect distribution.

For our strictly progressive process (see Section 4.2) we can make use of (4.2) to accommodate the cluster level random effect and covariate and write

$$P(Z(a_{ijr}) = s_{ijr} | Z(a_{ij,r-1}) = s_{ij,r-1}; u_j, x_j) = \sum_{h = s_{ij,r-1}}^{s_{ijr}} B(s_{ij,r-1}, h, s_{ijr}) \exp(-u_j \lambda_h e^{x_j' \beta} \triangle a_{jr})$$

where $\triangle a_{jr} = a_{jr} - a_{j,r-1}$ is the lag between the $(j-1)$st and jth assessment times for tank j and s_{ijr} represent the state occupied for the ith individual at time a_{jr} in cluster j. Note $B(\cdot, \cdot, \cdot)$ is defined as in (4.3).

With individual level panel data, the likelihood contribution given u_j in cluster j can be written as

$$\mathscr{L}_j(\lambda, \beta | u_j, x_j) \propto \prod_{i=1}^{n_j} \prod_{r=1}^{R_j} P(Z(a_{ijr}) = s_{ijr} | Z(a_{ij,r-1}) = s_{ij,r-1}; u_j, x_j). \tag{4.7}$$

Based on the form of (4.2) we can rewrite (4.7) as

$$\mathscr{L}_j(\lambda, \beta | u_j, x_j) \propto \prod_{i=1}^{n_j} \left[\sum_{h_0=s_{ij0}}^{s_{ij1}} \sum_{h_1=s_{ij1}}^{s_{ij2}} \cdots \sum_{h_{R_j-1}=s_{ij,R_j-1}}^{s_{ijR_j}} \left\{ \prod_{r=1}^{R_j} B(s_{ij,r-1}, h_{r-1}, s_{ijr}) e^{-u_h \lambda_{jh_{r-1}} \triangle a_{jr}} \right\} \right].$$

We can then get the marginal distribution by averaging over the random effect

$$\mathscr{L}_j(\lambda, \beta, \phi) \propto \int_0^\infty \mathscr{L}_j(\lambda, \beta | u_j, x_j) dG(u_j; \phi).$$

A closed-form of marginal likelihood is obtainable if there exists a Laplace transform $v_\phi(\cdot)$ for the random effect distribution.

When only aggregate data are available at the cluster level, we must again marginalize over the complete tables as in (5.7). This summation is infeasible because the number of possible realizations of individual paths increases at a prohibitive rate with the number of assessment times and the cluster size n_j, even for progressive models. We therefore consider an alternative approach based on a composite likelihood. Specifically we consider a two-way contribution over $[0, a_{jr}]$ given u_j in cluster j.

4.4.1 Two-way Composite Likelihood

Here we consider composite likelihood contributions based on data at times a_0 and a_{jr} for $r = 1, \ldots, R_j$. For a particular r in cluster j we obtain

$$\mathscr{L}_{jr}(\lambda, \beta | u_j, x_j) \propto \prod_{i=1}^{n_j} P(Z(a_{jr}) = s_{ijr} | Z(a_{j0}) = s_{ij0}; u_j, x_j) \tag{4.8}$$

where

$$\mathscr{L}_{jr}(\lambda, \beta | u_j, x_j) \propto \prod_{i=1}^{n_j} \sum_{h_i=s_{ij0}}^{s_{ijr}} B(s_{ij0}, h_i, s_{ir}) e^{-u_j \lambda_{jh_i} a_{jr}} . \tag{4.9}$$

Integrating (4.8) over the random effect and taking the product of all such terms for $r = 1, \ldots, R_j$ gives a joint probability and composite likelihood for panel data setting which can be written as

$$\mathscr{L}_j(\lambda, \beta, \phi) \propto \prod_{r=1}^{R_j} \left\{ \int_0^\infty \mathscr{L}_{jr}(\lambda, \beta | u_j, x_j) dG(u_j; \phi) \right\} . \tag{4.10}$$

If $v_\phi(\cdot)$ is the Laplace transform of the random effect distribution, a closed-form for the integral is obtained by replacing each exponential factor by the Laplace transform to obtain

$$\mathscr{L}_j(\lambda, \beta, \phi) \propto \prod_{r=1}^{R_j} \left\{ \sum_{h_1 = s_{1j0}}^{s_{1jr}} \sum_{h_2 = s_{2j0}}^{s_{2jr}} \cdots \sum_{h_{n_j} = s_{n_j, j0}}^{s_{n_j jr}} \prod_{i=1}^{n_j} B(s_{ij0}, h_i, s_{ijr}) v_\phi \left(\sum_{i=1}^{n_j} \lambda_{jh_i} a_{jr} \right) \right\} . \tag{4.11}$$

When data are aggregated and only marginal totals are available at each assessment time, the observed data composite likelihood is obtained by replacing each term in curly brackets in (4.11) with its sum over all possible $K \times K$ matrices of transition counts between a_0 and a_r to give

$$L_j(\lambda, \beta, \phi) \propto \prod_{r=1}^{R_j} \sum_{N(a_{jr}) \in \mathscr{N}_{jr}} \mathscr{L}_{jr}(\lambda, \beta, \phi) \tag{4.12}$$

for each cluster j where $N_{j1l}(a_r) = \sum_{i=1}^{n_j} I(Z_{ij}(a_{jr}) = l | Z_{ij}(a_{j0}) = 1)$ and $\mathscr{N}_{jr} = \{N_j(a_{jr}) : N_{j1.}(a_{jr}) = n_j, N_{j.l}(a_{jr}) = M_{jl}(a_{jr}), \forall\ l\}$. The overall composite likelihood is obtained by multiplying composite likelihood contributions of the form (4.12) over all J clusters.

4.4.2 A Simulation Study for Random Effect Model

Here we consider a strictly progressive process with all individuals starting at state 1. We set λ_{12} such that $P(Z_i(1) = 1 | Z_i(0) = 1) = 0.135$, We set $\lambda_{23} = 1.1\lambda_{12}$, $\lambda_{23} = 1.1^2\lambda_{12}$ and $\lambda_{45} = 1.1^3\lambda_{12}$. The random effect $U_j \sim \text{Gamma}(\text{mean} = 1, \text{var} = \phi)$ where $\phi = 0.4$ and 0.8. To assess the effect of cluster size, we have used 25, 50 and 100 clusters with 10 and

30 individuals per cluster in the simulation for a total of 500 simulations under 4 common assessment times. Results are displayed in Table 4.3. We see that the ESE and ASE are in alignment and the empirical coverage probabilities are well within the nominal level. Note that the finite sample empirical biases of the parameters under the random effect model decreases as the number of clusters increases. As the number of clusters increase and/or number per cluster increase, we see the resulting decrease in the standard errors.

4.5 Application

Lepidopsetta polyxystra is also known as the northern rock sole which are widely studied for understanding larval transport and associated nursery grounds in the Bering Sea (Laurel et al., 2014). As part of a study on the role of temperature on the development of the organism, fish larvae were placed in 15 different tanks under four constant temperatures of 2, 4, 7 or 10°C respectively. The fish larvae metamorphosis through egg, first, second, third stage of maturation, and finally enter an adult fourth stage. At each temperature, tanks were supplied with nutritionally enriched rotifers (*Branchionus plicatilis*), after which the number of fish larvae were recorded along with their development stages. Two distinct unevenly spaced assessment schedules were employed; see Section 4.1.2. We consider data from all 15 tanks and conduct an analysis of the aggregate counts recorded at the unevenly spaced assessment times based on both a robust marginal and a two-way random effect model.

We present results from fitting models without covariates, and with temperature as a continuous covariate and as a factor variable. For the model without covariates the parametric estimates of the state entry time distributions are then superimposed on the nonparametric composite likelihood estimates obtained by the pooled-adjacent violators algorithm (PAVA) (Ayer et al., 1955) to assess the goodness of fit of our proposed models. This nonparametric estimate was obtained by constructing a dataset of "pseudo-individuals" from each tank by treating each assessment time as corresponding to a different individual. Let $C_{(j1)} < C_{(j2)} < ... < C_{(jR)}$ denote R ordered assessment times for tank j. Let f_{jr} denote the number of pseudo-individuals in tank j who have experienced the event of interest at time $C_{(jr)}$ and those who had not failed by C_{jr} as $n_{jr} - f_{jr}$. An isotonic regression of $(f_{j1}/n_{j1}, ..., f_{jR}/n_{jR})$ with weights $(n_{j1}, ..., n_{jR})$ gives

$$1 - \hat{\mathscr{F}}(C_{(jr)}) = \max_{u \leq r} \min_{v \geq r} \left(\frac{\sum_{h=u}^{v} f_{jl}}{\sum_{h=u}^{v} n_{jr}} \right). \tag{4.13}$$

Under a working independence assumption (across time and across tanks) we pool all such data to obtain the nonparametric estimate of the state entry time distribution.. Figure 4.3 shows the nonparametric estimates superimposed with the marginal composite likelihood estimates on the left column and the marginal probabilities based on the random effect model on the right column. The dashed line is the lower and uppoer limits of the 95% confidence intervals. From Figure 4.3 we see that both methods agree well with the nonparametric estimates. For the random effect approach, we see that there is (as yet) unexplained tank-to-tank variation with $\widehat{\phi} = 0.28$ (see Table 4.4). It is natural to question what the source of this biological variation could be. To explore this we next fit regression models using the only covariate available, which is the temperature of the tank.

In order to assess the effect of different temperatures in tanks, we have also fit a model with temperature as a covariate (see Table 4.4). The coefficient of temperature for the marginal model when it is treated as having a linear effect gives a relative increase in the transition intensity of 1.29 (95% CI 1.23, 1.35) for each degree increase in temperature. The p-value of the test of the null hypothesis is p<0.001 suggesting that progression rates through the developmental stages are affected by the temperature of the environment.

When temperature is treated as a categorical covariate with 4 level, we select 2°C as the baseline temperature such that β_1, β_2 and β_3 represent the effects of 4°C, 7°C and 10°C respectively. It is also possible to compute the effects of the temperature change from 2°C to 4°C, 2°C to 7°C and 2°C to 10°C by computing $\exp(\log(1.29) * (4 - 2)) = 1.66$, $\exp(\log(1.29) * (7 - 2)) = 3.57$, and $\exp(\log(1.29) * (10 - 2)) = 7.67$ which are in broad agreement with the estimates from the model treating temperature as a continuous covariate.

The estimates from the random effects model are similar to the marginal approach when we add in the temperature effect. When adjusting for temperature the estimate of ϕ from the random effects model is zero for both categorical and linear covariates, which reflects the fact that temperature explains all of the residual tank-to-tank variation from the null model.

4.6 Discussion

We have described a composite likelihood-based method for the analysis of clustered aggregate developmental data. The computational feasibility of this approach hinges on the progressive nature of the process which is characteristic of most growth cycles, and the fact that all organisms were observed from the start of the first stage. Use of composite likelihood greatly reduces the size of the sample space that must be marginalized over to compute the probabilities based on the marginal frequencies.

Marginal models and random effect models were used to accommodate clustering of rates within tanks. Estimation of the parameters under the marginal formulation did not involve estimation of the dependence parameter of the copula as this was more of a nuisance parameter in the present setting. Moreover the apparent need to accommodate heterogeneity between clusters seems minimal once the temperature effect was accounted for. The plots of the state entry time distributions based on the available data exhibited good agreement with the nonparametric estimates using the pooled adjacent violators algorithm and so models with exponential sojourn times appear reasonable for the data at hand.

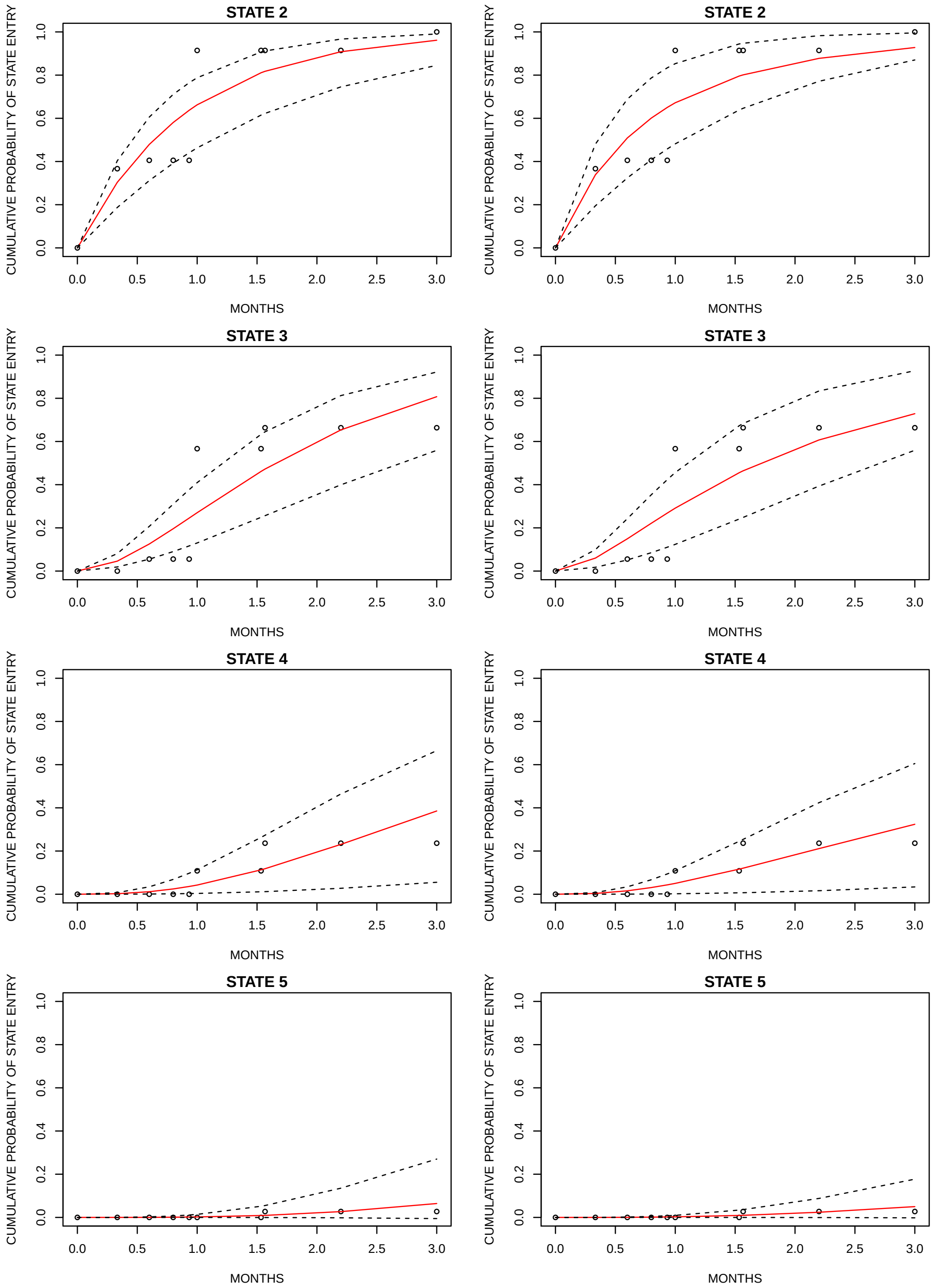


Figure 4.3: Nonparametric estimates(represented by dots) vs. parametric estimates from the marginal model on the left column and the random effect model via composite likelihood on the right column; the dotted line repreents a 95% confidence interval for the parametric fits.

		$\tau = 0.1$										$\tau = 0.2$									
		N/cluster=10					N/cluster=30					N/cluster=10					N/cluster=30				
Clusters	$\log\theta$	EBIAS	ESE	ASE	ASE[1]	ECP	EBIAS	ESE	ASE	ASE[1]	ECP	EBIAS	ESE	ASE	ASE[1]	ECP	EBIAS	ESE	ASE	ASE[1]	ECP
25	$\log\lambda_{12}$	<0.001	0.095	0.090	0.042	93.6 %	0.002	0.071	0.073	0.024	95.4 %	<0.001	0.114	0.110	0.042	93.8 %	-0.006	0.100	0.097	0.024	94.0 %
	$\log\lambda_{23}$	-0.005	0.095	0.094	0.063	93.8 %	0.002	0.057	0.054	0.037	93.4 %	-0.002	0.099	0.097	0.063	93.6 %	<0.001	0.065	0.069	0.037	96.6 %
	$\log\lambda_{34}$	¡0.001	0.095	0.090	0.042	93.6 %	-0.005	0.069	0.067	0.052	93.0 %	-0.010	0.121	0.118	0.091	93.0 %	-0.010	0.072	0.070	0.052	93.4 %
	$\log\lambda_{45}$	-0.011	0.116	0.116	0.091	93.9 %	-0.005	0.069	0.067	0.052	93.0 %	-0.010	0.121	0.118	0.091	93.0 %	-0.010	0.072	0.070	0.052	93.4 %
50	$\log\lambda_{12}$	-0.005	0.070	0.065	0.030	93.0 %	0.002	0.051	0.052	0.017	95.8 %	-0.006	0.079	0.078	0.030	94.8 %	-0.003	0.075	0.073	0.024	93.8 %
	$\log\lambda_{23}$	-0.003	0.064	0.064	0.045	94.8 %	¡0.001	0.041	0.039	0.026	93.6 %	-0.004	0.065	0.065	0.045	95.2 %	¡0.001	0.044	0.043	0.036	94.2 %
	$\log\lambda_{34}$	-0.003	0.088	0.083	0.065	93.0 %	-0.003	0.048	0.048	0.037	95.4 %	-0.011	0.084	0.084	0.065	94.6 %	-0.004	0.051	0.051	0.037	95.6 %
	$\log\lambda_{45}$	0.004	0.123	0.118	0.096	94.2 %	0.003	0.065	0.067	0.055	96.4 %	-0.001	0.116	0.115	0.096	94.2 %	-0.002	0.069	0.067	0.055	94.4 %
100	$\log\lambda_{12}$	-0.002	0.048	0.046	0.021	93.0 %	-0.001	0.037	0.037	0.012	94.0 %	-0.006	0.057	0.055	0.021	93.6 %	-0.001	0.051	0.049	0.012	94.4 %
	$\log\lambda_{23}$	-0.002	0.044	0.045	0.045	95.2 %	¡0.001	0.028	0.028	0.018	94.4 %	-0.002	0.047	0.047	0.031	95.2 %	-0.001	0.031	0.030	0.018	93.8 %
	$\log\lambda_{34}$	¡0.001	0.059	0.059	0.045	95.0 %	-0.002	0.034	0.035	0.026	96.0 %	¡0.001	0.059	0.059	0.045	94.6 %	-0.003	0.035	0.036	0.026	95.2 %
	$\log\lambda_{45}$	-0.002	0.082	0.081	0.067	94.4 %	0.001	0.047	0.048	0.039	95.2 %	-0.005	0.084	0.082	0.067	93.0 %	-0.001	0.046	0.048	0.039	95.2 %

[1] naive ASE

Table 4.2: Empirical performance of estimators for 500 simulations under a composite likelihood via marginal model.

		$\phi = 0.4$								$\phi = 0.8$							
		N/cluster=10				N/cluster=30				N/cluster=10				N/cluster=30			
Clusters	$\log \theta$	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
25	$\log \lambda_{12}$	0.004	0.167	0.166	93.3 %	-0.008	0.145	0.141	93.4 %	-0.014	0.215	0.210	92.7 %	-0.026	0.216	0.210	93.4 %
	$\log \lambda_{23}$	0.019	0.160	0.167	93.5 %	-0.007	0.149	0.146	93.7 %	0.012	0.231	0.218	93.5 %	-0.008	0.207	0.207	93.6 %
	$\log \lambda_{34}$	0.021	0.187	0.189	93.9 %	0.005	0.154	0.157	94.2 %	0.020	0.245	0.237	93.2 %	0.002	0.206	0.228	94.0 %
	$\log \lambda_{45}$	0.035	0.227	0.226	93.6 %	0.004	0.163	0.164	94.6 %	0.027	0.257	0.268	94.1 %	-0.009	0.227	0.230	95.2 %
	$\log \phi$	0.225	0.636	0.639	96.3 %	0.110	0.428	0.420	95.4 %	0.139	0.483	0.452	95.4 %	0.130	0.376	0.384	94.4 %
50	$\log \lambda_{12}$	-0.009	0.157	0.156	93.6 %	-0.008	0.145	0.141	93.4 %	-0.009	0.157	0.156	93.6 %	-0.013	0.147	0.143	93.4 %
	$\log \lambda_{23}$	-0.001	0.112	0.112	95.8 %	-0.003	0.103	0.101	93.2 %	<0.001	0.163	0.159	94.6 %	-0.008	0.153	0.152	93.0 %
	$\log \lambda_{34}$	0.009	0.131	0.135	95.0 %	-0.003	0.115	0.109	93.2 %	0.005	0.169	0.172	93.6 %	0.002	0.148	0.145	94.2 %
	$\log \lambda_{45}$	0.022	0.152	0.159	94.2 %	0.007	0.116	0.116	94.4 %	0.008	0.173	0.189	94.8 %	0.002	0.154	0.149	93.6 %
	$\log \phi$	0.095	0.398	0.408	96.6 %	0.046	0.283	0.275	94.2 %	0.081	0.289	0.305	96.2 %	0.059	0.237	0.242	95.0 %
100	$\log \lambda_{12}$	<0.001	0.086	0.086	94.2 %	<0.001	0.708	0.710	94.1 %	-0.009	0.105	0.127	94.8 %	-0.004	0.092	0.095	95.6 %
	$\log \lambda_{23}$	0.003	0.084	0.084	95.8 %	0.002	0.073	0.072	93.3 %	-0.003	0.114	0.116	95.4 %	0.002	0.095	0.101	96.8 %
	$\log \lambda_{34}$	-0.003	0.096	0.096	94.4 %	0.002	0.074	0.077	94.9 %	-0.003	0.114	0.128	95.6 %	0.005	0.095	0.103	95.6 %
	$\log \lambda_{45}$	<0.001	0.116	0.117	93.4 %	<0.001	0.080	0.084	94.5 %	0.007	0.132	0.140	94.4 %	0.005	0.097	0.105	96.4 %
	$\log \phi$	0.042	0.277	0.274	94.6 %	0.027	0.194	0.191	94.1 %	0.038	0.196	0.211	96.4 %	0.021	0.171	0.168	94.0 %

Table 4.3: Empirical performance of estimators for 500 simulations under a two-way composite likelihood with random effects model.

	Marginal Model						Random Effect Model					
	EST	95% CI	pvalue	EST	CI	pvalue	EST	CI	pvalue	EST	CI	pvalue
	No Covariate											
	Marginal						Two-way					
$\log \lambda_{12}$	0.08	(-0.35, 0.51)					0.28	(-0.43, 0.99)				
$\log \lambda_{23}$	-0.05	(-0.38, -0.28)					-0.09	(-0.46, 0.28)				
$\log \lambda_{34}$	-0.80	(-1.62, 0.02)					-1.07	(-1.91, -0.23)				
$\log \lambda_{45}$	-1.73	(-3.32, -0.14)					-2.27	(-3.96, -0.58)				
ϕ	–	–					0.28	(0.02, 3.89)				
	Regression											
	Categorical			Linear			Categorical			Linear		
$\log \lambda_{12}$	-0.80	(-1.08, -0.53)		-1.13	(-1.43, -0.82)		-0.80	(-1.10, -0.51)		-1.14	(-1.44, 0.83)	
$\log \lambda_{23}$	-1.20	(-1.55, -0.85)		-1.51	(-1.90, -1.11)		-1.20	(-1.56, -0.84)		-1.52	(-1.92, -1.12)	
$\log \lambda_{34}$	-2.21	(-2.82, -1.60)		-2.50	(-3.11, -1.89)		-2.21	(-2.70, -1.70)		-2.54	(-3.16, -1.91)	
$\log \lambda_{45}$	-3.88	(-5.66, -2.10)		-4.19	(-6.02, -2.36)		-3.88	(-5.84, -1.92)		-4.17	(-5.95, -2.40)	
ϕ	–	–		–	–		0.00	(0.00, 0.00)		0.01	(0.00, 0.02)	
$\exp(\beta_1)$	2.47	(1.88, 3.25)	<0.001	–	–		2.47	(1.88, 3.27)	<0.001	–	–	
$\exp(\beta_2)$	3.99	(2.59, 6.15)	<0.001	–	–		3.99	(2.58, 6.17)	<0.001	–	–	
$\exp(\beta_3)$	8.93	(6.18, 12.90)	<0.001	–	–		8.91	(6.86, 11.61)	<0.001	–	–	
$\exp(\beta)$	–	–		1.29	(1.23, 1.35)	<0.001	–	–		1.29	(1.23, 1.35)	<0.001

Table 4.4: Estimates from fitting the marginal models with robust variance estimation and the random effect models via two-way composite likelihood for the Northern rock sole.

Chapter 5

Cost-effective Design with Aggregation and Tracking

5.1 Introduction

In many growth and developmental studies organisms are arranged in tanks or other types of enclosure and repeatedly examined over time to acquire information on developmental stages. Examples include studies of plant growth (Gouno et al., 2011), metamorphosis of fish or amphibians (Laurel et al., 2014), or small arthropods (Munholland and Kalbfleisch, 1991). The maturation process can usually be naturally modelled using multistate processes.

In some contexts it can be difficult to identify individual organisms. In studies of hornworms for example (Borror and White, 1970), the larvae are both mobile and indistinguishable. Gouno et al. (2011) also reported on a growth study of *arabidopsis thaliana* where the data are recorded in aggregated form. In such cases the available data consists only of the counts of the number of organisms in the different developmental stages at each assessment time. This form of aggregation is also common when the only available data are published in tabular form.

There has been much discussion on methods for dealing with aggregate data. MacRae (1977) first introduced the nonlinear generalized least squares approach and briefly mentioned methods for exact maximum likelihood approach for aggregate data. Kalbfleish and Lawless (1983) introduced a weighted least squares approach for estimating transition intensities from aggregate data. We build upon a likelihood approach in this paper and consider strictly progressive Markov processes appropriate for growth data. Computational challenges may arise as the number of assessment times and individuals increase, so we propose composite likelihood (Varin et al., 2011) as an appealing alternative in such cases. When organisms are organized in different tanks (i.e. clusters), tank-to-tank variation must be taken into consideration. Chapter 4 use composite likelihood to handle such data based on both marginal methods with robust variance estimation, and a random effects model.

The focus of this paper is on the optimal design for studies involving multiple tanks/clusters; we adopt the marginal approach of Chapter 4 for aggregate data. In some contexts tracking of individuals is possible but incurs a cost. We also consider cost-effective design by addressing the situation in which some tanks contain organisms to be tracked individually over time, while other tanks may be designated to provide only aggregate counts in the different developmental stages at different assessment times. Sample size calculations are derived and cost-effective allocation of tanks to these two observation schemes is also considered.

The remainder of this contribution is organized as follows. In Section 5.2 we define notation and describe a composite likelihood for clustered Markov processes which we use to characterize growth of individual organisms and to accommodate dependence in progression rate within tanks. Large sample results and methods of inference for both tracking and aggregate observation schemes are given in Section 5.2. Sample size criteria are developed to meet design objectives and cost-effective allocation of tanks to the tracking and aggregate observation schemes are developed in Section 5.3. Simulation studies are also carried out in Section 5.4 along with an illustration involving the growth and development of Northern rock sole. Concluding remarks are made in Section 5.5.

5.2 Notation and Likelihood

5.2.1 Composite Likelihood for Clustered Panel Data

We consider strictly progressive multistate models suitable for studying maturation processes. Suppose that observations are made on a group of individuals who act independently of one another, with each individual passing through states according to a multistate process with state space $\{1, 2, \ldots, K\}$. We let $Z_j(t)$ denote the state occupied by individual j at time t and $\{Z_j(s), 0 < s\}$ be the multistate process.

Let $\mathscr{H}_j(t) = \{Z_j(s), 0 \leq s < t\}$ denote the history of the process for individual j at time t and let

$$\lambda_k(t|\mathscr{H}_j(t)) = \lim_{\triangle t \downarrow 0} \frac{P(Z_j(t + \triangle t^-) = k + 1 | Z_j(t^-) = k, \mathscr{H}_j(t))}{\triangle t} \tag{5.1}$$

denote the $k \to k+1$ transition intensity, $k = 1, \ldots, K-1$. For Markov processes the intensity does not depend on the history in which case we write (5.1) as $\lambda_k(t)$. Given a $K \times K$ transition intensity matrix $\Lambda(t)$ with $(k, k+1)$ entry $\lambda_k(t)$, diagonal entry $-\lambda_k(t)$, for $k = 1, \ldots, K-1$ and zeros elsewhere, by product integration (Cook and Lawless, 2018) the $K \times K$ transition probability matrix is

$$P(s,t) = \prod_{(s,t]} \{1 + \Lambda(u)du\} \tag{5.2}$$

with (k, l) entry $P(Z(t) = l|Z(s) = k)$ for $k \leq l$. If observations are made at times $0 = a_{j0} < a_{j1} < ... < a_{jR_j}$ for individual j, panel data denoted by $\{(Z_j(a_{jr}), a_{jr}), r = 1, 2, ..., R_j\}$ are obtained. Kalbfleisch and Lawless (1985) develop a Fisher-scoring algorithm for fitting the likelihood which is implemented in the R function *msm* (Jackson, 2011).

Now consider a setting with I tanks of organisms with n_i individuals in tank i, $i = 1, \ldots, I$. Let $0 = a_{i0} \leq \cdots \leq a_{iR_i}$ denote the common assessment times for all $j = 1, \ldots, J$ individuals in tank i, $i = 1, \ldots, I$. Diao and Cook (2014) formulate a copula-based model for correlated Markov processes which accommodate dependence between processes within clusters and retain the marginal Markov property for each process. With progressive processes, within-cluster dependence can be modeled in terms of sojourn or state entry times through copula functions. We consider a class of Archimedian copulas (Nelsen, 2006) of the form $C(u_1, u_2, \ldots, u_{n_i}; \eta) = G^{-1}(G(u_1; \eta) + \cdots + G(u_{n_i}; \eta))$ where $G : [0, 1] \to [0, \infty)$ is a continuous, strictly decreasing and convex generator function with dependence parameter η and $G(1; \eta) = 0$ (Nelsen, 2006). To induce a dependence, we select the first transition time (i.e. the entry time to state 2) and note that a dependence is induced within clusters for the subsequent state entry times. Specifically, we let T_{ij2} denote the entry time to state 2 for individual j in tank i and $T_{i2} = (T_{i12}, \ldots, T_{in_i2})'$ denote the vector of all state 2 entry times in tank i, $i = 1, \ldots, I$. We adopt the Clayton copula (Nelsen, 2006) and use Kendall's τ as a measure of dependence where

$$\tau = 1 + 4 \int_0^1 \frac{G(u; \eta)}{G'(u; \eta)} du.$$

We formulate the joint survivor function for T_{i2} by linking all marginal survivor functions $\mathscr{F}_{ij}(t_{ij2}; \lambda_1) = P(T_{ij2} \geq t_{ij2}) \exp(-\lambda_1 t_{ij2})$ via the Clayton copula as

$$\mathscr{F}(t_{i2}; \lambda_1, \eta) = \left(\mathscr{F}(t_{i12}; \lambda_1)^{-\eta} + \cdots + \mathscr{F}(t_{in_i2}; \lambda_1)^{-\eta} - (n_i - 1)\right)^{1/\eta}.$$

Diao and Cook (2014) describe an alternative approach where the association in the absorption times is modeled instead of earlier state entry or sojourn times, but the principle of inducing a dependence between multistate processes within a cluster by linking a particular time is in the same spirit.

Consider the case with a cluster level covariate x_i, $i = 1, \dots, I$ and let

$$\lim_{\triangle t \downarrow 0} \frac{P(Z_{ij}(t + \triangle t^-) = k+1 | Z_{ij}(t^-) = k, \mathscr{H}_j(t))}{\triangle t} = \lambda_k \exp(x_i' \beta)$$

which we denote more compactly as λ_{ik}, $k = 1, \dots, K-1$. If $\alpha_k = \log \lambda_k$, $k = 1, \dots, K-1$, $\alpha = (\alpha_1, \dots, \alpha_{K-1})'$ and $\beta = (\beta_1', \dots, \beta_{K-1}')'$, we then let $\theta = (\alpha', \beta')'$. Under a working independence assumption and a panel observation scheme (i.e. with individual tracking) the composite likelihood is

$$L_1(\theta) \propto \prod_{i=1}^{I} \prod_{r=1}^{R_i} L_{1ir}(\theta) \tag{5.3}$$

where

$$L_{1ir}(\theta) \propto \prod_{j=1}^{n_i} \prod_{k \le l} P(Z_{ij}(a_{ir}) = l | Z_{ij}(a_{i,r-1}) = k, x_i; \theta) \tag{5.4}$$

and $Z_{ij}(t)$ is the state occupied by individual j in tank i at time t. We then define

$$S_{1ir}(\theta) = \sum_{j=1}^{n_i} \sum_{k \le l} \frac{\partial \log P(Z_{ij}(a_{ir}) = l | Z_{ij}(a_{i,r-1}) = k, x_i; \theta)}{\partial \theta} \tag{5.5}$$

and we let $S_{1i}(\theta) = (S_{1i1}(\theta), \dots, S_{1iR_i}(\theta))$ be a $p \times R_i$ matrix. We let $\widehat{\theta}$ denote the solution to $S_1(\theta) = \sum_{i=1}^{I} \sum_{r=1}^{R_i} S_{1ir}(\theta) = 0$.

A robust sandwich variance estimate is required to ensure valid inference under this working independence assumption. Under standard regularity conditions (White, 1982)

$$\sqrt{I}(\widehat{\theta} - \theta) \to N(0, \mathscr{A}_1^{-1}(\theta) \mathscr{B}_1(\theta) \mathscr{A}_1^{-1}(\theta)) \tag{5.6}$$

where $\mathscr{A}_1(\theta) = -E\{\sum_{r=1}^{R_i} \partial S_{1ir}(\theta)\}$ and $\mathscr{B}_1(\theta) = E\{S_{1i}(\theta)S'_{1i}(\theta)\}$. The matrices $\mathscr{A}_1(\theta)$ and $\mathscr{B}_1(\theta)$ can be estimated empirically by

$$\widehat{A}_1 = -I^{-1}\sum_{i=1}^{I}\sum_{r=1}^{R_i}\frac{\partial S_{1ir}(\theta)}{\partial \theta'} \mid_{\theta=\widehat{\theta}}$$

and

$$\widehat{B}_1 = I^{-1}\sum_{i=1}^{I} S_{1i}(\theta)S'_{1i}(\theta) \mid_{\theta=\widehat{\theta}},$$

and tests regarding elements of θ or associated 95% confidence intervals are constructed based on the estimated covariance matrix $\widehat{A}_1^{-1}\widehat{B}_1\widehat{A}_1^{-1}$.

5.2.2 Composite Likelihood for Correlated Aggregate Data

Under the Markov property for a single individual process considered on its own, the stage occupied at time a_{ir} only depends on the stage occupied at $a_{i,r-1}$. With aggregate data we only need to consider two consecutive assessment times, and the joint distribution is built up as a product of the conditional probabilities. However, as the number of assessment times and individuals per tank increase, the likelihood becomes computationally challenging. That motivates use of a composite likelihood approach where we adopt a working independence assumption and consider contributions from the marginal frequency data observed at each time point as arising independently from the data at different time points from the same tank.

Here we consider data from the baseline assessment to each of the followup assessment times. Thus for two assessment times $a_{i0} = 0$ and a_{ir}, the missing information in the aggregate data are $N_i(a_{ir})$, the vector containing all counts $N_{i1l}(a_{ir}) = \sum_{j=1}^{n_i} I(Z_{ij}(a_{ir}) = l|Z_{ij}(a_{i0}) = 1)$ for $l = 1, ..., K$ and $i = 1, \ldots, I$. With a strictly progressive process and $P(Z_{ij}(a_{i0}) = 1) = 1$ and we let $N_{i1l}(a_{ir}) = M_{il}(a_{ir})$ corresponds to the number of individuals occupying state l at time a_{ir} in tank i. We can then obtain the composite likelihood

$$L_2(\theta) \propto \prod_{i=1}^{I} \prod_{r=1}^{R_i} L_{2ir}(\theta) \tag{5.7}$$

and

$$L_{2ir}(\theta) \propto P(M_i(a_{ir}) | M_{i1}(a_{i0}) = n_i, x_i; \theta)$$

where $M_i(a_{ir}) = (M_{i1}(a_{ir}), \ldots, M_{iK}(a_{ir}))'$.

Robust sandwich variance estimates are adopted to ensure valid inference. The estimating equations corresponding to the composite likelihood is

$$S_2(\theta) = \sum_{i=1}^{I} \sum_{r=1}^{R_i} S_{2ir}(\theta)$$

where $S_{2ir}(\theta) = \partial \log L_{2ir}(\theta)/\partial\theta$. Since the contributions of (5.7) are valid likelihood contributions $E\{S_2(\theta)\} = 0$ and the solution is denoted by $\widetilde{\theta}$. Again, under standard regularity conditions (White, 1982), we can then construct the robust sandwich variance as

$$\sqrt{I}(\widetilde{\theta} - \theta) \to N(0, \mathscr{A}_2^{-1}(\theta)\mathscr{B}_2(\theta)\mathscr{A}_2^{-1}(\theta)) \tag{5.8}$$

where $\mathscr{A}_2(\theta) = -E\{\sum_{r=1}^{R_i} \partial S_{2ir}(\theta)\}$ and $\mathscr{B}_2(\theta) = E\{S_{2i}(\theta)S_{2i}'(\theta)\}$ with $S_{2i}(\theta) = (S_{2i1}(\theta), \ldots, S_{2iR_i}(\theta))$. The matrices $\mathscr{A}_2(\theta)$ and $\mathscr{B}_2(\theta)$ can be estimated empirically by

$$\widehat{A}_2 = -I^{-1} \sum_{i=1}^{I} \sum_{r=1}^{R_i} \frac{\partial S_{2ir}(\theta)}{\partial \theta'} \mid_{\theta=\widetilde{\theta}}$$

and

$$\widehat{B}_2 = I^{-1} \sum_{i=1}^{J} S_{2i}(\theta)S_{2i}'(\theta) \mid_{\theta=\widetilde{\theta}} .$$

5.3 Study Design

In this section we discuss the cost-effect design of a prospective study in which a Markov model can characterize dynamic features of the process with some clusters providing repeated aggregate data, and others providing longitudinal responses at the individual level. Note that the expected information for both panel and aggregate data will be computed in a robust sandwich form due to the working independence assumption from the composite likelihood (see Section 5.2). We let I_1 denote the number of tanks for assigned to the panel observation scheme and I_2 denote the number of tanks providing only repeated aggregate data. Without loss of generality we suppose tanks $1, \ldots, I_1$ are under the panel and tanks $I_1+1, \ldots, I_1+I_2$ are aggregate observation schemes. The composite likelihood resulting from pooling the data from the panel and aggregate data observation schemes is $L(\theta) = L_1(\theta)L_2(\theta)$ where $L_1(\theta) = \prod_{i=1}^{I_1} \prod_{r=1}^{R_i} L_{1ir}(\theta)$ and $L_2(\theta) = \prod_{i=I_1+1}^{I_1+I_2} \prod_{r=1}^{R_i} L_{2ir}(\theta)$. We let $f = I_1/I$ denote the proportion of tanks that are under panel observation scheme. We let n denote the number of individuals per tank which is fixed and common across all tanks. The cost of observation per individual is C_1 and C_2 for panel and aggregate data observation schemes respectively. The asymptotic robust variance of the maximum composite likelihood estimator is then

$$\mathscr{G}(\theta) = \mathscr{A}(\theta)^{-1}\mathscr{B}(\theta)\mathscr{A}(\theta)^{-1}$$

where

$$\mathscr{A}(\theta) = f\mathscr{A}_1(\theta) + (1-f)\mathscr{A}_2(\theta),$$

$$\mathscr{B}(\theta) = f\mathscr{B}_1(\theta) + (1-f)\mathscr{B}_2(\theta)$$

with the component matrices from (5.6) and (5.8) and

$$\sqrt{I}(\bar{\theta} - \theta) \sim N(0, \mathscr{G}(\theta))$$

with $\bar{\theta}$ being the estimate of θ. Given a target parameter of interest represented by the qth element of θ, the optimal cost-effective design involves allocation of tanks subject to the cost constraint B satisfying

$$\min \left[\mathscr{G}(\theta)\right]_{qq} + \rho[nI(fC_1 + (1-f)C_2) - B] \tag{5.9}$$

where ρ is a Lagrange multiplier. If the interest lies in more than one parameter, we can adopt other optimal allocation methods such as the D-optimality which is widely used in experimental design studies (John and Draper, 1975).

Here we give an example of the cost-effective design for clustered data under a specified setting. Let $n = 10$ for each tank i and let $X_i \sim \text{Bern}(0.5)$ be a tank level covariate, $i = 1, \dots, I$. We use a 5-state progressive process as in the case of the maturation stages of Northern rock sole. We assume 4 follow-up assessment times (not including a_0) and the assessment times are evenly spaced between 0 and 1. We set λ_{12} such that $P(Z_{ij}(1) = 1 | Z_{ij}(0) = 1) = 0.135$. We then set $\lambda_{23} = \lambda_{12}w$, $\lambda_{34} = \lambda_{12}w^2$, and $\lambda_{45} = \lambda_{12}w^3$ with $w = 1.1$ indicating an increasingly rapid progression through the more advanced states, and set $\beta = \log 1.2$. The data is generated such that the entry times to state 2 within each tank are correlated under a copula model (see Section 5.2); the subsequent sojourn times are generated from an exponential distribution. In this example we adopt the Clayton copula with Kendall's τ set to 0 (for independence) or 0.2.

Under the above setting, we now consider the case where the interest lies in study design with the goal is to achieve a pre-specified precision set to 0.01 for the estimator of the regression coefficient. Given the pre-specified variance, Figure 5.1 shows the percentage of aggregate tanks needed to achieve that variance as a function of the cost when Kendall's τ is 0 (left column) and Kendall's τ is 0.2 (right column). Note that when the cost ratio is 1, λ_{12} increases then decreases again. This is due to the fact that our model is strictly progressive and all units start in state 1. Under this particular situation, aggregate data gives similar amount of information as panel data. Moreover, we see that λ_{45} has a strictly increasing curve for cost ratio $= 1$ which corresponds to the fact that aggregate data is losing information comparing to panel data. Note that increasing Kendall's τ increases the cost to achieve the pre-specified variance. Figure 5.2 displays the trade-off between the optimal allocation of tanks and the associated asymptotic variance asymptotic variance when we decrease the budget but keep the constraint that the total number of tanks is the same. The number of tanks is fixed at a number such that we can achieve the pre-specified variance under panel observation scheme. Here we used cost ratio $C_2/C_1 = 0.5$ for illustration purposes. Again, we plot the results in Figure 5.2 for Kendall's τ 0 (left column) and Kendall's τ 0.2 (right column).

Note that we have also superimposed a blue line mimicking the asymptotic variance from a simulation study under the same setting as outputted from Figure 5.2. Moreover, we have done 100 simulation to assess the empirical biases (Ebias), empirical standard errors (ESE) and the robust standard errors (ASE) for β. We see a good agreement between the blue lines (simulated ASE) broadly match of the red (expected asymptotic variance) from Figure 5.3.

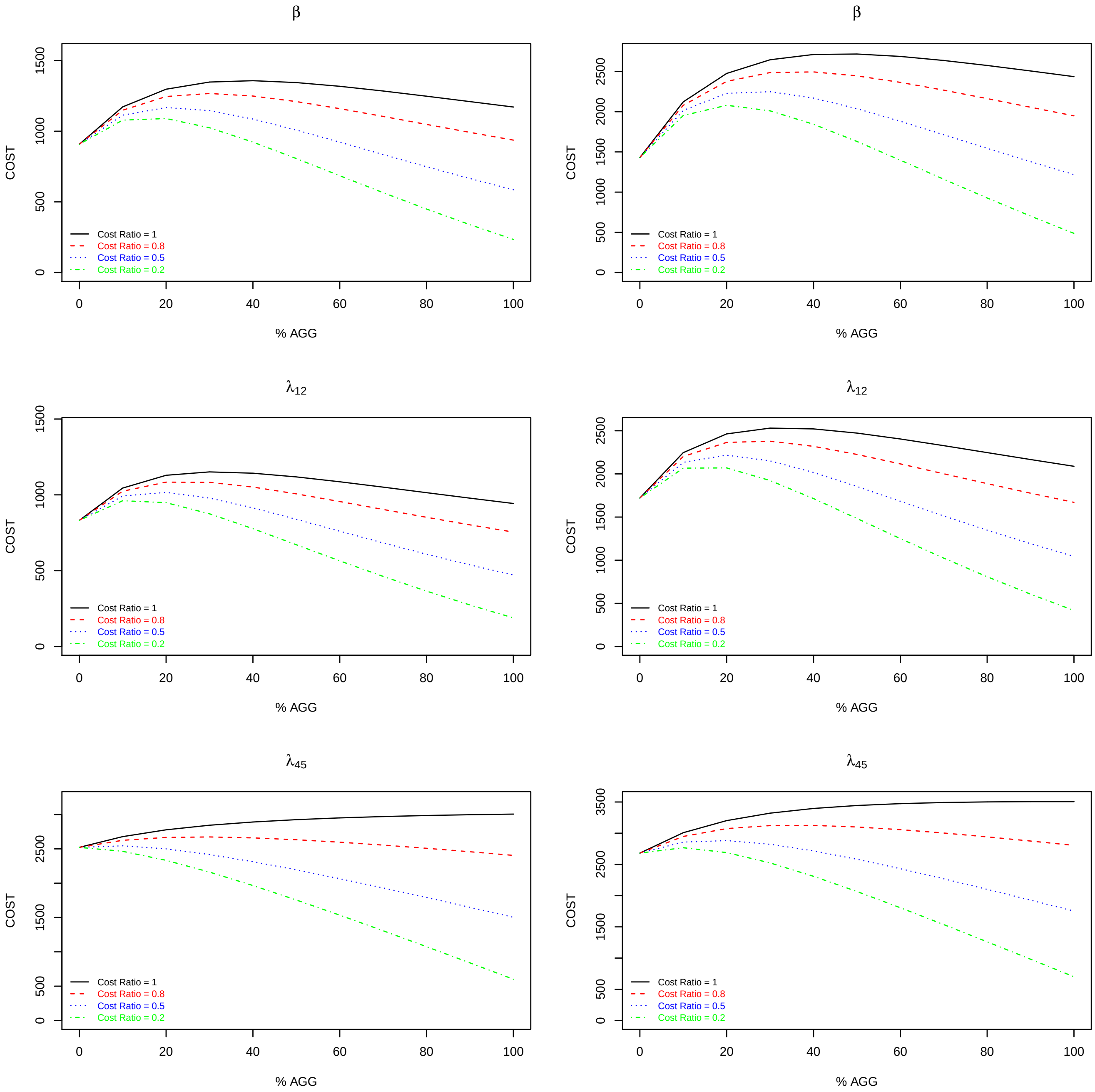


Figure 5.1: Plot of the trade-off between cost and % of aggregated tanks needed to achieve a pre-specified variance with Kendall's $\tau = 0$ on the left column and Kendall's $\tau = 0.2$ on the right column.

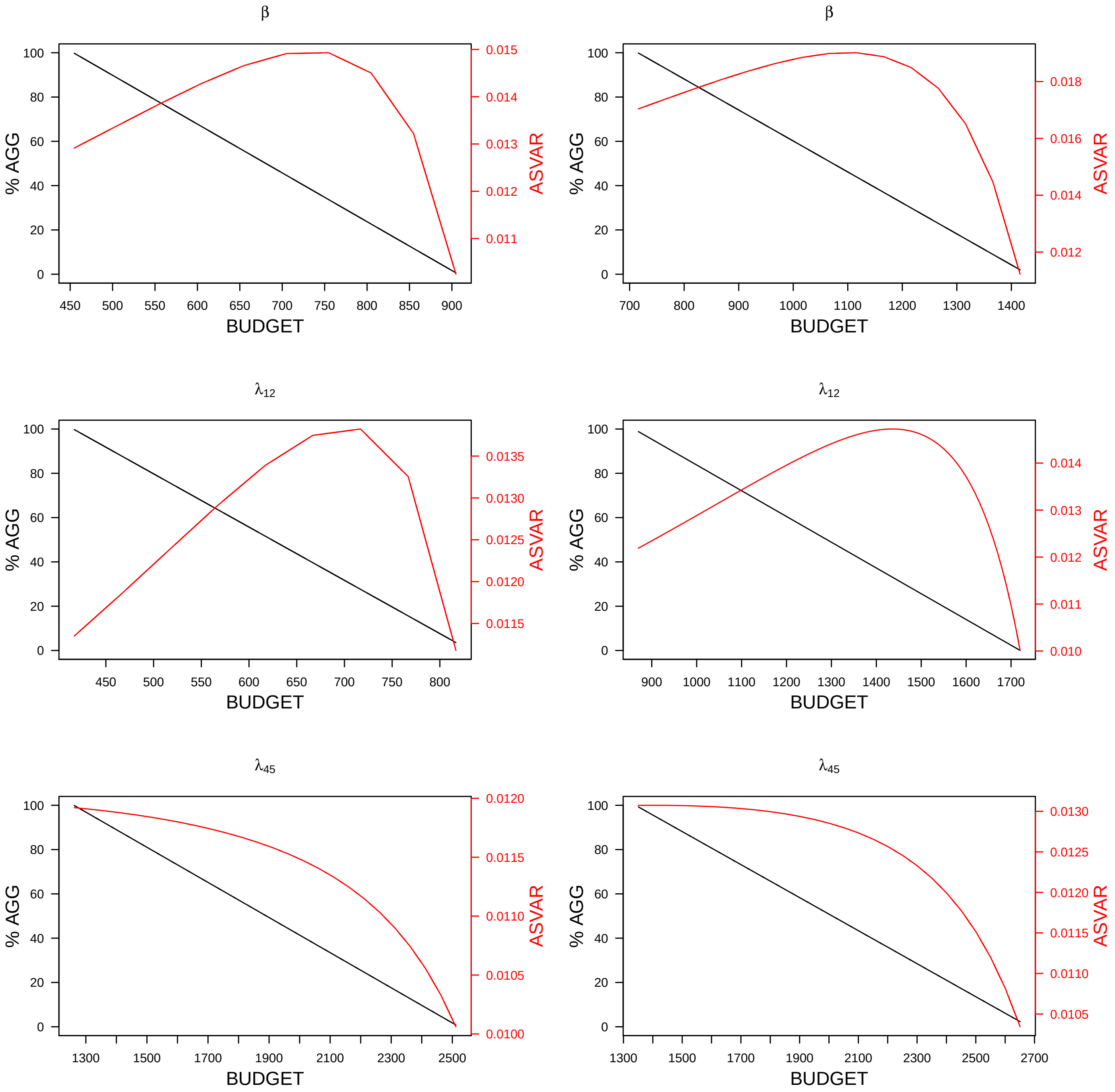


Figure 5.2: Plot of the trade-off between the optimal allocation of % aggregated tanks and their associated asymptotic variance subject to a fixed budget and number of tanks with a cost ratio of 0.5 with Kendall's $\tau = 0$ on the left column and Kendall's $\tau = 0.2$ on the right column.

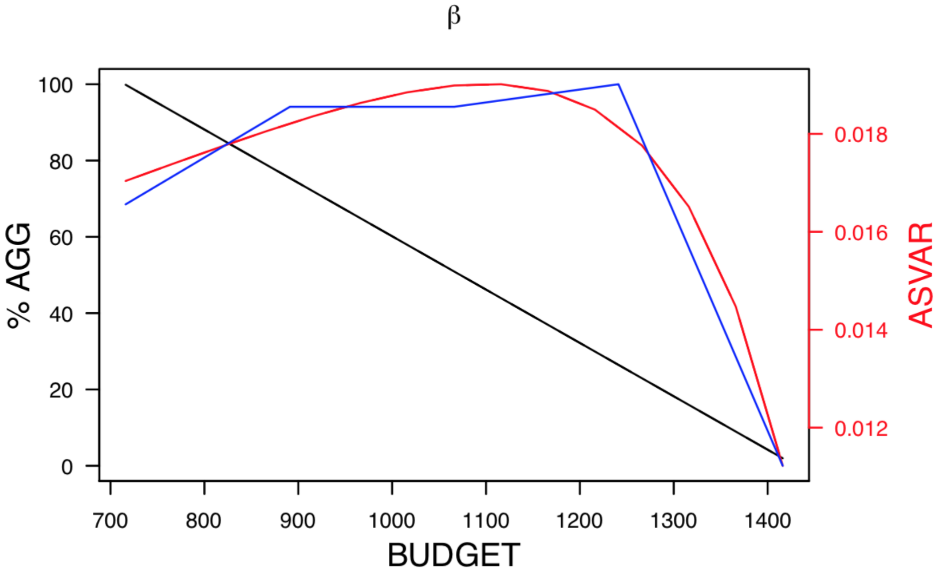


Figure 5.3: Empirical performance of estimators for 100 simulations under a mixture of panel and composite likelihood via marginal model according to a proportion vs. the expected asymptotic variance.

5.4 An Illustrative Study of Northern Rock Sole

We consider the study of the development and growth of Northern rock soles (*Lepidopsetta polyxystra*) (Laurel et al., 2014). The larvae were distributed across 15 tanks under four different temperatures and two observation schemes. Tanks 1-7 were assessed under scheme 1 on days 24, 28, 46, 66 and 90 days while tanks 8-15 were assessed under scheme 2 on days 10, 18, 30 and 47. In many tanks assessment did not occur on the scheduled days leading to incomplete data. Of interest was the effect of temperature on the growth and development of the larvae. Specifically, the larvae pass through a sequence of progressive stages and their development was scored by the degree of observed tail flexion using the criteria established by Hawkyard et al. (2014). Stage 1 of the larvae is characterized by a straight notochord (no flexion); stage 2 larvae have straight notochord with the appearance of caudal peduncle ‘node’ near the posterior end; larvae in stage 3 have a bent notochord with caudal peduncle formation near the posterior end; stage 4 larvae have bent notochord and initial envelopment of the notochord by the caudal peduncle; and stage 5 is characterized by the full envelopment of the notochord by the caudal peduncle with only a remnant of bent notochord still visible (Laurel et al., 2014). Larvae are hard to be identified and may be subjected to a higher cost and error. Laurel et al. (2014) sampled 10 fishes randomly from each tank to potentially avoid identification problems and for convenience. Our method proposed in this paper could provide insights on the optimal cost-effective design for such prospective study.

We first estimate the parameters θ with the constraint $\lambda_{45} = 0.4\lambda_{34}$ due to the sparsity of stage 4 to 5 transitions. We adopt the marginal composite likelihood methods for aggregate data treating temperature as a continuous variable as in Chapter 4. The estimates are shown in Table 5.1. The expected information matrices can then be constructed from the estimates. For demonstration purposes we only consider the observation scheme 1 as our assessment times. We consider both cases when Kendall's $\tau = 0.1$ and 0.2. We set the cost per fish for tracking to be \$1 and this could be multiplied by a factor to accommodate different costs. Figure 5.4 shows the trade-off between cost and % aggregated tanks need to achieve a pre-specified variance which we set to 0.01. We see that when Kendall's τ increase, there's a increase in price needed to achieve the same pre-specified variance. Note that when the cost ratio is about 0.2, we see that going 100% aggregate cost less than 100% panel while achieving the pre-specified variance level. We then let the number of tanks fixed at the number needed to achieve a pre-specified variance of 0.01 when all the tanks are under panel observation scheme which in turn gives us a maximum cost. We then decrease the maximum cost while keep the number of tanks the same and observe the trade-off between optimal allocation of % aggregated tanks and the associated asymptotic variance as shown in Figure 5.5. Note that when Kendall's τ increase, we see an increase both in the asymptotic variance and the budget.

5.5 Discussion

We have described a cost-effective optimal design method based on clustered panel and aggregate data. Aggregate data may be subjected to a lower cost and effort when monitoring organisms. Having aggregate data may also prevent possible misclassification or measurement error when the organisms are hard to identify. The method proposed here gives insight on the trade-off between number of aggregate tanks and panel tanks needed in order to achieve a user-desired variance tolerance. Design can also be considered in terms of power of tests of the cluster-level covariate effects, or other features of the multistate process such as mean sojourn times or median time to maturation. Depending on the cost ratio and the user-desired variance tolerance, one can gain insights on such prospective study with the optimal cost-effective design.

	EST.	S.E.	95% C.I.
$\log\lambda_{12}$	-1.116	0.154	(-2.066, -1.028)
$\log\lambda_{23}$	-1.494	0.201	(-1.888, -1.100)
$\log\lambda_{34}$	-2.575	0.313	(-3.188, -1.962)
$\exp(\beta)$	1.283	0.028	(1.229, 1.339)

Table 5.1: Results of fitting the composite likelihood model under aggregate data setting for the growth of Northern rock soles.

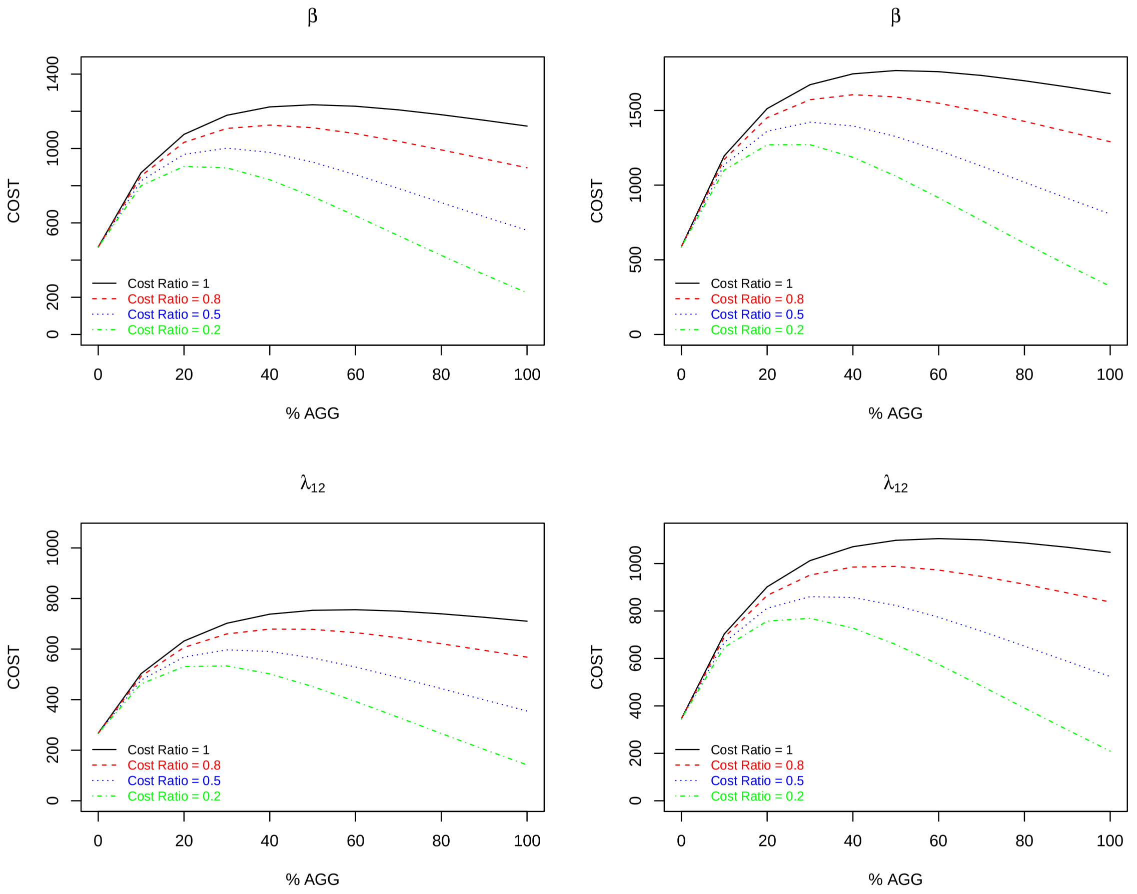


Figure 5.4: Plot of the trade-off between cost and % of aggregated tanks need to achieve a pre-specified variance with Kendall's $\tau = 0.1$ on the left column and Kendall's $\tau = 0.2$ on the right column.

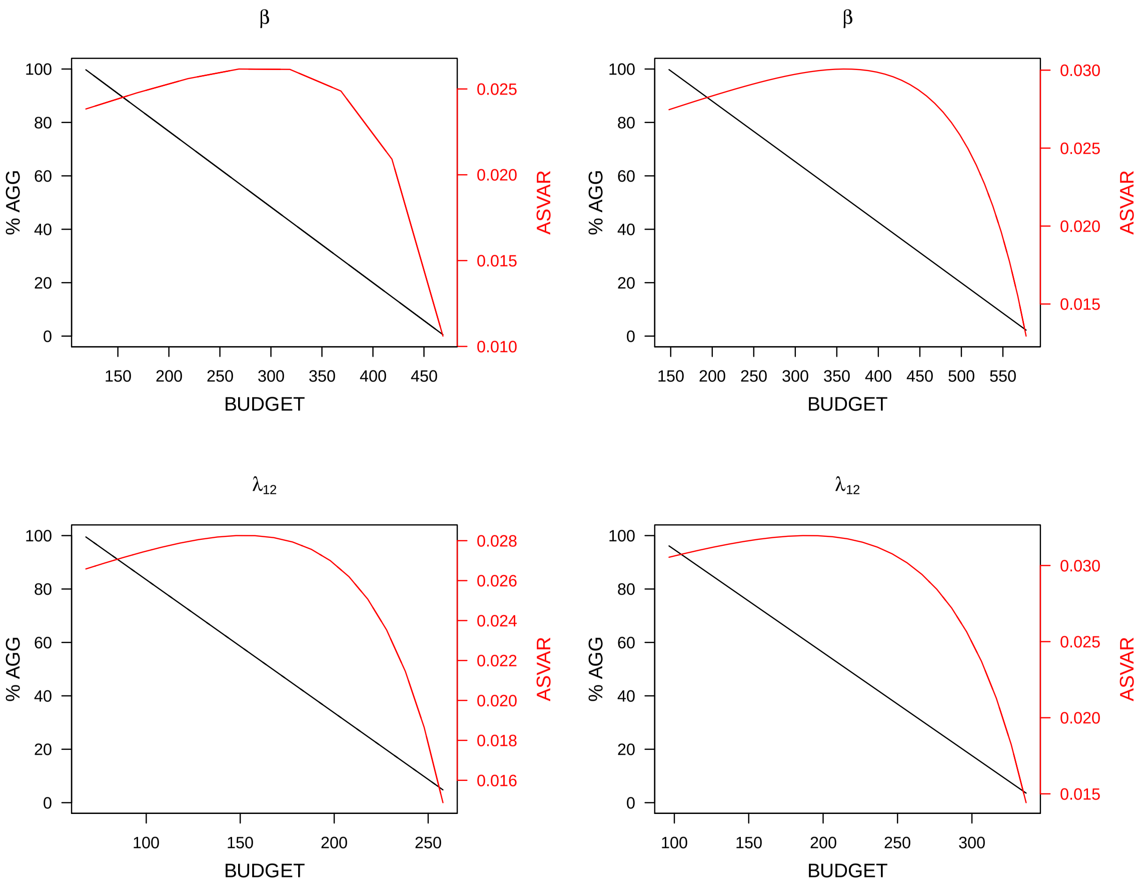


Figure 5.5: Plot of the trade-off between the optimal allocation of % aggregated tanks and their associated asymptotic variance subject to a fixed budget and number of tanks with a cost ratio of 0.25 with Kendall's $\tau = 0.1$ on the left column and Kendall's $\tau = 0.2$ on the right column.

Chapter 6

Future and Ongoing Work

6.1 Finite Mixture Models for Multistate Panel Data

In Chapter 2 attention was restricted to the case in which the transition intensities for the unilateral classes (left or right) are the same as the corresponding are in the bilateral class. We also assume that the marker effects are the same for two unilateral classes. It is reasonable to assume these constraints on scientific grounds in this context, but in other settings it may be desirable to relax these constraints to obtain more flexible models and test the plausibility of these assumptions. Fitting of this more flexible model may be feasible with larger datasets and estimability issues may arise if the dataset is not sufficient to support the estimation of different margins.

We have assumed that the observation process satisfies the sequential missing at random assumption of Hogan et al. (Hogan et al., 2004). An alternative approach to our analysis would be to predict the presence and nature of back involvement at a particular point in the disease course based on direct multinomial regression; in this case inverse intensity of visit weights (Lin et al., 2004) are required to adjust for the selection bias arising from the need to restrict attention to individuals who can be definitively classified at the landmark time. These may also be required when assessing the predictive accuracy if a fitted model with a validation sample. We do not explore this here as the modeling of a nonsusceptible fraction is less natural in a marginal framework such as this; see Cook and Lawless (2018).

6.2 Marginal Mixture Models for Multivariate Interval-censored Times

6.2.1 Accommodating Higher Dimensional Responses

The model in Chapter 3 was studied in the context of bivariate processes. In some settings interest may lie in modeling higher dimensional interval-censored data while accommodating a non-susceptible fraction and understanding the association between latent marginal processes. Suppose J events may be realized. In this case the multivariate binary model for the susceptibility indicators naturally extends, where one may let

$$\psi_{ijj'} = \frac{P(Z_{ij}=1, Z_{ij'}=1|X_i)P(Z_{ij}=0, Z_{ij'}=0|X_i)}{P(Z_{ij}=1, Z_{ij'}=0|X_i)P(Z_{ij}=0, Z_{ij'}=1|X_i)} \tag{6.1}$$

denote the odds ratio for type j and j' susceptibility indicators. These odds ratios may be different for each possible pair j and j', but we note that while there are several models that can be considered, there are constraints on the admissible odds ratios with multivariate binary data (Liang et al., 1992).

For the failure times, let $\tau_{jj'}$ denote Kendall's τ for the association between T_{ij} and $T_{ij'}$ given $Z_{ij} = Z_{ij'} = 1$. A multivariate Gaussian copula function seems the most natural to consider in this setting since the pairwise dependencies in failure times may differ across pairs given joint susceptibility.

The likelihood, two-stage and estimating function approaches of Section 3 may all be extended and employed to deal with higher dimensional failure time models. Composite likelihood may also offer an appealing and computationally convenient option in this setting (Varin, 2008), with options for pairwise and higher-order forms. With a higher dimensional response, the risk of mis-specifying the dependence structure is greater, so the robustness of the composite likelihood, two-stage approach to estimation or the estimating function approaches may make them preferable to maximum likelihood.

6.2.2 Analysis of Current Status Data

Many large cohort studies are being conducted around the world with a view to collecting data on disease prevalence, incidence and progression. Such cohort studies often involve complex survey design which involves recruiting individuals from the population according to a stratified sampling scheme. Specifically individuals may be recruited according strata defined by age and sex and furnish information on their disease status (present or absent) at the time of first contact. This current status data represents a special case of the intermittent observation scheme of Section 3 where disease onset times are either left- or right-censored. The pooled adjacent violators algorithm (PAVA) is a widely used method for obtaining nonparametric maximum likelihood estimates of the survivor function or onset time distribution based on current status data (Ayer et al., 1955) . The corresponding estimator for $\mathscr{F}(\cdot)$ is a step function which can jump at each observed inspection time, or a subset of inspection times, but cannot jump at any other time.

Here we can consider a two-stage approach to fitting the model of Chapter 3 to current status data. We can adopt PAVA to estimate $\mathscr{F}(\cdot)$ and the susceptibility model for each process, and then use the estimates from stage 1 to obtain estimates of the association parameters in stage 2. Let $Y_{ij} = I(T_{ij} < C_i)$ indicate that the type j event was known to occur for individual i. Let $\alpha_1 = (\beta', \mathscr{F}_j'(\cdot), j = 1, 2)'$. The complete date composite likelihood at Stage 1 can then be written as

$$\mathscr{L}_1(\alpha_1) \propto \prod_{i=1}^{n}\prod_{j=1}^{2} \big([1-\mathscr{F}_j(C_i)]\,\mu_{ij}\big)^{z_{ij}y_{ij}} \big([\mathscr{F}_j(C_i)\mu_{ij}]^{z_{ij}}\,[1-\mu_{ij}]^{1-z_{ij}}\big)^{1-y_{ij}} \tag{6.2}$$

and

$$\begin{aligned}\log\mathscr{L}_1(\alpha_1) = &\sum_{i=1}^{n}\sum_{j=1}^{2} z_{ij}\log\mu_{ij} + (1-z_{ij})\log(1-\mu_{ij})\\ &+ z_{ij}\left[y_{ij}\log(1-\mathscr{F}_j(C_i)) + (1-y_{ij})\log\mathscr{F}_j(C_i))\right]\end{aligned} \tag{6.3}$$

At the rth iteration of the EM algorithm, we will have

$$\begin{aligned}Q_{1ij}(\beta_j;\alpha_1^r) &= \eta_{ij}^r\log\mu_{ij} + (1-\eta_{ij}^r)\log(1-\mu_{ij})\\ Q_{2ij}(\mathscr{F}_j(\cdot);\alpha_1^r) &= \eta_{ij}^r\left[y_{ij}\log(1-\mathscr{F}_j(C_i)) + (1-y_{ij})\log\mathscr{F}_j(C_i)\right]\end{aligned}$$

where

$$\eta_{ij}^r = P(Z_{ij}=1|Y_{ij}=0,D_i;\alpha_1^r) = \frac{[\mathscr{F}_j(C_i)]\,\mu_{ij}}{[\mathscr{F}_j(C_i)]\,\mu_{ij} + (1-\mu_{ij})}, \quad j=1,2.$$

Optimization of (6.2) can be carried out by adapting the PAVA. Specifically, for each process j, the number of individuals who are estimated to be at risk at the kth inspection time in the mth iteration is $\widehat{\triangle}_{jk}^m$ where $\widehat{\triangle}_{jk}^m = \sum_{i=1}^n I(C_i = C_k)(Y_{ij} + (1-Y_{ij})\eta_{ij}^m)$. An isotonic regression of $(r_1/\widehat{\triangle}_{j1}^m, ..., r_K/\widehat{\triangle}_{jK}^m)'$ with weights $(\widehat{\triangle}_{j1}^m, ..., \widehat{\triangle}_{jK}^m)'$ gives

$$\widehat{\mathscr{F}}^{(m+1)}(C_k) = \max_{u\le k}\min_{v\ge k}\left(\frac{\sum_{l=u}^{v}\sum_{i=1}^{n} r_l}{\sum_{l=u}^{v}\widehat{\triangle}_{jk}^m}\right).$$

With the estimates from stage 1 then, we are able to obtain the association parameters by maximizing (3.5). Survey weights can also be incorporated into the likelihood or the EM algorithm for the nonparametric estimation of the onset time distribution based on the PAVA, this would be appropriate for cross-sectional studies employing a complex survey design.

6.2.3 Other Extensions

There are several directions for further development of the proposed model in Chapter 3. Smoothed estimates of the marginal densities for the failure times can be developed (Hjort and Jones, 1996; Li et al., 1997) or smoothed estimates of baseline hazards can be obtained by local likelihood (Betensky et al., 1999). Smoothing of multivariate failure time distributions based on interval-censored data can also be carried out (Braun and Stafford, 2016) but this will not yield parsimonious measures of dependence in the failure times.

Often it is most natural to consider covariate effects on the latent susceptibility indicators but if there is interest in modeling covariate effects on the failure times semiparametric proportional hazards or additive models can be considered.

6.3 Analysis of Aggregate Data from Clustered Multistate Processes

In Chapter 4 we introduced a composite likelihood approach to handle aggregate data using either marginal methods with robust variance estimation, or random effects models to accommodate clustering of transition rates within tanks. In principle, however, one could consider relaxing the working independence assumption within tanks in the marginal approach to estimate the dependence parameter as well.

We restricted attention to time homogeneous transition intensities, but this can be relaxed easily to accommodate a piecewise-constant form. This was not done in the application because of the close alignment of the empirical estimates and the estimates based on the proposed model. Extensions may be developed for recurrent processes or processes involving a terminal (e.g. death) state which can be entered at any time during the maturation process, but settings involving multistate models with reversible transitions are much more difficult to handle even under the panel observation scheme.

6.4 Cost-effective Design with Aggregation and Tracking

In some settings it may be feasible to tag organisms to enable tracking of individuals, but this may incur a cost. If it is possible, it may be of interest to consider the cost-benefit of tracking individual organisms. The framework we have described in Chapter 5 can be generalized in a number of ways. In some settings it may only be possible to record aggregate data at certain phases of the development process (i.e. at the larval stage) but it may be possible to tag or otherwise identify organisms when they are more developed. In this case aggregate data may be available at early stages but tracking of individuals may yield panel observations once a certain stage of the life cycle has been reached. Another interesting variation of this design is to allow timing of assessments to differ between tanks. Some tanks, for example, may be examined more frequently at the early stages of the life cycle and others may be examined more frequently at later stages. Optimal allocation of the tanks to these observation schedules can also be considered.

In Chapter 5 we have considered a cost-effect design for a strictly progressive model that deals with growth studies. In a completely different setting this issue arises in school-based studies of health knowledge, attitudes and behaviour among youth. Here tracking of individuals may require greater effort to get ethics approvals in comparison to repeat cross-sectional studies, which offer data more like the aggregate data in our setting. However school-based studies also feature immigration and emigration which mean any models based on marginal aggregate summaries must accommodate the fact that some new students may have entered the school and some may have left; such data may be available from school administrators.

References

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The Annual of Mathematical and Statistics*, 26:641 – 647, 1955.

P. H. Bennett and P. H. N. Wood. Population studies of the rheumatic diseases. *Amsterdam, Excerpta Medica*, pages 477 – 478, 1968.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36:192 – 236, 1974.

R. Betensky and D. Finkelstein. A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, 18:3089 – 3100, 1999.

R. A. Betensky, J. C. Lindsey, L. M. Ryan, and M. P. Wand. Local EM estimation of the hazard function for interval-censored data. *Biometrics*, 55:238 – 245, 1999.

D. Boos. On generalized score tests. *Journal of the American Statistical Association*, 46: 327 – 333, 1992.

D. L. Borror and R. E. White. *A Field Guide to the Insects of Amnerica North of Mexico.* Boston: Houghton Mifflin, 1970.

W. J. Braun and J. E. Stafford. Multivariate density estimation for interval-censored data with application to a forest fire modelling problem. *Environmetrics*, 27:345 – 354, 2016.

A. T. Cate. Maximum likelihood estimation of the Markov chain model with macro data and the ecological inference model. *CPB Netherlands Bureau for Economic*, 2014.

N. Chatterjee and J. Shih. A bivariate cure-mixture approach for modeling familial association in diseases. *Biometrics*, 57:779 – 786, 2001.

R. J. Cook and J. F. Lawless. *The Statistical Analysis of Recurrent Events*. Springer, New York, 2007.

R. J. Cook and J. F. Lawless. *Multistate Models for the Analysis of Life History Data*. Springer, New York, 2018.

R. J. Cook, J. D. Kalbfleisch, and G. Y. Yi. A generalized mover-stayer model for panel data. *Biostatistics*, 3:407 – 420, 2002.

R. J. Cook, G. Y. Yi, and K. A. Lee. A conditional Markov model for clustered progressive multistate processes under incomplete observation. *Biometrics*, 60:436 – 443, 2004.

R. J. Cook, B. J. White, G. Y. Yi, and K. A. Lee. Analysis of a nonsusceptible fraction with current status data. *Statistics in Medicine*, 27:2715 – 2730, 2008a.

R. J. Cook, L. Zeng, and K. A. Lee. A multistate model for bivariate interval-censored failure time data. *Biometrics*, 64:1100 – 1109, 2008b.

D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91:729 – 737, 2004.

M. Crowder and D. Stephens. On inference from Markov chain macro-data using transforms. *Journal of Statistical Planning and Inference*, 141:3201 – 3216, 2011.

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

L. Diao and R. J. Cook. Composite likelihood for joint analysis of multiple multistate processes via copulas. *Biostatistics*, 15:690 – 705, 2014.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

V. T. Farewell. A model for a binary variable with time censored observations. *Biometrika*, 64:43 – 46, 1977.

V. T. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041 – 1046, 1982.

J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

H. Frydman. Maximum likelihood estimation in the mover-stayer model. *Journal of American Statistical Association*, 79:632 – 638, 1984.

C. Fuchs and J. B. Greenhouse. The EM algorithm for maximum likelihood estimation in the moverstayer model. *Biometrics*, 44:605 – 613, 1988.

R. C. Gentleman, J. F. Lawless, J. C. Lindsey, and P. Yan. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13:805 – 821, 1994.

D. Gladman, D. Ibañez, and M. B. Urowitz. Systemic lupus erythematosus disease activity index 2000. *Rheumatology*, 29:228 – 291, 2002.

D. D. Gladman and V. Chandran. Observational cohort studies: lessons learnt from the University of Toronto psoriatic arthritis program. *Rheumatology*, 50:25 – 31, 2010.

E. Gouno, L. Courtrai, and M. Fredette. Estimation from aggregate data. *Computational Statistics and Data Analysis*, 55:615 – 626, 2011.

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

J. G. Hanly, M. L. Russell, and D. D. Gladman. Psoriatic spondyloarthropathy: a long term prospective study. *Ann. Rheumatol. Dis.*, 47:386 – 393, 1988.

M. Harron, R. Winchster, J. T. Giles, E. Heffernan, and O. FitzGerald. Certain class I HLA alleles and haplotypes implicated in susceptibility play a role in determining specific features of the psoriatic arthritis phenotype. *Clinical and Epidemiological Research*, 75: 155 – 162, 2016.

M. Hawkyard, B. Laurel, and C. Landon. Rotifers enriched with taurine by microparticulate and direct enrichment methods influence the growth and metamorphic stage of northern rock sole (*Lepidopsetta polyxystra*) larvae. *Aquaculture*, 425:157–151, 2014.

D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *Ann. Statist.*, 4:2244 – 2253, 1991.

N. Hjort and M. Jones. Locally parametric nonparametric density estimation. *Annals of Statistics*, 24:1619 – 1647, 1996.

J. W. Hogan, J. Roy, and C. Korkontzelou. Handling drop-out in longitudinal studies. *Statistics in Medicine*, 23:1455 – 1497, 2004.

P. Hougaard. *Analysis of Multivariate Survival Data*. Springer, 2012.

J. A. Husted, B. D. Tom, V. T. Farewell, C. T. Schentag, and D. D. Gladman. A longitudinal study of the effect of disease activity and clinical damage on physical function over the course of psoriatic arthritis: Does the effect change over time? *Arthritis and Rheumatology*, 56:840 – 849, 2007.

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

S. Jiang and R. J. Cook. Score tests based on a finite mixture model of markov processes under intermittent observation. *Submitted to Statistics in Medicine*, 2018a.

S. Jiang and R. J. Cook. Analysis of multivariate interval-censored failure times with dependent susceptibility. *Submitted to Statistics in Biosciences*, 2018b.

S. Jiang and R. J. Cook. Analysis of aggregate data from clustered multistate processes via composite likelihood. *Submitted to Biometrics*, 2018c.

S. Jiang and R. J. Cook. Cost-effective design of growth studies with aggregation and tracking. *Submitted to Biometrics*, 2018d.

H. Joe. *Multivariate Dependence Concepts*. Chapman and Hall, London, 1997.

R. C. S. John and N. R. Draper. D-Optimality for regression designs: A review. *Technometrics*, 17:15–23, 1975.

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

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data.* John Wiley and Sons, 2002.

J. D. Kalbfleish and J. F. Lawless. Estimation in Markov models from aggregate data. *Biometrics*, 39:907 – 919, 1983.

M. Y. Kim and X. Xue. The analysis of multivariate interval-censored survival data. *Statistics in Medicine*, 21:3715 – 3726, 2002.

Y.-J. Kim. Cure rate model with bivariate interval censored data. *Communication in Statistics – Simulation and Computation*, 0:1 – 9, 2016.

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data.* Springer, New York, 2003.

K. F. Lam and H. Xue. A semiparametric regression cure model with current status data. *Biometrika*, 92:573 – 586, 2005.

B. Laurel, C. Danley, and S. Haines. The effects of temperature on growth, development and settlement of northern sole larvar (*Lepidopsetta polyxystra*). *Fisheries oceanography*, 23(6):495–505, 2014.

J. F. Lawless. *Statistical Models and Methods for Lifetime Data, 2nd edition.* John Wiley and Sons, Hoboken, 2003.

S. T. Leatherdale, K. S. Brown, V. Carson, R. A. Childs, J. A. Dubin, S. J. Elliot, and C. M. Sabiston. The COMPASS study: A longitudinal hierarchical research platform for evaluating natural experiments related to changes in school-level programs, policies and built environment resources. *BMC Public Health*, 14:331, 2014.

L. Li, T. Watkins, and Q. Yu. An em algorithm for smoothing the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics*, 24: 531 – 542, 1997.

K. Y. Liang, S. L. Zeger, and B. Qaqish. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54:3 – 40, 1992.

H. Lin, D. O. Scharfstein, and R. A. Rosenheck. Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:791 – 813, 2004.

S. V. D. Linden, H. A. Valkenburg, and A. Cats. Evaluation of diagnostic criteria for ankylosing spondylitis. a proposal for modification of the New York criteria. *Arthritis Rheum.*, 27:361 – 368, 1984.

B. G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:221 – 239, 1988.

B. G. Lindsay, G. Y. Yi, and J. Sun. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21:71 – 105, 2011.

S. R. Lipsitz and N. M. Laird. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika*, 78:153 – 160, 1991.

C. Loader. Local likelihood density estimation. *Annals of Statistics*, 24:1602 – 1618, 1996.

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

E. C. MacRae. Estimation of time-varying Markov processes with aggregate data. *Econometrica*, 45:183 – 198, 1977.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London New York, 1989.

P. L. Munholland and J. D. Kalbfleisch. A semi-Markov model for insect life history data. *Biometrics*, 47:1117 – 1126, 1991.

R. B. Nelsen. *An Introduction to Copulas*. Springer, New York, 2006.

A. O'Keeffe, B. Tom, and V. Farewell. Mixture distributions in multistate modelling: some considerations in a study of psoriatic arthritis. *Statistics in Medicine*, 32:600 – 619, 2013.

A. Pasanisi, S. Fu, and N. Bousquet. Estimating discrete markov models from various incomplete data schemes. *Computational Statistics and Data Analysis*, 56:2609 – 2625, 2012.

T. Peng. Fitting semiparametric cure models. *Computational Statistics and Data Analysis*, 41:481 – 490, 2003.

T. Peng, K. B. G. Dear, and K. C. Carrier. Testing for the presence of cured patients: a simulation study. *Statistics in Medicine*, 20:1783 – 1796, 2001.

R. L. Prentice and L. P. Zhao. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 57:825 – 839, 1991.

P. Rahman, D. D. Gladman, R. J. Cook, Y. Zhou, G. Young, and D. Salonen. Radiological assessment in psoriatic arthritis. *British journal of rheumatology*, 37:760 – 765, 1998.

J. P. Reinecke, J. S. Buckner, and S. R. Grugel. Life cycle of laboratory-reared tobacco hornworms *Manduca sexta*, a study of development and behavior, using time-lapse cinematography. *The Biological Bulletin*, 158:129–140, 1980.

B. L. Riggs, H. W. Wahner, W. L. Dunn, R. B. Mazess, K. P. Offord, and L. J. Melton. Differential changes in bone mineral density of the appendicular and axial skeleton with ageing: relationship to spinal osteoporosis. *Journal of Clinical Investigation*, 67:328–335, 1981.

G. A. Satten. Estimating the extent of tracking in interval-censored chain-of-events data. *Biometrics*, 55:1228 – 1231, 1999.

J. Shih and T. A. Louis. Inference on the association parameter in copula models for bivariate survival data. *Biometrics*, 51:1384 – 1399, 1995.

J. Sun. *The Statistical Analysis of Interval-censored Failure Time Data*. Springer, New York, 2006.

R. Sutradhar and R. J. Cook. Analysis of interval-censored data from clustered multistate processes:application to joint damage in psoriatic arthritis. *Journal of Royal Statistical Society*, 57:553 – 566, 2008.

M. J. Sweeting, D. D. Angelis, K. R. Neal, M. E. Ramsay, W. L. Irving, M. Wright, L. Brant, H. E. Harris, T. H. S. Group, and H. N. R. S. Group. Estimated progression rates in three united kingdom hepatitis c cohorts differed according to method of recruitment. *Journal of Clinical Epidemiology*, 59:144 – 152, 2006.

J. P. Sy and J. M. G. Taylor. Estimation in a Cox proportional hazards cure model. *Biometrics*, 56:227 – 236, 2000.

D. Tolusso and R. J. Cook. Second-order estimating equations for the analysis of clustered current status data. *Biostatistics*, 10:756 – 772, 2009.

S. L. Tyas, J. C. Salazar, D. A. Snowdon, M. F. Desrosiers, K. P. Riley, M. S. Mendiondo, and R. J. Kryscio. Transitions to mild cognitive impairments, dementia, and death: Findings from the nun study. *American Journal of Epidemiology*, 165:1231 – 1238, 2007.

C. Varin. On composite marginal likelihoods. *Advances in Statistical Analysis*, 92:1 – 28, 2008.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 2011.

W. Wang and A. A. Ding. On assessing the association for bivariate current status data. *Biometrika*, 87:879 – 893, 2000.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1 – 25, 1982.

Y. Wu and R. J. Cook. Variable selection and prediction in biased samples with censored outcomes. *Lifetime Data Analysis*, 24:72 – 93, 2018.