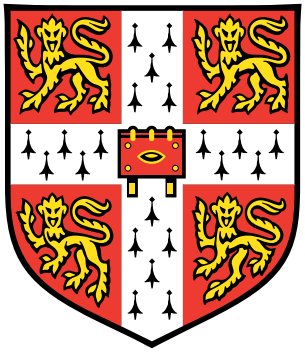
# Model diagnostics in multi-state models of biological systems

A dissertation submitted to the University of Cambridge
Department of Pure Mathematics and Mathematical Statistics
for the degree of Doctor of Philosophy
October 2007

Andrew Charles Titman



Fitzwilliam College
University of Cambridge

# Declaration

All writing and analysis in this thesis was carried out by myself, between October 2004 and October 2007. To my knowledge, all work is original except where referenced. No work is included which results from collaboration with others. None of the work contained in this thesis has been submitted for a degree or other qualification at any other university.

# Acknowledgements

My sincerest thanks go to my supervisor Linda Sharples for her excellent help and guidance throughout the project. Thanks also to Chris Jackson for many insightful comments and Vern Farewell for adeptly directing the course of the thesis at key stages. I am also grateful to Susan Pitts at the Statistical Laboratory for additional support.

This work was financed by a grant awarded by the UK Medical Research Council.

## Summary

Multi-state models can be used to explain categorical longitudinal data. In medical applications it is common to have only panel data which may be observed at time intervals that are irregular and vary among subjects. In addition the observed states may be subject to misclassification error. This thesis concentrates on multi-state models in the context of panel observed data.

Fitting models to such data is difficult, both in terms of computation and parameter identifiability, unless assumptions about the process and the mechanism for misclassification are made. It is often assumed that the process is Markov, with many studies also based on time homogeneity. For models with state misclassification a hidden Markov model structure in which, conditional on the true states, the observed states are independent, is usually assumed. Similarly, patients are commonly considered homogeneous, at least conditional on known covariates.

Methods for assessing the appropriateness of these assumptions are relatively underdeveloped, particularly for models with irregular observation schemes, exact death times or misclassified observed states. This thesis concentrates on the development of diagnostics for such models. Whilst the focus is on general methods, two datasets relating to chronic disease in post-transplant patients are analysed as motivating examples.

A review of the existing literature and a thorough investigation of particular diagnostic tools is given in the first two chapters. This shows that existing methods are limited in the range of models they can be applied to and the type of model departures they can detect. Chapter 3 therefore develops a general goodness-of-fit test for Markov and hidden Markov models, extending previous work by Aguirre-Hernández and Farewell.

Chapter 4 assesses the effect of some types of model misspecification on inference, using asymptotic approximations. Since informal and general goodness-of-fit tests lack power in detecting certain types of misspecification, the remainder of the thesis concentrates on developing methods for fitting time dependent models to panel observed data. Methods using piecewise-constant intensities or parametric forms can be applied for time inhomogeneous Markov models. Such methods can also be used in semi-Markov models, but implementation is difficult and only possible in a limited range of cases. However, it is shown that, by allowing the time in each state to follow a phase-type distribution, semi-Markov models can be expressed as a type of hidden Markov model. This allows a very wide range of models to be fitted with relative ease.

# List of Abbreviations

| | |
|---|---|
| AH/F | Aguirre-Hernández, R and Farewell V.T. (2002) |
| AIC | Akaike Information Criterion |
| BFGS | Broyden-Fletcher-Goldfarb-Shanno optimisation algorithm |
| BOS | Bronchiolitis obliterans syndrome |
| CAV | Cardiac allograft vasculopathy |
| cdf | cumulative distribution function |
| CLT | Central limit theorem |
| EM | Expectation-Maximisation |
| $FEV_1$ | Forced expiratory volume in one second |
| HMM | Hidden Markov model |
| HSMM | Hidden semi-Markov model |
| IHD | Ischemic heart disease |
| i.i.d. | Independent, identically distributed |
| LL | log-likelihood |
| MCEM | Monte-Carlo Expectation-Maximisation |
| MCMC | Markov chain Monte-Carlo |
| MH | Metropolis-Hastings |
| MI | Multiple Imputation |
| mle | maximum likelihood estimate |
| MVN | Multivariate normal |
| ODE | Ordinary differential equation |
| pdf | probability density function |
| QSD | Quasi-stationary distribution |

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Overview

Multi-state models are an approach to analysing categorical longitudinal data. These types of model are used particularly in medical applications in which stages or levels of a disease are represented by the states in the model. Such models have been used in a wide range of medical applications, for instance HIV/AIDS [3, 55, 68, 95, 115], human papillomavirus [14, 72], breast cancer [43, 62, 102], psoritic arthritis [30], liver cirrhosis [8], dementia [69], diabetic retinopathy [79, 94] and smoking prevention [29, 70]. The discrete states of the model may either represent clinically defined stages of a disease, e.g. number of damaged joints in patients with psoriatic arthritis, or alternatively be a discretisation of a continuous marker, e.g. CD4 count in patients infected with HIV. Multi-state models have been applied to chronic diseases affecting post-transplantation patients [65, 124]. In this thesis datasets relating to development of cardiac allograft vascopathy (CAV) in post-heart-transplant patients and bronchiolitis obliterans syndrome (BOS) in post-lung-transplant patients will be used to illustrate the methods.

Whilst there are many applications in which complete history is observed, at least up to right censoring [61], these models are particularly useful when the data are collected under panel observation. In this case the data are of the form of a series of observations $x_1, \ldots, x_N$ at a discrete set of sampling times $t_1, \ldots, t_N$, which may vary between subjects. In this setting it is particularly important to have a model that is sufficiently simple to ensure that the likelihood function is tractable and the parameters are identifiable. Markov models provide a reasonably flexible class of models which can be fitted to such

data. Moreover, hidden Markov models allow the possibility of misclassification of the observed states to be accommodated. Recently, with improvements in computing power and developed software, such models have become easier to fit, particularly if an assumption of time homogeneity is made. However, these approaches make strong assumptions about the process. Currently methods for assessing the appropriateness of these model assumptions are not well developed.

This thesis seeks to appraise existing, and develop new, methods for testing model fit, and identify areas of model departure in multi-state models. Specifically the focus is on panel observed data, including the cases where the observed state is subject to misclassification and when the time of entry into an absorbing state is known. It is assumed that the baseline intensities, covariate effects and, where appropriate, misclassification probabilities, may be quantities of interest and not necessarily nuisance parameters. The thesis is limited to methods for diagnosing parametric models. Non-parametric and semi-parametric methods will only be explored as tools for assessing parametric model fit. Particular emphasis is placed on diagnostics for time homogeneous Markov models since these are the most common multi-state models fitted and have the strongest assumptions. Importance is also placed on developing specific tests for time dependent alternatives.

The remainder of the chapter gives an introduction to the theory and notation to be used in the remainder of the thesis. There is then a review of the existing literature related to assessment of model fit in multi-state models and methods for fitting models with time dependent transition intensities. Chapter 2 firstly introduces the CAV and BOS datasets to be used in the thesis. The remainder of the chapter considers in more detail some of the existing informal diagnostics for model fit, applying them to the CAV and BOS datasets. Chapter 3 develops existing work by Aguirre-Hernández and Farewell [6] on a general Pearson-type goodness-of-fit test for Markov models. The test is extended to allow application on misclassification hidden Markov models. It is shown that the existing test is not appropriate for datasets with exact death times. A modified test is developed to deal with this case.

In chapter 4 there is an investigation of the potential effects of model misspecification in some specific cases. Chapter 5 reviews and develops methods for fitting time inhomogeneous Markov and semi-Markov models. In chapter 6 an approach to fitting semi-Markov models in which the sojourn times have phase-type distributions is developed. The final chapter gives an overview of the results and a discussion of areas of further work.

## 1.2 Multi-state models

Typically in multi-state models with irregular sampling schemes, the movement between a discrete set of states $\mathcal{S} = \{1, \ldots, R\}$ is governed by a continuous time stochastic process $X(t)$ which takes values in $\mathcal{S}$. The simplest multi-state model is the survival model in which subjects begin in the state 'alive' and progress to the absorbing state 'death'. Other simple multi-state models include the illness-death model in which subjects can progress from 'well' to 'death' possibly via a state 'ill'. Typically in this situation the interest will be in determining either the rate of progression to 'ill' or the relative rate to death from 'ill' compared to 'well'.

It is usual to define a multi-state model by its matrix of *transition intensities*, $Q(t, \mathcal{F}_t)$, with $(r, s)$ entry

$$q_{rs}(t, \mathcal{F}_t) = \lim_{\delta t \downarrow 0} \frac{\mathbb{P}(X(t + \delta t) = s | X(t) = r, \mathcal{F}_t)}{\delta t}$$

where $\mathcal{F}_t$ represents the history (or filtration) of the process up to time $t$. The *transition probabilities* are defined as

$$p_{rs}(t_1, t_2, \mathcal{F}_{t_1}) = \mathbb{P}(X(t_2) = s | X(t_1) = r, \mathcal{F}_{t_1}).$$

Some further assumptions are usually necessary when fitting multi-state models to data. This is particularly true when the data are collected under panel observation. This means that, at best, transition times are interval censored, so that a transition time is known to have occurred within a certain interval. More often it will also mean that the precise number and nature of the transitions will not be known. For example figure 1.1 shows the possible evolution of a multi-state process, the discrete observation scheme causes the second sojourn in state 1 to be missed.

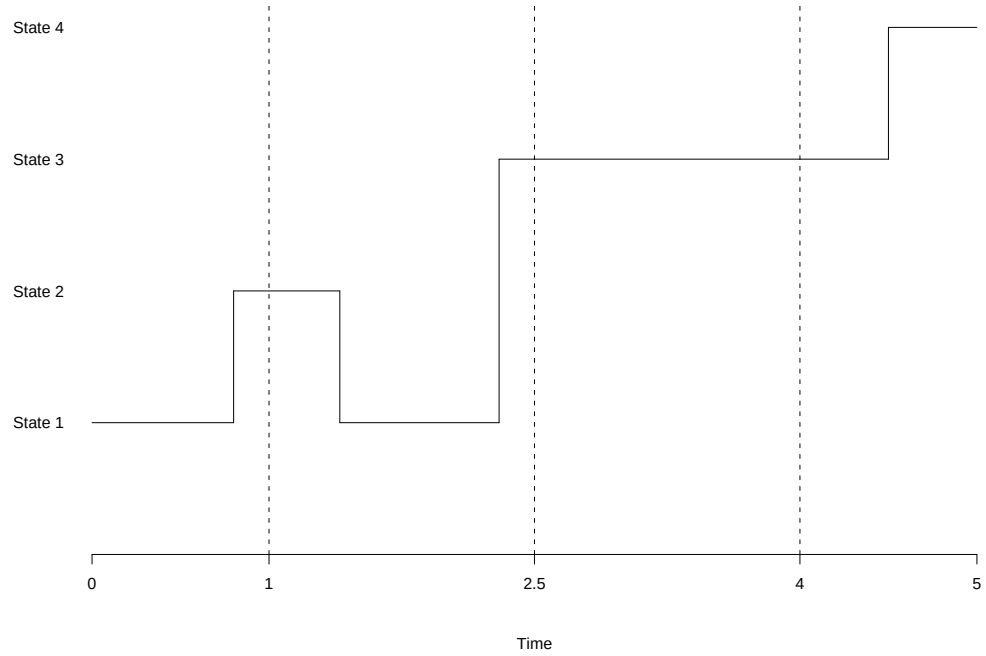
Below we introduce some notation and describe the standard assumptions made in multi-state models.

### 1.2.1 General Markov model

The Markov assumption is that the future evolution of the process only depends on the current state of the process. In terms of the transition intensities this implies

$$q_{rs}(t, \mathcal{F}_t) = q_{rs}(t) = \lim_{\delta t \downarrow 0} \frac{\mathbb{P}(X(t + \delta t) = s | X(t) = r)}{\delta t}$$

Figure 1.1: Example of panel observation of a multi-state process. The process is observed three times.




so that transition intensities vary with time, but do not depend on the past history of the process.

A particular advantage of a Markov model is that the likelihood for a series of discrete observations, assuming the observation scheme was uninformative, can be expressed simply as a product of transition probabilities

$$L = \prod_{i=0}^{N-1} p_{x_i,x_{i+1}}(t_i, t_{i+1}).$$

The transition probability matrix for a Markov model satisfies the forward Kolmogorov equations [34],

$$\frac{dP(t_1,t)}{dt} = P(t_1,t)Q(t), \tag{1.1}$$

subject to initial condition $P(t_1,t_1) = I$ where $P(t_1,t)$ is the matrix with $(r,s)$ entry

$p_{rs}(t_1, t)$. For most $Q(t)$, the Kolmogorov equations define a system of first order non-linear differential equations which often cannot be solved analytically.

### 1.2.2 Time homogeneous Markov models

A time homogeneous Markov model has the further property that

$$Q(t) = Q_0, \forall t$$

for some constant matrix $Q_0$. This implies that the *sojourn time* within a particular state, $r$, has an exponential distribution with rate parameter $\sum_{s \neq r} q_{rs}$ where $q_{rs}$ is the (r,s) entry of $Q_0$.

In this setting we have that the transition probabilities only depend on the interval between times $t_1$ and $t_2$ and not on $t_1$ itself. Equation 1.1 becomes

$$\frac{dP(t)}{dt} = P(t)Q_0 \tag{1.2}$$

subject to the condition $P(0) = I$. The solution to this is

$$P(t) = \exp(tQ_0) = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q_0^n. \tag{1.3}$$

The matrix exponential in this equation can be calculated using the eigen-decomposition of $Q_0$. Let $D$ be a diagonal matrix of the eigenvalues and $U$ the matrix with the corresponding eigenvectors as columns. Provided the eigenvalues are distinct, $U$ is invertible and $Q_0 = UDU^{-1}$. Then

$$\exp(tQ_0) = U \exp(tD) U^{-1}.$$

Transition probabilities can in some cases be calculated through direct integration rather than computation of matrix exponentials, for instance when subjects cannot return to a state once they have left it.

Maximum likelihood estimation of model parameters can be done by numerical optimisation. Derivative free algorithms such as Nelder-Mead [98] are available. However, in general the first derivatives are not difficult to compute, it is often advantageous to use a quasi-Newton algorithm such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method incorporating the first derivatives [120]. Kalbfleisch and Lawless [70] gave a Fisher-scoring algorithm using the first derivatives and using the expectation of the second derivatives,

which can be calculated from the likelihood and first derivatives. However this latter approach cannot be used in the common situation where time of entry into the absorbing state is known exactly. Whilst analytic expressions for the second derivatives are available [80], their computation is too time consuming to merit inclusion in the optimisation process. Nor are they necessary in the computation of the observed Fisher information as the numerical Hessian matrix can be computed as part of the Nelder-Mead or BFGS algorithm.

### 1.2.3 Semi-Markov models

The Markov assumption is restrictive and not necessarily realistic. Time homogeneous semi-Markov models assume that the trajectory of the process depends only on the amount of time spent in the current state, allowing the sojourn times in each state to have an arbitrary distribution. In terms of the transition intensities this implies

$$q_{rs}(t, \mathcal{F}_t) = q_{rs}(t_r) = \lim_{\delta t \downarrow 0} \frac{\mathbb{P}(X(t+\delta t) = s | X(t) = r, t_r)}{\delta t}$$

where $t_r < t$ is the time at which state $r$ was entered.

The likelihood for a time homogeneous semi-Markov model for panel observed data is difficult to evaluate since the time of entry into the current state is generally unknown. The Markov property does not apply so all observations for a subject have to be evaluated together. The transition probabilities between observed states, must be indexed $p_{rs}(t, u)$ where now $t$ denotes the time between observations and $u$ represents the time spent in the current state. For progressive models, where a subject cannot re-enter a state once it has left it, these transition probabilities can be calculated by considering all possible paths between the states and calculating the probability of each path. For bi-directional models, the number of potential paths is infinite and analytic solutions to the transition probabilities do not exist in general.

Further generalisations of the semi-Markov model can be made. An inhomogeneous semi-Markov model allows the transition intensities to depend on time since initiation of the process as well as the time elapsed since entry into the current state. When the transition times are interval censored and data on precise transitions are missing, there is unlikely to be sufficient information to allow non-homogeneous semi-Markov models to be fitted.

### 1.2.4 Hidden Markov models

A further reason for the invalidity of the Markov property on the observed states may be the presence of state misclassification.

A hidden Markov model (HMM) consists of an underlying unobserved Markov process $X(t)$. The observed response $O(t_1), \ldots, O(t_N)$ at sampling times $t_1, \ldots, t_N$ is given through some error distribution $e(r, x)$ such that

$$\mathbb{P}(O(t_i) = x | X(t_i) = r) = e(r, x).$$

Moreover it is usually assumed that the response only depends on $X(t_i)$ and conditional on

$$X(t_1), \ldots, X(t_N),$$

the responses $O(t_1), \ldots, O(t_N)$ are independent. In this thesis we will concentrate on the special case where the response takes a discrete set of values. Typically this implies that the $e(r, x)$ define multinomial cell probabilities. This is a *misclassification* hidden Markov model if the discrete set of values is the same as the state space of $X(t)$. The response probabilities $e(r, s)$ can then be thought of as misclassification probabilities.

Observations from a hidden Markov model do not obey the Markov property. This makes computation of the likelihood more difficult. The likelihood for a subject observed at times $t_1, \ldots, t_N$ is given by

$$\begin{aligned} L &= \mathbb{P}(O_1, \ldots, O_N) \\ &= \sum \mathbb{P}(O_1, \ldots, O_N | X_1, \ldots, X_N) \mathbb{P}(X_1, \ldots, X_N) \end{aligned}$$

where the sum is over all possible states $X_1, \ldots, X_N$. However, the Markov property in the underlying states and the conditional independence of $\{O_1, \ldots, O_N\}$ can be exploited so that the likelihood can be expressed as

$$L = \sum_{X_1} \mathbb{P}(O_1|X_1)\mathbb{P}(X_1) \sum_{X_2} \mathbb{P}(O_2|X_2)\mathbb{P}(X_2|X_1) \ldots \sum_{X_N} \mathbb{P}(O_N|X_N)\mathbb{P}(X_N|X_{N-1}).$$

For a misclassification HMM, $\mathbb{P}(O_i = s | X_i = r) = e_{rs}$ and $\mathbb{P}(X_i = s | X_{i-1} = r) = p_{rs}(t_i - t_{i-1})$. If we therefore define matrices $M_1 \ldots M_N$ where $M_i$ is a $R \times R$ matrix with $(r, s)$ entry

$$e_{s,O_i} p_{rs}(t_i - t_{i-1})$$

and $t_0 = 0$, then the likelihood can be written as a matrix product

$$L = \pi_{\mathbf{0}} M_1 M_3 \dots M_N \mathbf{1} \tag{1.4}$$

where $\pi_{\mathbf{0}}$ is the vector of initial state probabilities and $\mathbf{1}$ is a vector of ones of length $R$. $\pi_{\mathbf{0}}$ may not be known and can instead be included as unknown parameters [14]. Alternatively, in recurrent models, it is sometimes assumed that the Markov process is in equilibrium.

The likelihood can be maximised numerically using derivative free optimisation algorithms. This approach has the advantage of giving the Fisher information of the observed likelihood through the numerical Hessian. However, such derivative free algorithms may require many evaluations of the likelihood in order to converge to the optimum, particularly if the choice of starting value is poor. The most common method of maximising the likelihood among discrete-time, hidden Markov models is the Baum-Welch or Forward-Backward algorithm [12]. This is a type of Expectation-Maximisation algorithm. The algorithm is easily extended to continuous-time Markov models [14]. The EM algorithm tends to be faster at reaching a neighbourhood of the optimum. However, full convergence may be slower. Moreover, under the EM algorithm approach it is more difficult to obtain the observed Fisher information.

## 1.3 Particulars of multi-state modelling

There are particular features of data or model structure that can have significant influences on the ways in which model appropriateness can be assessed. Most methods only work in a subset of these cases. This section serves to clarify the terminology which will be used throughout.

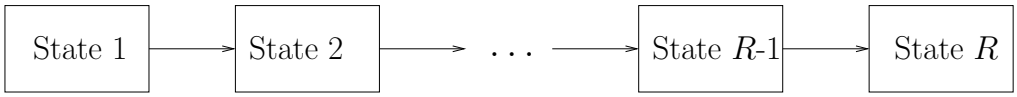
### 1.3.1 Model structure

The types of transitions between states that are allowed by a model have implications for inference. The main features of a multi-state model structure that affect statistical modelling are summarised here.

#### Unidirectional models

*Unidirectional* models consist of one simple chain of states. Subjects begin in state 1 and can only progress through the states sequentially until an absorbing state $R$ (figure 1.2.

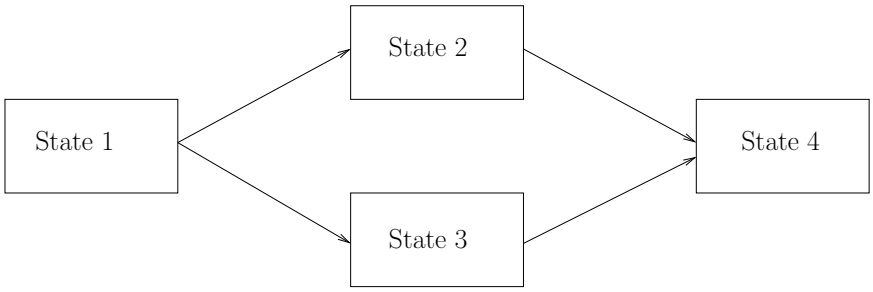
Figure 1.2: Example of a unidirectional model




The survival model, in which state 1 is living and state 2 is dead, is the simplest example of a unidirectional model.

### Progressive models

Figure 1.3: Example of a progressive model

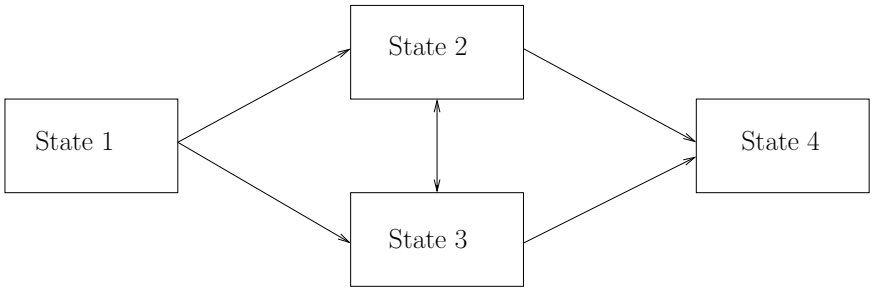



Unidirectional models are a simple example of the wider class of *progressive* models. Progressive models take the form of a directed acyclic graph. They can allow a choice of ways out of a state, but once a subject has left a state, it cannot return. Figure 1.3 depicts an example of a progressive model. A common example of a progressive model is the three state chronic disease model where subjects begin in state 1 (healthy) from which they can progress to either state 2 (diseased) or state 3 (death). From state 2 they may only progress to state 3.

### Bi-directional models

*Bi-directional* models contain an absorbing state but can allow transitions in either direction between some of the transient states. An example of this is a three state disease model where subjects may recover from disease. Some authors refer to these models as being reversible. It is possible however, to get this confused with the quite different concept of time reversibility in Markov chains. A stationary Markov chain $\{X(t) : -\infty < t < \infty\}$ is said to be time reversible if the reversed process, $Y(t) = X(-t)$, is the same stochastic

Figure 1.4: Example of a bi-directional model. Cycles between states 2 and 3 are possible.




process (i.e. it has the same transition intensities as $X(t)$). A Markov chain is stationary if it is in its equilibrium or stationary distribution. The stationary distribution, $\pi$, of a time homogeneous continuous time Markov chain satisfies the equation

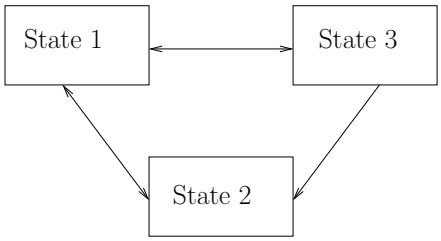
$$\pi = \pi P(t)$$

for any $t \geq 0$ and satisfies the equation

$$\pi Q = 0.$$

**Recurrent models**

Figure 1.5: Example of a recurrent model




*Recurrent* models do not have an absorbing state, and include states which are *recurrent*, in the sense that the probability the process will eventually return to a state is 1. Figure 1.5 depicts a recurrent model. The simplest example of a recurrent model is the two state illness-recovery model where state 1 represents healthy and state 2 represents illness.

Unidirectional models are the easiest of the above to implement. This is because, at worst, the transition times of the process are interval censored. The uncertainty surrounding the

true state trajectory in progressive models presents more difficulties. Recurrent and bi-directional models present the greatest challenges for implementation. Estimates of the transition intensities between states that are connected in both directions are particularly difficult to obtain and are heavily dependent on assumptions about the process (i.e. the Markov property or time homogeneity). For recurrent models, one may often be able to assume the process is stationary (i.e. subjects begin in the stationary distribution of the process) which can make implementation easier.

### 1.3.2 Observation scheme

The sampling scheme from which panel observed data arise can have a great impact on model fit and assessment.

*Balanced observation*: All subjects are observed at a set series of times, $t_1, ..., t_n$, the simplest case is *regular balanced observation* where the times are, $t, 2t, ..., nt$ so that all time intervals between observations are of length $t$. This observation scheme would be appropriate for experimental studies such as randomised controlled trials.

*Irregular observation*: Subject $i$ has their own set of observation times $t_{1i}, ..., t_{n_i i}$. Such an observation scheme may arise if subjects missed scheduled observation times or if observations arose from irregular clinic visits.

#### Joint modelling of survival and disease screening

It is typically assumed that the observation scheme is independent of the underlying process. One important exception is death (or another absorbing event) for which observation only takes place at that time because a death has occurred. In these situations the exact transition time to death will be observed, but the state occupied directly preceding death will not be known.

For a time homogeneous Markov process, the likelihood contribution of a death (state $R$) observed at time $t$ after an observation in state $r$ is given by

$$\sum_{k}^{R-1} p_{rk}(t) q_{kR}. \tag{1.5}$$

The inclusion of death as an absorbing state may also result in censored observations. When mortality of subjects is followed-up until the end of study (administrative censoring)

there will typically be a gap between the last observation time at which a subject's state could be observed, $t_N$ and the last time at which it could be known that they had died, $t_E$. At the end of the study it is known whether a subject is still alive but not their precise state if alive. For a time homogeneous Markov process, these censored observations have likelihood contribution

$$\sum_{s \neq R} p_{rs}(t_E - t_N) = 1 - p_{rR}(t_E - t_N)$$

where $R$ is the absorbing state and $r$ is the previous observed state.

### 1.3.3 Covariates

Variables associated with transition intensities are commonly assumed to have a multiplicative effect of the form

$$q_{rs} = q_{rs}^{(0)}(t) \exp\left(\beta_{rs}^T \mathbf{z}\right)$$

where $q_{rs}^{(0)}(t)$ is the baseline intensity at time $t$, $\beta_{\mathbf{rs}}$ is the covariate effect vector and $\mathbf{z}$ the covariate vector. For a time homogeneous process $q_{rs}^{(0)}(t) = q_{rs}^{(0)}$. Each transition intensity can have a separate set of covariate effects. Sometimes covariates vary with time. If this time variation is deterministic, for instance age, the resultant process is a time inhomogeneous Markov model, even if the baseline intensities are not dependent on time. The transition intensities could be written as

$$q_{rs}(t) = q_{rs}^{(0)} \exp\left(\beta_{rs}^T \mathbf{z}(\mathbf{t})\right)$$

but as with other time inhomogeneous models, the transition probabilities will be difficult to compute. Often time dependent covariates are not deterministic. In this situation the covariate status of the subject will usually only be observed at the same time points as the process. Most approaches to this problem have been to assume the covariate stays constant between observations [77], so that

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

For misclassification hidden Markov models, in addition to covariates on the transition intensities, it is also possible to have covariates on misclassification probabilities. The

usual parametrisation is the standard multinomial logit model such that

$$e_{r1}(z_i) = \frac{1}{\sum_{k=2}^{R} \exp(z_i \beta_{rk})}$$

and

$$e_{rs}(z_i) = \frac{\exp(z_i \beta_{rs})}{\sum_{k=2}^{R} \exp(z_i \beta_{rk})}, \quad s > 1.$$

Time dependent covariates affecting the misclassification probabilities are not problematic because it is only the value of the covariate at the observation time that is relevant.

## 1.4 Literature review of techniques for assessing model fit

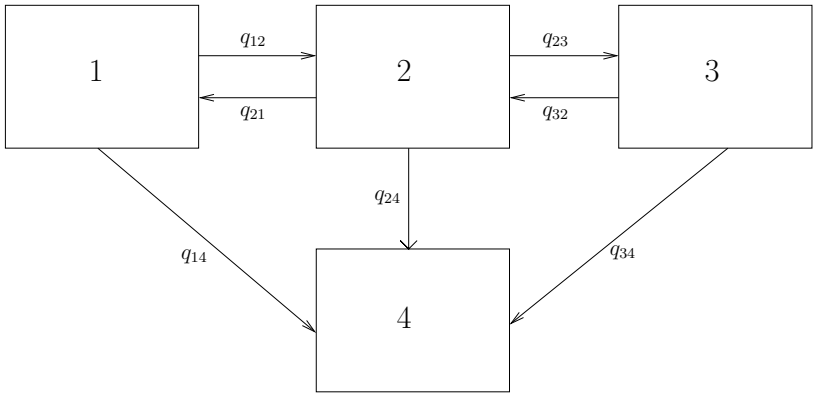
Assessing the validity of a time homogeneous Markov or hidden Markov model should incorporate a range of techniques. The fit of the model can be assessed by testing each of the specific assumptions of the model individually and by general goodness-of-fit tests. Irregular sampling times, continuous covariates and exact death times all present additional challenges. Hidden Markov models where the observed data have misclassification error have additional assumptions that need to be tested. Moreover the structure of the data is different so alternative methods are required.

### 1.4.1 The Markov assumption

The Markov assumption, that the future evolution of the process depends only on the current state and not past history, is key to many analyses, and, even when not strictly appropriate it can provide a base case analysis against which to assess other models. However, it is difficult to test the assumption explicitly for panel observed data and little methodology has been developed to test it. A method suggested by Kay [73] involves creating data for the exact transition times between states using interpolation. A test can then be performed on this completed data. For instance, consider a disease model where death is the absorbing state and which includes state 1 and state 2 and transitions between them are possible in both directions (figure 1.6). Let $x$ be the time spent in state 2 during last sojourn from state 1. We can fit a model where the intensity $q_{12}$ is given by $\lambda_0 \exp(\beta x)$ and test $H_0 : \beta = 0$. This would assess the assumption that the transition rate

to death from 1 is unaffected by the previous sojourn time. However, the accuracy of any conclusions from this test depend on the accuracy with which the exact transition times can be determined through interpolation, i.e. where observations are frequent relative to transition times. Application of this test has tended to be in cases where complete observation can be assumed [103].

Figure 1.6: 4 state model used by Kay and in Grüger *et al*




Healy and De Gruttola [59], working with a progressive model, compared the time to next transition for subjects who were in a particular state for two consecutive measurements with the same time for those who had just jumped from a previous state. They used a log-rank test to assess whether there was a significant difference between the two groups. Essentially this tests whether the particular state has a non-exponential distribution. This method only tests one part of the model and only a subset of the data contributes to the analysis. The method is also only applicable to progressive models. The standard log-rank test [107] should also only be applied to right-censored data. If it is only possible to leave the state being tested by entering an absorbing state (e.g. death), whose time of entry is known exactly, then the test will involve right-censored data. More generally the data will be interval-censored and an alternative test would be required.

Foulkes and De Gruttola [48] used logistic regression to consider whether any change in state between the first two observation points was an indicator for any change in state between the second and third observation points. Regular or nearly regular time intervals between observations are required for this approach to be effective.

### 1.4.2 Homogeneity of transition intensities through time

A key characteristic of time homogeneous Markov models is that the transition intensities remain constant through time. This assumption can be tested using piecewise constant transition intensities, as originally used by Faddy [45]. A formal likelihood ratio test of the two models can be used as a test for time independence. As an alternative Kalbfleisch and Lawless [70] suggest fitting the parametric time-dependent model: $q_{rs}(t) = q_{rs} \exp(-\lambda t)$ and perform a likelihood ratio test on the hypothesis that $\lambda = 0$. More generally they show that any process $X(t)$ with time dependent intensity matrix $Q(t) = Q_0 g(t;\lambda)$ can be fitted for any non-negative function $g(t;\lambda)$. By taking

$$s = \int_0^t g(u;\lambda)du$$

we can get a process $X(s)$ which is time homogeneous with intensity matrix $Q_0$. Both tests based on piecewise constant intensities, and those based on a functional dependence on time, require some choice to be made in order to be performed. For piecewise constant intensities, the number and location of the change points must be determined. For functional dependence, the function $g(t;\lambda)$, must be specified. Piecewise constant intensities allow a more general alternative model and may therefore be more powerful in detecting departures from time homogeneity in many cases, although the effectiveness for a particular dataset may be heavily dependent on the change points chosen.

De Stavola [128] presented a method for testing for local departures from homogeneity by the use of local score tests. The particular alternative to be tested was a process with time dependent transition intensities given by

$$q_{rs}(t) \approx q_{rs} + \epsilon t.$$

where $\epsilon$ is small. The conditional probabilities for this alternative can be found by solving the forward Kolmogorov equations, but this is not a straightforward task. An approximate power series solution, in powers of $\epsilon$ can be used, where it is assumed terms of order $\epsilon^2$ or above can be discarded. A score test for $\epsilon = 0$ can then be applied. This approach was also advocated by Gentleman, Lawless, Lindsey and Yan [55] who extended the use of local score tests to time dependent intensities of the form

$$q_{rs}(t) \approx q_{rs} t^{\beta-1}$$

where in this case the power series solution is expanded about $\beta = 1$. The advantage of these score tests over piecewise constant hazards is that only the time homogeneous model needs to be fitted.

### 1.4.3 Comparison with non-parametric estimates

In cases where the model has an absorbing state for which the time of entry is known exactly (and time of initiation of the process) Kaplan-Meier product limit estimates of the survival function can be compared to survival estimated from the fitted Markov models. A large degree of general disagreement between the two curves can be taken as informal evidence against the Markov model. Comparison of Markov model-fitted and Kaplan-Meier survival is a commonly used technique [55, 78, 85]. Gentleman *et al* put 95% piecewise confidence intervals on the fitted survival curve for the Markov model to allow a slightly more formal assessment. In some cases approximate hypothesis tests have been applied to the observed discrepancies, either the Hollander-Proschan test [60] as in papers by Pérez-Ocón *et al* [102, 105, 106] or a likelihood ratio style test detailed in Lawless (1982) [83] again by Pérez-Ocón *et al* [104]. Kaplan-Meier estimates are only valid for the assumption of homogeneous subjects. Siannis *et al* fitted a Cox proportional hazards model [35] to the survival data in the presence of covariates. Each set of covariate values will give a different pair of estimated curves, so it is necessary to consider the fit for a selected range of values. In chapter 2 of this thesis the analysis of survival curves and the use of formal tests based on them will be discussed further.

Neither the individual sojourn time distributions within each transitional state, nor the times to absorption, can be assessed using Kaplan-Meier plots because the event times aren't generally known, nor can the times to absorption if they are interval censored. Longini *et al* [90] discussed the possibility of comparing the Markov model with non-parametric alternatives. Sophisticated methods of non-parametric estimation using the principle of *self-consistent estimators* [132] have been used to provide plots that the time homogeneous Markov model can be considered against.

Self-consistent algorithms are a class of Expectation-Maximisation algorithms. The technique involves identifying the finite set of time points for which the transition intensities under the maximised non-parametric estimate can have positive values. It is then shown that under the maximised likelihood these points satisfy a set of equations. For instance such equations may be of the form

$$q_{rs}(t_i)\mathbb{E}\{\#\ \text{in state}\ \ r\ \ \text{at}\ \ t_i - \ |\mathcal{D}\} = \mathbb{E}\{\#\ \text{transitions from}\ \ r\ \ \text{to}\ \ s\ \ \text{at}\ \ t_i|\mathcal{D}\}$$

where $q_{rs}(t_i)$ is the estimated transition intensity at time $t_i$ and $\mathcal{D}$ represents the data. The algorithm begins with an arbitrary set of estimates for $Q(t)$ and proceeds iteratively by calculating the expectations based upon the estimated parameters and the data, and

updating the estimates of $Q(t)$ based on this.

However, the range of models for which self-consistent algorithms have been applied is limited. Frydman developed methods for non-parametric estimation of non-homogeneous Markov models, in the case of a three-stage unidirectional model [49] and a three-state disease model in which the state from which death was entered was known [50]. It is more typical for the state immediately before death to be unknown. Approximate non-parametric estimation for a time inhomogeneous Markov model was achieved by Gaüzère *et al* [52]. This used imputation rules in order to impute the times of transition and whether or not a transition occurred. The problem then becomes estimation with right-censored data so that Nelson-Aalen or Kaplan-Meier estimates can be used. Subsequent work by Gaüzère [53] and Frydman and Szarek [51] has extended the self-consistency method to allow the exact non-parametric mle to be computed when the state immediately before death is unknown. However there are no developed methods for other model patterns or semi-parametric models for covariates.

Some methods for non-parametric estimation for semi-Markov models exist. De Gruttola and Lagakos fit a three-state unidirectional model with application to AIDS induction time [40], Satten and Sternberg provide a more general method for unidirectional models with an arbitrary number of states and allow for unknown initiation times [116, 129]. However these methods for semi-Markov models involve choosing an arbitrary discrete set of time points, at which transitions can take place, effectively requiring the process to take place in discrete time. Datta and Sundaram [38] developed a method for non-parametric estimation of state occupation probabilities but only in the special case of *current status data* - this is when disease status is only known at the time of initiation and the time of censoring.

Penalised likelihood methods for multi-state models [25, 26] are closely related to the non-parametric approach. It is argued that a biological process is unlikely to have rapid changes in hazard. Therefore, the transition intensities are estimated by maximising a penalised likelihood function

$$pl(q(.)) = l(q(.)) - \sum_{r,s} \kappa_{r,s} \int_0^T \frac{\partial^2 q_{rs}(u)}{\partial t^2} du$$

where $q(.)$ is the set of intensity functions, which vary with time $t$, $T$ is last observed time in the data and $\kappa_{rs}$ is the constant that determines the level of smoothing for the transition intensity relating to transitions between states $r$ and $s$. In practice, it is not possible to obtain functions which exactly maximise the penalised likelihood. Instead spline functions

are used to estimate the transition intensities. Penalised likelihood methods have been applied to both time inhomogeneous Markov and semi-Markov models [68, 69].

### 1.4.4 Contingency table based methods

Contingency table methods provide an assessment of the overall fit of the assumed model. Kalbfleisch and Lawless [70] dealt with data in which there was balanced observation. In addition covariates were categorical. In this setting, model fit can be assessed by considering observed and expected transition frequencies, either through a likelihood ratio test or the asymptotically equivalent Pearson $\chi^2$ statistic. De Stavola [128] applied the Pearson $\chi^2$ form of the test to breast cancer data. Chan and Muñoz-Hernández [17] gave a rigorous proof of the asymptotic null distribution.

Chen and Sen [21] argued that Pearson chi-square tests have low power, particularly when the degrees of freedom are large and that the asymptotic null distribution cannot be applied when counts in parts of the table are small. They therefore proposed as an alternative the use of a Cochran-Mantel-Haenszel statistic. However, the null distribution they gave for their statistic is only appropriate for testing a fully specified model and is therefore of limited practical use.

Aguirre-Hernández and Farewell [6] presented what was essentially an extension of Kalbfleish and Lawless's Pearson chi-square method, to cope with the common situation of irregular observation times and continuous covariates. Simulations suggest that the null distribution of this statistic is reasonably well approximated by the analogous $\chi^2_{d-p}$ distribution when there are no continuous covariates and has a slightly inflated mean when continuous covariates are present. The Aguirre-Hernández and Farewell test can allow testing in quite a general range of situations. Chapter 3 of the thesis is devoted to extensions of this method to an even wider range of cases.

Gentleman *et al* [55] suggest prevalence or transition counts as an approach to assessing overall goodness-of-fit when subjects have irregular and unique observation times, so that a Pearson-chi square test is not possible. Prevalence counts involve comparing the overall *state occupancies* at a fixed set of times with those expected by the fitted model, whilst transition counts involve comparing the observed number of *transitions* between states between fixed time points with those expected by the model. Unfortunately, if observation times are indeed irregular, the observed counts will not be available for a particular subject. Therefore some interpolation is necessary to construct the counts. Gentleman *et al*

suggest to simply assume that a subject has remained in the same state as their previous observed state. They state that provided subjects are observed sufficiently frequently any bias should be minimal. The fit of prevalence or transition counts can be assessed using statistics comparing observed against expected such as the chi-squared statistic, but these statistics will not have a $\chi^2$ null distribution. The use of prevalence counts is common [95] and a full discussion on their use is in chapter 2.

To some extent contingency table based goodness-of-fit tests can provide some power in testing the Markov assumption, but this power will be dependent on the particular groupings chosen in the table. If observations are grouped by more of the previous history than just the previous state, more power for testing the Markov assumption should be obtained. Such methods have not been used in the literature to solely test the validity of a Markov model. Bureau *et al* used them to consider the fit of a misclassification hidden Markov model. Kang *et al* [72] categorised observations by the complete pattern of observed states when comparing the fit of a Markov to a semi-Markov model.

**Residual based measures**

Kosorok and Chao [79] proposed the use of *summary residuals* for each observation. These quantities consider the expected state probabilities of an observation based upon the last observation and covariates. The basic idea is to consider the categorical observed state e.g. $1, ..., N$ as a random variable with mean and variance that are functions of the transition intensities, and then make the appropriate transformation to give a random variable with mean 0 and variance 1. If the model is correct and for a fully specified parameter vector, the summary residuals will have this mean and variance and moreover they will be uncorrelated. The authors assert that if the model is correct, these properties will also approximately hold when using a parameter vector, $\hat{\theta}$, fitted from the data. The residuals can be plotted against quantities such as time in study to assess time homogeneity, or against covariates of interest to assess whether the covariate model (e.g. assumptions of log-linear effects) is an adequate fit. The use of residual plots does not seem to have been adopted by subsequent authors. Their strength in assessing the covariate model make them potentially useful. In Chapter 2 they are considered in more detail.

Chen and Sen [21] adopted a related model validation procedure. Given the state $X_j(t_{i-1}) =$

$r$ observed at time $t_{i-1}$ for subject $j$, a binary random variable is defined as follows

$$\delta_{ij} = \begin{cases} 1 & X_j(t_i) = k_i \\ 0 & \text{otherwise} \end{cases}$$

where

$$k_i = \{s : p_{rs}(t_{i-1}, t_i) = max_m \ p_{rm}(t_{i-1}, t_i)\}.$$

So $\delta_{ij}$ is 1 if the observed state equals the most likely state and 0 otherwise. Summing the $\delta_{ij}$ for all intervals and subjects gives a validation score for the model

$$C = \sum_j^N \sum_i^{n_j} \delta_{ij}.$$

Intuitively, a large value of $C$ is an indication of good model fit. By defining

$$Z = \frac{(C - \mu)}{V}$$

where $\mu = \sum_j^N \sum_i^{n_j} \theta_{ij}$ and $V^2 = \sum_j^N \sum_i^{n_j} \theta_{ij}(1 - \theta_{ij})$ and $\theta_{ij} = \mathbb{P}(X_j(t_i) = k_i | X_j(t_{i-1}))$, it can be shown that for a fully specified model $Z \to \mathrm{N}(0, 1)$, and this could be used as the basis of a goodness-of-fit test. However, the distribution of $Z$ is not standard normal when the model parameters are estimated from the data. As there is no penalty for the number of parameters in the model, using this criterion would tend to favour over-fitted models. The same authors presented a variation on this approach in a subsequent paper [22].

### 1.4.5 Homogeneity of parameters across the subject population

The simplest test of the assumption of homogeneity of parameters across the subject population is to model (additional) covariate effects where they are available [73]. In particular we can let the transition intensities be functions of covariates

$$q_{rs}(z) = q_{rs} \exp(\beta^T z)$$

Often however, lack of subject homogeneity may be due to unobserved data, covariates or individual frailties. A simple case is where the assumption that all subjects are in the same state at time zero is relaxed and instead there is an initial mixture of state occupation probabilities [136].

Satten [117] introduced a model for *tracking*, a form of random effects where each subject has an individual frailty that acts as a multiplier on their entire intensity matrix so that subjects who progress quickly through one state are also likely to progress quickly through the others. A likelihood ratio test can be used to test 'tracking' against the standard Markov model. This procedure is covered further in chapter 2.

More sophisticated random effects models are difficult to fit except in special cases such as the two state recurrent 'disease / disease-free' model [28]. Cook, Yi and Lee [30] gave a procedure for fitting models with a multivariate log-normal random effects distribution on the parameter vector. The ideal formulation would have each subject's individual transition intensities governed by a random effects vector $G_i$,

$$\log(G_i) \sim \mathrm{MVN}(0, \Sigma),$$

where $\Sigma$ is the unknown covariance matrix. Each component of $G_i$ then corresponds to the multiplicative factor on one of the transition intensities. This allows subjects to have more sophisticated correlations between their sojourn times than in the case of tracking. For instance $\Sigma$ may be such that subjects who progress rapidly through the first state may have slower progression through a subsequent state.

In order to calculate the likelihood for this model, integrals of the form

$$\int_{\mathcal{U}} \left( \prod_{j=0}^{N-1} p_{x_j, x_{j+1}}(t_{j+1} - t_j; u) \right) g(u) du \tag{1.6}$$

where $u$ is an individual realisation from the random effects vector, $g(u)$ is the multivariate log-normal density and $x_0, \ldots, x_N$ are the observed states at times $t_0, \ldots, t_N$ for a subject who was observed $N$ times. This integral is intractable because each value of $u$ implies a different eigen-decomposition of the intensity matrix. The integral is over the whole space of $u$ so will be multi-dimensional. Evaluating such integrals by quadrature or Monte Carlo techniques is likely to be too time consuming for practical use.

Instead, Cook *et al* rely on approximating the multivariate log-normal for $G_i$ by a discrete distribution on multiple points. This allows equation (1.6) to be expressed as a sum

$$\sum_{m} \left( \prod_{j=0}^{N-1} p_{x_j, x_{j+1}}(t_{j+1} - t_j; u_m) \right) g_m \tag{1.7}$$

where $g_m$ represents the probability of the point $u_m$.

While this is far more effective than attempting to evaluate the integrals directly, it is still computationally intensive. In Cook *et al*'s application, 64 points were used in the discrete distribution. This means calculating the likelihood for one set of parameters is around 64 times slower than the analogous model without random effects.

*Mixed* Markov models, where the subjects' heterogeneity is defined through differences in the actual pattern of their underlying transition intensity matrices, are another possible alternative. The basic case is mover-stayer models [18] in which an unknown proportion of the sample are not at risk of developing disease. Cook and Kalbfleisch generalised the mover-stayer model to allow the progression of any particular subject to be limited to a particular state in a progressive model [29].

For all these alternative random effects models, the likelihood can be calculated. Thus we can test the basic assumptions of the homogeneous Markov model against an alternative using a likelihood ratio test, comparing the respective likelihoods in the two models.

### 1.4.6 Misclassification hidden Markov models

In misclassification HMMs, the observed data no longer satisfy the Markov property. This means that some of the techniques used in Markov models cannot be immediately transferred over. The basic concepts such as considering observed and expected quantities are still valid but it is necessary to modify the methods.

Satten and Longini [115] took the 'ability of the model to predict the next observation' as a criterion for model fit. This is analogous to considering observed and expected transitions in a Markov model except that now, it is necessary to consider all observations of a subject up to the time of interest, in order to calculate the probabilities. Since in a misclassification HMM the previously observed state might be highly unrepresentative of the state of the underlying process (e.g. if we observe a 1 after a long sequence of 2s in a progressive model), the authors did not categorise subjects by last observed state. Instead they constructed a table that is in some respects analogous to the idea of prevalence counts. However, instead of categorising by the occupancy at a particular time, they instead considered the average observed occupancy for each subject in particular time intervals. This approach was used by Jackson and Sharples [65]. The method is also applicable to more general hidden Markov models where the observations are continuous.

Bureau and Shiboski [14] were interested in recurrent binary outcomes and provided two novel approaches to assessing model fit. They considered observed and expected counts based not just on the last observed state, but on the sequence of previous states. This allows more thorough testing of the assumptions of independent misclassification that is a major feature of the model. However, this approach may also be applicable for Markov models in further testing issues like the Markov property. Chi-squared style deviances can be calculated on the resultant contingency table, but they cannot be compared to known null distributions.

Their second method is a graphical approach that requires simulation. The idea is to try to estimate the empirical distribution of the sojourn times in the observed process, and compare the resultant Kaplan-Meier estimates with equivalent estimates from simulated data. This is a somewhat obscure concept because the observed process can only be defined at the discrete observation times. Therefore the shape of the Kaplan-Meier estimate will be heavily dependent on the sampling distribution. This type of plot is discussed in more detail in section 2.6.2.

### 1.4.7 Literature for fitting time dependent models

An effective method of assessing the fit of a time homogeneous Markov model is to fit an alternative model allowing some kind of time dependency in the transition intensities. The resulting likelihood ratio statistic can be used to test the assumption of time homogeneity. Moreover, if the fit of a time homogeneous model is poor, one natural step is to seek a time dependent model to better explain the data. Methods of fitting time dependent multi-state models are therefore highly relevant to model assessment.

#### Time-inhomogeneous Markov models

The most common approach to fitting a time dependent model is to use a Markov model with piecewise constant intensities [45, 114]. For this model, the likelihood is only slightly harder to calculate than a time homogeneous model. However a particular drawback of this approach is that a choice has to be made of where the change points in the hazards should occur. We may also need to choose the number of change points. Various reasons for the choice of change points for such models have been given. These include clinical reasons [123], reference to the overall empirical estimate of the hazard of death [105], or to ensure that roughly equal numbers of observations occur in each region [73]. Mathieu

*et al* [95] constrained themselves to two time periods and then considered which of a range of potential cut-off times gave the best likelihood. Ocaña-Riola *et al* [101] present an algorithm based upon starting with a maximum number of intervals and successively merging the intervals until no further merging can occur without a significant deterioration in the likelihood. A full account of piecewise-constant hazard intensities is given in chapter 5.

An approach by Chen, Bernard and Sen [20] is analogous to piecewise constant intensities. Their approach is only really applicable in situations where there is a common set of potential sampling times for all subjects, $t_1, ..., t_n$, but that individual subjects miss some of these observation times. No transition intensities are estimated directly. Instead, for each interval $(t_i, t_{i+1}]$ the transition probability matrix $P(t_i, t_{i+1})$ is estimated using multinomial regression, covariates can be accommodated by assuming linearity via a logit link function. The likelihood contribution for subjects who miss a series of observations between $t_i$ and $t_{i+r}$ is just

$$\prod_{s=0}^{r-1} P(t_{i+s}, t_{i+s+1}).$$

This method becomes infeasible if there are many potential sampling times as the number of parameters to be estimated becomes too large.

There are relatively few examples in the literature of time inhomogeneous Markov models with continuous transition intensities fitted to panel observed data. This is primarily a reflection of their general difficulty. If the transition intensity matrix $Q(t)$ is time dependent then the Kolmogorov forward equations

$$\frac{dP(t_1, t_2)}{dt} = P(t_1, t_2)Q(t_2)$$

are a system of non-linear differential equations, and have no closed form solution, except in the case noted by Kalbfleisch and Lawless [70], and referred to in section 1.4.2, where $Q(t) = Q_0 g(t)$, for some function $g(t)$ which is monotonically increasing with t, so that there exists an operational time for which the differential equation is linear.

This result can be applied to the case of a Markov model with Weibull transition intensities provided the shape parameter is the same for each intensity [99].

If each transition intensity is of the form

$$q_{rs}(t) = \alpha\lambda_{rs}(\lambda_{rs}t)^{\alpha-1},$$

where $\alpha$ is the common Weibull shape parameter and $\lambda_{rs}$ is the transition specific rate parameter, then in the notation above we can take the $(r, s)$ entry of $Q_0$ to be $\lambda_{rs}^{\alpha}$ and then

$$g(t) = \alpha t^{\alpha-1}.$$

If the Weibull shape parameters are allowed to differ, closed form expressions for the transition intensities are not available.

For progressive models, the solution to the forward equations can still be found using numerical integration. Pérez-Ocón *et al* [105] fitted a Markov process with piecewise Weibull hazard functions to a three-state disease model. In a similar vein, Hsieh *et al* [62] fitted a three-state unidirectional model and Chen *et al* [19] fitted a five-state unidirectional model, in which the first transition intensity in each model was a Weibull hazard function. In these cases, an intractable integral was only required in the calculation of one of the transition probabilities, and moreover this integral is only of one dimension. Hence it was feasible to use quadrature to numerically calculate the likelihood. For models with a greater number of states, there will be more transition probability functions that are not available in closed form, and the numerical integrals required to evaluate them will be multi-dimensional. This makes such a method of fitting rather impractical.

Anisimov *et al* [11] fitted a hidden Markov model with three recurrent states. The model included a time varying covariate and so was time inhomogeneous. This covariate was assumed constant between observations, therefore standard approaches to fitting the model could have been applied. Instead the transition probabilities were calculated by approximately solving the forward equations by numerical recursion (Euler method). In principle, this technique of solving non-linear differential equations could be used to fit more complicated time dependent Markov models. This approach will be explored in chapter 5 of this thesis.

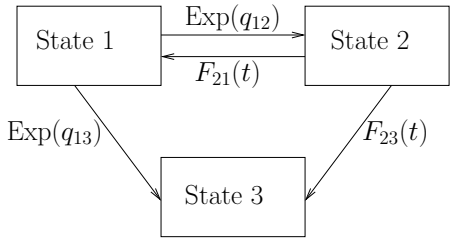
### Semi-Markov models

Semi-Markov models for panel observed data are more difficult to fit than time inhomogeneous Markov models. The Markov property no longer applies so the likelihood for a subject cannot be factorised into a product of transition probabilities between observed states. This crucial point seems to have been missed by Ruiz-Castro and Pérez-Ocón [113]. Some other attempts to fit semi-Markov models have used pseudo-likelihood [23] or

minimum chi-square estimation [109]. Lawless and Yan [84] noted that for a progressive three-state model, fitting a semi-Markov model can be feasible, for instance by numerical integration.

Having previously emphasised the lack of methods for non-Markov panel data [71], a full likelihood approach was presented by Kang and Lagakos [72] for a model with reverse transitions. The authors note that if at least one of the states has an exponential sojourn time distribution (i.e. it is Markov), the likelihood is simplified. Rather than having to consider the entire history of a subject, it is possible to factorise into a product of sojourns away from the exponential state(s). In their example they had a three-state disease model where recovery was possible and state 1 was exponential (figure 1.7).

Figure 1.7: Model used in Kang and Lagakos (2007). $F_{21}(t)$ and $F_{23}(t)$ depend on time $t$ since entry into state 2.




Transition probabilities could be expressed in terms of a sum of the number of intermediate visits to state 1. The terms in this sum can be calculated recursively as convolutions that need to be computed numerically. For a general sojourn time distribution where sojourn times arbitrarily close to zero are possible, the sum is infinite. Therefore, to allow the likelihood to be computed, the authors restrict themselves to sojourn time distribution where there is a 'guarantee time' $G$, such that the hazard of jumping is zero up to time $G$. This ensures the sums have a finite number of terms. Even with these restrictions the method would still seem to be computationally intensive.

The use of Markov Chain Monte Carlo (MCMC) methods for approximating the likelihood for non-Markov process has been outlined by Chen, Xie and Liu [24]. Stopping-time resampling methods allow importance sampling to be used. They consider a two-state model with reverse transitions and Weibull sojourn time distributions. In the example only the likelihood with respect to one unknown parameter is considered. It is not clear how practical such methods would be in multi-parameter settings.

Crespi *et al* [36] were able to fit a semi-Markov model to a two-state model with reverse transitions using full likelihood. This was achieved by assuming that the process was described by a time-homogeneous Markov birth-death process, in which state 0 corresponded to state 0 in their model, but that states 1 or greater all were observed as state 1. This allows standard methods for misclassification hidden Markov models to be applied to calculate the likelihood. In effect, the sojourn time in state 1 has been allowed to take on a phase-type distribution. A phase-type distribution is any distribution which arises from considering the time from initiation to absorption in a Markov chain. This idea was also employed by Ruiz-Castro *et al* [113] to calculate transition probabilities, $p_{ij}(t, s)$, for a semi-Markov model, although it was necessary in their case to assume entry into state $i$ at time $t$, for them to be valid. The flexibility and the relative computational simplicity of the phase-type method make it very attractive. This approach will be explored in detail and developed in chapter 6.

## 1.5 Conclusion

In summary, the existing methods for assessing goodness-of-fit in multi-state models have been limited in scope, either by being specific to only one type of model departure, by being informal, or by being only applicable to certain types of data or model. Similarly, whilst piecewise-constant hazard intensities allow time inhomogeneous Markov models to be fitted to a wide range of models, other methods for models with time dependent transition intensities are limited to particular cases. In particular, many methods require that the process is progressive.

# Chapter 2

# Informal Diagnostic Tools

This chapter discusses various methods used to assess model suitability and aims to identify the most useful. Section 2.1 introduces the two main datasets which are used to illustrate the methods in this and subsequent chapters. The methods assessed are quite varied in nature, in general they are ordered in the chapter by complexity, with the simpler first. However methods solely or primarily applicable to models without misclassification are dealt with before a separate section on methods for hidden Markov models.

## 2.1 Data

To illustrate and apply the methods used in this and subsequent chapters, two transplant related datasets are used. The first relates to screening for chronic disease in heart transplant recipients and the second to a marker of dysfunction in lung transplant recipients.

### 2.1.1 Cardiac allograft vasculopathy data

Cardiac allograft vasculopathy (CAV) is one of the main causes of death among long-term survivors of heart transplantation. The data include 596 post heart transplant patients who had their transplants between 1979 and 2000. Patients were followed up until March 2005. CAV is a chronic disease involving narrowing of the arteries (stenosis). This narrowing process is assumed to be irreversible. Accurate diagnosis of the disease can be achieved through the use of intravascular ultrasound. However, this is prohibitively expensive, therefore angiography is instead used to assess the disease. On the basis of the

angiogram patients were classified as either normal (0% stenosis), having mild CAV (up to 30% stenosis), moderate CAV (30-70% stenosis), or severe CAV (more than 70% stenosis). The design of the study evolved somewhat over time. Early transplant patients were recalled for angiography annually. However, the protocol was altered so that patients were first recalled at 2 years after transplant. If any abnormality was detected at two years, they were invited to return annually. Otherwise the next recall was scheduled to be 4 years after transplant, with annual screenings thereafter. This sampling protocol is 'doctor's care' according to Grüger *et al*'s terminology [57] and as such is non-informative. Patients' actual observation times were quite irregular. Angiograms occurred only approximately around the year markers. More significantly, there were a considerable number of missed angiograms, especially beyond the first few years after transplant. In particular, many patients had no further angiograms after their first few. Potentially this patient self-selection in observations could be problematic. If for instance the patients who stopped having angiograms did so because their status remained normal, then this could lead to some bias in inference, with progression rates being overestimated. It is assumed that such patient self-selection is not taking place in the CAV data, or if it is, only to a negligible degree. This is realistic since heart transplantation involves denervation of the heart and so CAV is essentially a silent process.

The data consist of 1972 angiograms, 563 patients had at least one angiogram, with patients undergoing up to 14 (mean 3.2, median 2). 225 of the 596 patients died within the study period. The mean follow-up time per patient was 8.6 years (median 8.2 years, range 0.3-21.6 years).

Angiography is an imperfect measure of disease. Hence the classifications are subject to error. Table 2.1 gives the observed transitions of raw state (ordered from 1=normal to 4=severe CAV) between angiograms (including from transplant where disease status is assumed to be normal). As can be seen there is a significant amount of backward transitions which shouldn't occur if the clinical hypothesis of irreversibility is true and classification is accurate. It is particularly difficult to distinguish between mild CAV (state 2) and moderate CAV (state 3) and moreover there are only 141 observations of moderate CAV. Hence it seems advantageous to combine these two states. This produces a model of the underlying disease shown in figure 2.1. A new table of transitions is shown in table 2.2.

As angiography is considered to miss stenosis rather than over diagnose, a simple approach to modelling the data involves assuming the true state at an observation is the highest state

Figure 2.1: Four state disease model for CAV data

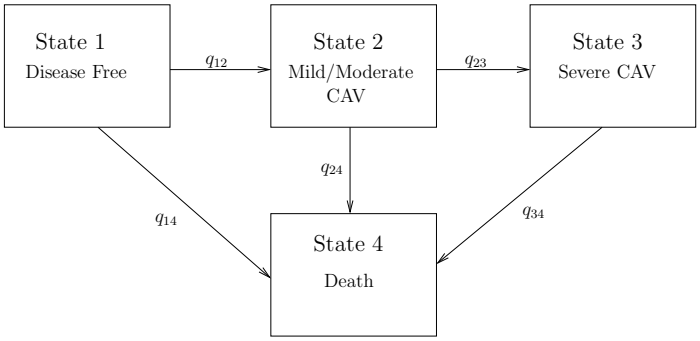



Table 2.1: Observed transitions between CAV states diagnosed at angiography

| Previous | State | | | |
|---|---|---|---|---|
| State | 1 | 2 | 3 | 4 |
| 1 | 1366 | 140 | 64 | 44 |
| 2 | 34 | 55 | 26 | 14 |
| 3 | 12 | 11 | 42 | 40 |
| 4 | 4 | 4 | 9 | 107 |

observed up to that time. However it can lead to some bias in the underlying estimates of transition intensities. Instead a misclassification HMM can be used. In this example misclassification is restricted to adjacent states.

There are various covariates which may influence the patient's progression through the states. These include the age of patient, the age of the donor, the sexes of the recipient and donor, and whether the recipient had preoperative ischemic heart disease (IHD). Sharples *et al* [124] analysed a previous version of these data and concluded that IHD and donor age were significant factors for disease onset, with onset being higher for those who had IHD and for patients with older donors. Thus these two variables will be included in our analyses.

Methods for assessing goodness-of-fit in Markov models do not tend to be universal in their applicability. As explained in the literature review in section 1.4, many are not applicable to the case of exact death times. Similarly covariates can be problematic and few methods for assessing the adequacy of the model for covariates are available. To allow

Table 2.2: Transitions between collapsed CAV states

| Previous | State | | |
|---|---|---|---|
| State | 1 | 2 | 3 |
| 1 | 1366 | 204 | 44 |
| 2 | 46 | 134 | 54 |
| 3 | 4 | 13 | 107 |

adequate examples of the methods used we shall therefore use variations of the CAV data. In particular, angiography tends to underestimate disease status. Hence to provide an example dataset for a Markov model, adjusted data will be used. These adjusted data involve assuming the true state is equal to the highest state observed up to that point.

### 2.1.2 Models for the CAV data

The models presented in this section were fitted using numerical likelihood maximisation, implemented in the **msm** [67] package in R.

#### Time homogeneous Markov model

Table 2.3 gives the parameter estimates for the models on the CAV data without misclassification, with state at observation time $t$ defined as the highest state observed up to (and including) time $t$. The intensities presented are rates per year. Two models were fitted, the first assuming patient homogeneity and the second including IHD and donor age as covariates on the transition intensity $q_{12}$, such that the transition intensity for patient $i$ is given by

$$q_{12} = q_{12}^{(0)} \exp\left(\beta_{12}^{(\mathrm{IHD})} \mathrm{IHD}_i + \beta_{12}^{(\mathrm{dage})} \mathrm{dage}_i\right).$$

Hence the baseline rates presented in table 2.3 for the model with covariates refer to a patient with donor age 0 and no IHD.

95% confidence intervals on the parameters are obtained by assuming asymptotic normality of the parameter vector with covariance matrix given by the inverse of the observed Fisher information matrix.

This shows that the death rate increases with disease severity and that once onset has

been observed progression to severe disease, and from severe disease to death, is rapid. IHD and donor age are confirmed as significant risk factors for disease onset.

Table 2.3: Model parameter estimates for CAV model without misclassification.

| Parameter | Without covariates | With covariates |
|---|---|---|
| $q_{12}$ | 0.094 (0.082,0.107) | 0.039 (0.027,0.057) |
| $q_{14}$ | 0.023 (0.017,0.030) | 0.022 (0.017,0.030) |
| $q_{23}$ | 0.200 (0.162,0.246) | 0.199 (0.162,0.246) |
| $q_{24}$ | 0.040 (0.022,0.073) | 0.041 (0.023,0.075) |
| $q_{34}$ | 0.146 (0.116,0.184) | 0.146 (0.116,0.184) |
| $\beta_{12}^{(\mathrm{IHD})}$ | | 0.446 (0.185,0.706) |
| $\beta_{12}^{(\mathrm{dage})}$ | | 0.022 (0.011,0.033) |
| -2 × LL | 3552.92 | 3524.57 |

**Time homogeneous misclassification hidden Markov model**

Table 2.4 gives the parameter estimates for the models on the CAV data with misclassification. This is consistent with the clinical hypothesis that angiography is more likely to under-estimate than over-estimate the severity of stenosis.

### 2.1.3 Bronchiolitis obliterans syndrome data

Bronchiolitis obliterans syndrome (BOS) is the irreversible, progressive airway obstruction and impairment of lung function occurring in post-lung-transplant patients. It is the major risk to long-term survival among lung transplant recipients. BOS status can only be reliably assessed histologically. In practice however, BOS is observed through the decline in lung function. The dataset involves 488 patients, of whom 242 were heart and lung transplant patients, 122 were double lung transplant patients and 124 were single lung transplant patients. This creates considerable heterogeneity within the data: patients who are given single lung transplants have a significantly poorer prognosis.

Lung function is measured using forced expiratory volume in one second ($\mathrm{FEV}_1$) measured in litres. The standard definition for stages of BOS severity is expressed in terms of decline in $\mathrm{FEV}_1$ relative to a post-transplantation baseline measure. 80% of baseline or above is

Table 2.4: Model parameter estimates for CAV model with misclassification.

| Parameter | Without covariates | With covariates |
|---|---|---|
| $q_{12}$ | 0.087 (0.073,0.102) | 0.033 (0.021,0.050) |
| $q_{14}$ | 0.021 (0.015,0.029) | 0.021 (0.015,0.029) |
| $q_{23}$ | 0.195 (0.146,0.260) | 0.190 (0.143,0.252) |
| $q_{24}$ | 0.053 (0.028,0.100) | 0.053 (0.029,0.099) |
| $q_{34}$ | 0.155 (0.120,0.201) | 0.155 (0.120,0.201) |
| $e_{12}$ | 0.027 (0.017,0.046) | 0.025 (0.015,0.042) |
| $e_{21}$ | 0.177 (0.112,0.259) | 0.186 (0.123,0.272) |
| $e_{23}$ | 0.066 (0.040,0.113) | 0.065 (0.038,0.108) |
| $e_{32}$ | 0.101 (0.050,0.191) | 0.102 (0.051,0.194) |
| $\beta_{12}^{(\mathrm{IHD})}$ | | 0.520 (0.234,0.807) |
| $\beta_{12}^{(\mathrm{dage})}$ | | 0.025 (0.013,0.037) |
| -2 × LL | 3964.46 | 3933.60 |

considered normal, 66-80% is BOS stage 1, 50-65% is BOS state 2 and below 50% is BOS stage 3. Table 2.5 gives the transitions between observed states in the data (figure 2.2). Again the process is considered irreversible but $\mathrm{FEV}_1$ is a much less specific marker than angiography.

The data include transitions from BOS stages 2 or 3 up to normal. $\mathrm{FEV}_1$ as a measure of lung function is highly sensitive to acute events (infections or rejections) and other fluctuations. Some information is available on the occurrence of infections and rejections for each patient, and it has been established that the presence of acute events is associated with acute decline in lung function. BOS is not defined until at least 6 months after transplant. Time is therefore taken from 6 months after transplant and patients are assumed to be disease free at that time.

An earlier version of this dataset has been analysed using a misclassification HMM [65]. The four classified BOS states could result in a 5 state disease model. However, there are insufficient data to reliably estimate the parameters. Hence three or four state models in which some of the intermediate states are combined are preferred. We focus on data with 4 states where state 1 represents $\mathrm{FEV}_1 \geq 80\%$, state 2 represents $65\% \leq \mathrm{FEV}_1 < 80\%$, state 3 represents $\mathrm{FEV}_1 < 65\%$ and state 4 represents death. Hence the structure of the

Table 2.5: Transitions between observed BOS states. States 0, 4 and C represent normal, dead and censored respectively.

| Previous | State | | | | | |
|---|---|---|---|---|---|---|
| State | 0 | 1 | 2 | 3 | 4 | C |
| 0 | 4244 | 497 | 99 | 25 | 73 | 139 |
| 1 | 283 | 1160 | 374 | 51 | 26 | 28 |
| 2 | 50 | 246 | 929 | 272 | 52 | 14 |
| 3 | 12 | 19 | 161 | 1654 | 140 | 16 |

Table 2.6: Transitions between collapsed observed BOS states. State C represents censored.

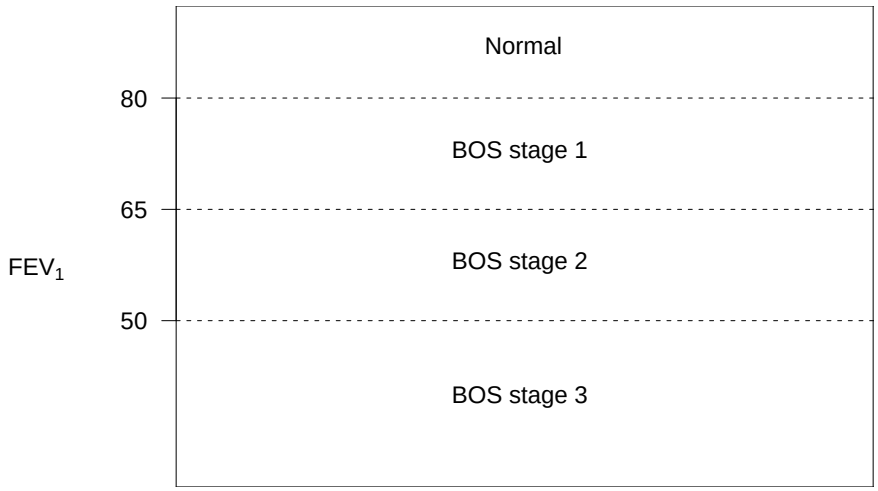
| Previous | State | | | | |
|---|---|---|---|---|---|
| State | 1 | 2 | 3 | 4 | C |
| 1 | 4244 | 497 | 124 | 73 | 139 |
| 2 | 283 | 1160 | 425 | 26 | 28 |
| 3 | 62 | 265 | 3016 | 192 | 30 |

BOS model is the same as the CAV model as shown in figure 2.1. Table 2.6 gives the number of transitions between these collapsed states.

A HMM with continuous observations has also been fitted to the raw $FEV_1$ measurements [65], but focus will be on misclassification models in this thesis and we shall not consider the raw $FEV_1$ measurements.

A particular feature of the data is the frequency of observations. Unlike the CAV data, where patients are generally observed at intervals of 1 year or more, in the BOS data the time between $FEV_1$ measurements is short, with a median of 37 days and mean of 72 days. Whilst there are 10076 transitions between observed states, there are only 488 patients and 291 deaths. The information on mortality is therefore quite limited.

Figure 2.2: Observed BOS states in terms of percentage of baseline $FEV_1$.




### 2.1.4 Models for the BOS data

#### Models without covariates on transition intensities

Four time homogeneous misclassification hidden Markov models are fitted to the data. Models 1 and 2 allow misclassification only to adjacent states, while models 3 and 4 allow misclassification also from true state 1 to observed state 3. Models 1 and 3 have no covariates, while models 2 and 4 have the presence of an acute event at the observation time as a covariate on the misclassification probabilities. The parameter estimates are given in tables 2.7 and 2.8 along with the value of -2 $\times$ the log likelihood in each case. The presence of acute events does seem to increase the probability of being misclassified to a higher state and conversely lowers the probability of being misclassified to a lower state. Allowing misclassification from state 1 to 3 also represents a very significant improvement in likelihood. Acute events were highly significant under this model as well. In each model

Table 2.7: Model parameter estimates for BOS models with adjacent misclassification

| | Model | |
|---|---|---|
| Parameter | 1 | 2 |
| $q_{12}$ | 0.313 (0.279,0.351) | 0.314 (0.280,0.352) |
| $q_{14}$ | 0.043 (0.031,0.058) | 0.042 (0.031,0.057) |
| $q_{23}$ | 0.321 (0.275,0.375) | 0.316 (0.270,0.370) |
| $q_{24}$ | 0.066 (0.043,0.102) | 0.072 (0.048,0.107) |
| $q_{34}$ | 0.328 (0.279,0.385) | 0.321 (0.273,0.377) |
| $e_{12}$ | 0.039 (0.032,0.046) | 0.319 (0.026,0.039) |
| $e_{21}$ | 0.196 (0.181,0.211) | 0.193 (0.179,0.208) |
| $e_{23}$ | 0.274 (0.258,0.291) | 0.274 (0.257,0.291) |
| $e_{32}$ | 0.012 (0.008,0.018) | 0.011 (0.007,0.018) |
| $a_{12}$ | | 1.878 (1.526,2.230) |
| $a_{21}$ | | -0.354 (-0.594,-0.114) |
| $a_{23}$ | | 0.781 (0.593,0.970) |
| $a_{32}$ | | -0.531 (-1.733,0.670 |
| -2 × LL | 13146.1 | 12985.89 |

the misclassification probabilities, $e_{ij}$ relate to the acute event effects by

$$\text{logit}(e_{ij}) = a_{ij}^{(0)} + a_{ij}\mathbb{1}\{\text{acute event}\}.$$

**Models with covariates on transition intensities**

Transplantation type is an important factor in the prognosis of post-transplant patients. Patients with single lung transplants are only receiving partial treatment. Their prognosis will depend on the reserve in the remaining native lung. If transplantation type is made a categorical covariate affecting transition intensities between states further significant improvements in the model fit are achieved. In the model with only adjacent misclassification -2×LL is improved by 92.7 from 10 additional parameters. A similar result is found for the model with misclassification from 1 to 3, in that case -2×LL is improved by 79.4 from 10 additional parameters. Rate of onset of disease is lowest for heart and lung patients, with the rate highest for single lung patients. In addition, transition rates

Table 2.8: Model parameter estimates for BOS models allowing misclassification from state 1 to state 3.

| | Model | |
|---|---|---|
| Parameter | 3 | 4 |
| $q_{12}$ | 0.264 (0.235, 0.300) | 0.262 (0.233, 0.295) |
| $q_{14}$ | 0.047 (0.035, 0.062) | 0.047 (0.035, 0.062) |
| $q_{23}$ | 0.382 (0.326, 0.448) | 0.377 (0.321, 0.442) |
| $q_{24}$ | 0.063 (0.037, 0.108) | 0.071 (0.044, 0.115) |
| $q_{34}$ | 0.328 (0.279, 0.385) | 0.320 (0.272, 0.377) |
| $e_{12}$ | 0.051 (0.044, 0.060) | 0.046 (0.037, 0.057) |
| $e_{13}$ | 0.009 (0.007, 0.013) | 0.008 (0.005, 0.011) |
| $e_{21}$ | 0.102 (0.089, 0.117) | 0.093 (0.077, 0.112) |
| $e_{23}$ | 0.310 (0.291, 0.330) | 0.315 (0.295, 0.338) |
| $e_{32}$ | 0.010 (0.007, 0.016) | 0.010 (0.006, 0.016) |
| $a_{12}$ | | 1.630 (1.341, 1.919) |
| $a_{13}$ | | 1.806 (1.157, 2.455) |
| $a_{21}$ | | -0.690 (-1.230, -0.150) |
| $a_{23}$ | | 0.864 (0.650, 1.079) |
| $a_{32}$ | | -0.647 (-1.961, 0.668) |
| -2 × LL | 12520.49 | 12312.66 |

to death from all states are significantly higher in single lung transplant patients. The effects for the model with adjacent state misclassification are given in table 2.9. In this cohort, post transplant life expectancy, conditional on surviving 6 months, is highest for heart and lung patients at 9.34 years (8.46-10.21), compared to 7.28 (5.91-8.71) for double lung and 5.67 (4.92-6.42) for single lung patients.

## 2.2 Comparison with empirical estimates

### 2.2.1 The empirical survivor curve

The use of Kaplan-Meier product limit estimates as an informal way of validating a fitted Markov model for data where the time of entry into the absorbing state is known exactly is common in the literature [55, 78]. The idea is straightforward. When the model implies that all subjects start in the same state at time zero and progress to an absorbing state, if the assumptions in the Markov model are correct, there should not be much disagreement between the empirical survival curve and the survival curve implied by the fitted Markov model. Determining whether the disagreement we observe is within allowable bounds is not so straightforward.

#### Intervals around the Markov curve

One common way of presenting the plot is to calculate the confidence interval about the estimated survival curve from the Markov model. We treat the quantity of interest $p_{1R}(t;\theta)$ at a fixed $t$ as a function of the model parameters, $\theta$, and use the delta method, with the maximum likelihood estimate $\hat{\theta}$ and the observed Fisher information matrix, and get point-wise confidence limits for the probability of being in state $R$ by time $t$. Calculating the standard error of $\text{logit}(p_{1R}(t;\theta))$ and back-transforming, will ensure that the resulting confidence intervals for $p_{1R}(t;\theta)$ lie between zero and one. The delta method is based on a Taylor series approximation which requires first derivatives. For progressive time homogeneous models, $p_{1R}(t;\theta)$ will be available in closed form, and easily differentiable. For models where reverse transitions are possible, closed form expressions for the transition probabilities are not available. However the matrix exponential for $P(t)$ can be differentiated directly to get expressions for the derivatives in terms of partial derivatives and the eigenvalues and eigenvectors of the intensity matrix $Q(\theta)$ [70].

Table 2.9: Model parameter estimates for BOS models including effect of transplant type on disease progression. Heart and Lung transplant patients are taken as baseline.

| | Model | |
|---|---|---|
| Parameter | Adjacent Misc | $1 \rightarrow 3$ misc |
| $q_{12}$ | 0.235 (0.201, 0.275) | 0.208 (0.177, 0.244) |
| $q_{14}$ | 0.026 (0.016, 0.041) | 0.026 (0.016, 0.041) |
| $q_{23}$ | 0.415 (0.343, 0.502) | 0.502 (0.410, 0.616) |
| $q_{24}$ | 0.024 (0.007, 0.087) | 0.025 (0.005, 0.135) |
| $q_{34}$ | 0.287 (0.237, 0.349) | 0.289 (0.236, 0.353) |
| $e_{12}$ | 0.021 (0.019, 0.033) | 0.034 (0.027, 0.054) |
| $e_{13}$ | | 0.005 (0.004, 0.010) |
| $e_{21}$ | 0.207 (0.179, 0.214) | 0.105 (0.071, 0.120) |
| $e_{23}$ | 0.242 (0.252, 0.296) | 0.279 (0.281, 0.357) |
| $e_{32}$ | 0.013 (0.007, 0.018) | 0.011 (0.006, 0.016) |
| $a_{12}$ | 1.892 (1.535, 2.248) | 1.658 (1.344, 1.973) |
| $a_{13}$ | | 1.960 (1.315, 2.605) |
| $a_{21}$ | -0.064 (-0.316, 0.188) | -0.346 (-0.885, 0.194) |
| $a_{23}$ | 0.753 (0.555, 0.950) | 0.810 (0.594, 1.025) |
| $a_{32}$ | -0.573 (-1.790, 0.644) | -0.653 (-1.976, 0.670) |
| $\beta_{12}^{(SL)}$ | 0.828 (0.558, 1.097) | 0.715 (0.441, 0.988) |
| $\beta_{14}^{(SL)}$ | 1.332 (0.645, 2.018) | 1.426 (0.787, 2.064) |
| $\beta_{23}^{(SL)}$ | -0.567 (-0.960, -0.174) | -0.631 (-1.037, -0.224) |
| $\beta_{24}^{(SL)}$ | 1.670 (0.290, 3.050) | 1.644 (-0.136, 3.425) |
| $\beta_{34}^{(SL)}$ | 0.406 (-0.017, 0.828) | 0.384 (-0.044, 0.811) |
| $\beta_{12}^{(DL)}$ | 0.652 (0.337, 0.967) | 0.402 (0.077, 0.727) |
| $\beta_{14}^{(DL)}$ | 0.739 (-0.170,1.649) | 1.025 (0.271, 1.779) |
| $\beta_{23}^{(DL)}$ | -0.848 (-1.357,-0.340) | -0.757 (-1.268, -0.246) |
| $\beta_{24}^{(DL)}$ | 1.358 (-0.109,2.824) | 1.107 (-0.899, 3.112) |
| $\beta_{34}^{(DL)}$ | 0.295 (-0.257,0.847) | 0.303 (-0.249, 0.854) |
| -2 $\times$ LL | 12893.19 | 12233.13 |

The calculation of derivatives can be quite intricate. An alternative, which may in fact be more accurate because it doesn't involve the linearisation inherent in the delta method, is to get the confidence limits of $p_{1R}(t;\theta)$ through simulation. We assume that

$$\hat{\theta} - \theta \sim \mathrm{N}(0, I(\hat{\theta})^{-1})$$

and use this to sample random vectors $\hat{\theta}^*$ for which $p_{1R}(t;\hat{\theta}^*)$ is calculated [3].

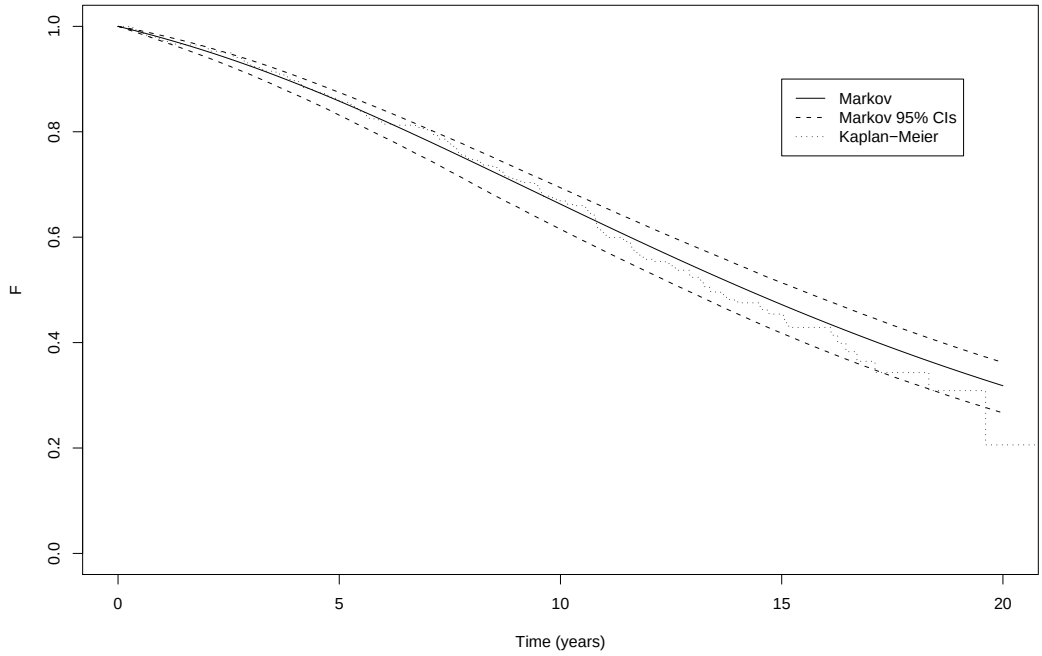
A third option for the calculation of standard errors is to use a non-parametric bootstrap. This has the advantage of not assuming asymptotic results. However, for complex models the necessity to re-fit the model for each bootstrap realisation makes the method less attractive.

If the model is correct and the sample size sufficient so that a multivariate normal approximation to the maximum likelihood estimate is reasonable then this method will give valid piecewise confidence intervals for the probability of survival. However, a criterion based upon whether the Kaplan-Meier estimate lies within this confidence interval, even for a chosen single point, will not give a test with 95% coverage. Clearly the Kaplan-Meier estimate has diminishing accuracy towards the end of the curve where the number of events is small. This will not be reflected in the confidence intervals for the Markov curve as the model assumptions allow confident extrapolation beyond the final event time. Figure 2.3 gives the comparison plot for the model for CAV data. The Kaplan-Meier curve lies outside of the Markov model's confidence band at around 19 years and has lower survival from around 10 years after transplantation.

**Intervals around the Kaplan-Meier curve**

An alternative is to use the imprecision of the Kaplan-Meier estimate and construct the $100(1-\alpha)\%$ confidence intervals [64], typically calculated using a normal approximation like Greenwood's formula. This approach is attractive because it is simple to perform. These point-wise confidence intervals would be correct if we were testing against an entirely specified Markov model (or indeed any other entirely specified curve). However, when the Markov model is fitted from the same data as the survival curve, the Markov curve will tend to be closer to the Kaplan-Meier curve. Hence, the 95% confidence intervals have more than 95% pointwise coverage. Given that we will not, in practice, be considering a single point, but rather multiple testing of a series of (highly correlated) points, the precise boundaries of the confidence interval are perhaps not of great importance. What

Figure 2.3: Comparison of the estimated survival curve for the CAV data under a Markov model with the empirical survival curve. 95% confidence intervals are around the Markov curve.



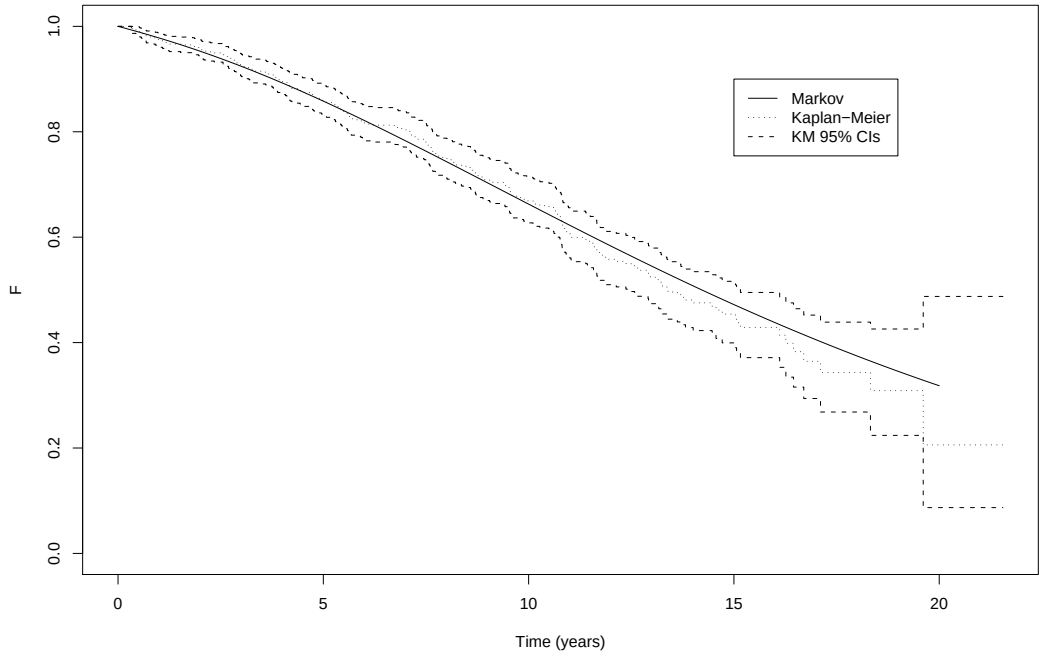

is important is that an idea of the relative uncertainty is known, and this is achieved for these intervals, unlike intervals which only consider the uncertainty of the Markov estimate. Figure 2.4 gives the comparison plot for the model for CAV data. Compared to figure 2.3, the confidence intervals are wider, in particular there is a far more marked widening after 15 years.

### Use of formal statistical tests

While most authors have been content to consider empirical comparisons as only an informal diagnostic, Pérez-Ocón *et al* [102, 105, 106], in a series of papers, attempted to put a formal p-value onto the overall discrepancy of the fitted Markov curve and the Kaplan-Meier estimate.

They used a test proposed by Hollander and Proschan [60]. For right-censored data from

Figure 2.4: Comparison of the estimated survival curve for the CAV data under a Markov model with the empirical survival curve. 95% confidence intervals are around the Kaplan-Meier curve.




a distribution $F(x)$ this tests the hypothesis $H_0 : F(x) = G_0(x)$.

Suppose we have data $z_1, \ldots, z_n$, where $Z_i = \min(Y_i, T_i)$, where $Y_i$ is the failure time and $T_i$ the censoring time for subject $i$, and let $\delta_i = 1$ if $z_i = y_i$ (if the observation is uncensored) and be zero otherwise.

The test statistic is

$$C = -\int R_G(y) d\hat{R}_F(y),$$

where $R_G(y) = 1 - G_0(y)$ and $\hat{R}_F(y)$ is the Kaplan-Meier estimate of the survivor function for $\mathbf{y}$. For computational purposes a simplified version is

$$C = \sum_{\delta_i = 1} R_G(Z_i) \hat{f}(Z_i)$$

where $\hat{f}(Z_i)$ is the jump of the Kaplan-Meier distribution at $Z_i$. Under $H_0$, $C$ has expec-

tation $\frac{1}{2}$. A consistent estimator for the variance of $C$ can be obtained from

$$\hat{\sigma}^2 = 16^{-1} \sum_{i=1}^{n} (\frac{n}{(n-i+1)}) \left( [(R_G(Z_{i-1})]^4 - [(R_G(Z_i)]^4 \right).$$

Hence under $H_0$

$$C^* = \frac{\sqrt{n}(C - \frac{1}{2})}{\hat{\sigma}}$$

has a standard normal distribution. A two-sided test of size $\alpha$ can then be constructed by rejecting if $|C^*| > \Phi^{-1}(1 - \frac{\alpha}{2})$, where $\Phi(y)$ is the cdf. of a standard normal distribution.

However, the distribution in the null hypothesis, $G_0(y)$, must be fully specified in order for the null distribution of the test to have the stated asymptotic distribution. When applied to a curve that has itself been fitted to the data it is clear that the null distribution of the statistic will be different, tending to have smaller values. The precise null distribution would be very difficult to determine. However, simulation can show what is likely to occur.

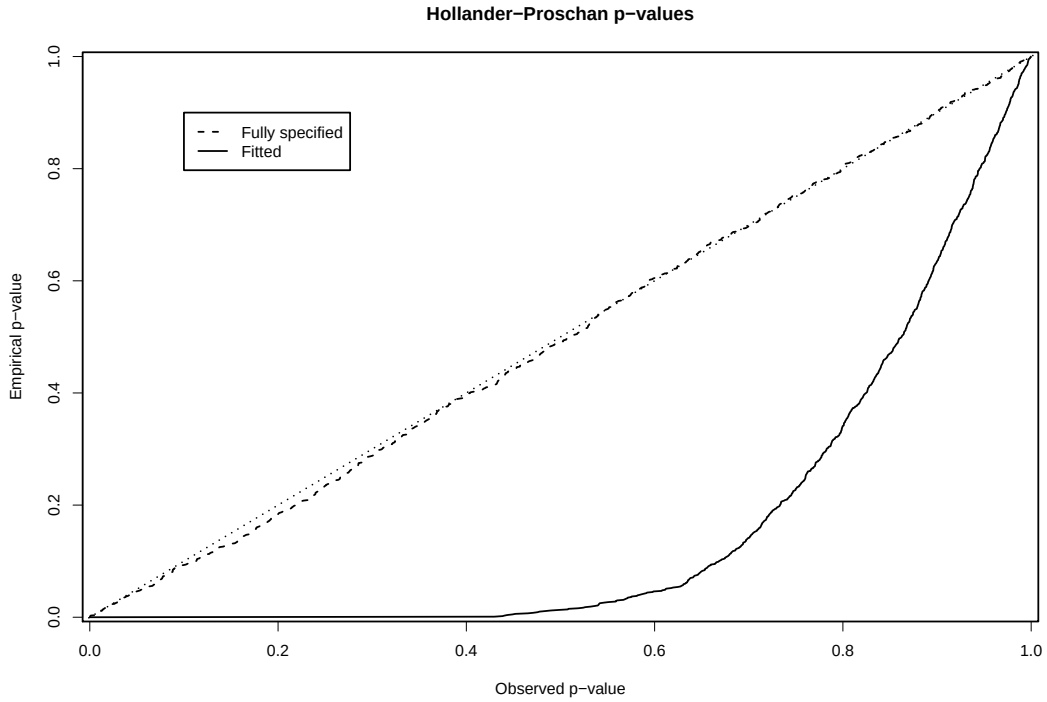
**Hollander-Proschan simulation study**

We simulate data from a three-state disease Markov model in which 500 subjects are observed at one year intervals until either death, the time of which is known exactly, or censoring, which is random with a $U[0, 40]$ distribution. The transition intensities were chosen such that around $\frac{2}{3}$ of subjects had an observed death. The Kaplan-Meier curve for the data is calculated. The Hollander-Proschan test statistic is calculated, firstly for the survival curve implied by the full specified model, and secondly using the survival curve implied by the Markov model with parameters fitted from the data.

As one would expect, the test on the fully specified model gives p-values that are very close to being $U[0, 1]$. In contrast, against the Markov curve fitted to the data, the null distribution is dramatically changed. From 1000 samples, the smallest p-value observed was 0.31, with the median being 0.85. Hence, using this test will probably lead to accepting models which are in fact a significantly poor fit (for instance if p-values between 0.05 and 0.3 are observed). The observed p-values are plotted in figure 2.5.

It seems unlikely that a closed form null distribution for a Holland-Proschan type of statistic could be found when the model is not fully specified. The distribution is likely to be dependent on how much of the dataset relates to the survival aspect directly (censoring and observed deaths) and how much only relates to survival indirectly (transitions between

Figure 2.5: Observed p-values for simulated Markov model data sets using the Hollander-Proschan test for the fully specified model and the fitted model. Observed lines should be near to the diagonal line.




transient states of the model). If a p-value is nevertheless desired, one possible approach would be to bootstrap the null distribution of the statistic, which would be time consuming for large datasets.

### Lawless (1982) test

A more suitable test in the situation of unknown parameters is that of Lawless [83]. This tests the hypothesis that the failure times come from some family of survival distributions, $F(t;\theta)$, where $\theta$ is to be estimated from the data. In the case of a Markov model where all subjects are in state 1 at time 0, $F(t;\theta) = 1 - p_{1R}(t;\theta)$. The test considers a set of times $0, t_1, \ldots, t_n$. The observed failure times are grouped according to which interval $I_i = (t_i, t_{i+1}]$ they lie within. The frequencies within each interval, $O_i$ comprise the

observed data. The expected frequency for the interval $I_i$ is calculated by taking

$$r_i \left( 1 - \frac{F(t_{i+1}; \hat{\theta})}{F(t_i; \hat{\theta})} \right) \tag{2.1}$$

where $r_i$ are the number of subjects observed to be at risk at time $t_i$. Each $O_i$ can then be thought of as being binomially distributed. A Pearson-chi-square test or an asymptotically equivalent likelihood ratio test, can then be performed on the data. However, the calculation for the expected frequencies is valid only if it can be assumed that subject's censoring is restricted to the points $t_1, \ldots, t_n$. If they can instead occur at any time, the expected counts given from equation 2.1 will be higher than the true value. A further drawback of the test is that even if the censoring assumptions were true, the null distribution of the statistic is known only to lie between $\chi^2_{n-p-1}$ and $\chi^2_{n-1}$, where $p$ is the dimension of $\theta$. This is because $\hat{\theta}$ is estimated from the full data, rather than the grouped data used in the test. For models with a large number of parameters compared to the number of intervals, this can mean there is a very large degree of uncertainty about the p-value. For instance Pérez-Ocón *et al* (2001) [104] considered a model with piecewise constant hazard intensities and covariate effects which had 36 parameters. Grouping the failure times in 3 month intervals led to a p-value for a test which could only be said to be bounded between 0.0006 and 0.8737.

Neither approach to formal testing seems completely effective. It is therefore most appropriate to only use comparisons with the empirical survival estimate as an informal assessment tool, using the 95% confidence intervals on the Kaplan-Meier curve as a guide.

**Continuous covariates case**

Kaplan-Meier product limits assume homogeneity between subjects. For categorical covariates, it may be reasonable to compare an individual Kaplan-Meier estimate for the subset of data with a particular covariate pattern. Using Cox-proportional hazards models instead of Kaplan-Meier estimates as the empirical benchmark has been proposed [125]. Each set of covariate values will give a different pair of estimated curves, so it is necessary to consider the fit for a selected range of values.

A note of caution needs to be taken when making a comparison using the Cox-proportional hazards model as the empirical estimate. In the homogeneous subject case the time homogeneous Markov model for the deaths is contained within the parameter space of the empirical Kaplan-Meier survivor curve, meaning that as the sample size tends to

infinity, the Kaplan-Meier estimate will converge in probability to the survivor function of the Markov model. It is not however the case that a Markov model with covariates is contained within the parameter space of the Cox-proportional hazards model. In the Markov model, the proportional hazards assumption is met for each of the individual transition intensities, so that

$$q_{rs}(z) = q_{rs}^{(0)} \exp(\beta_{rs}^T z).$$

However, this does not imply that the hazard of death given $X(0) = 1$ satisfies the proportional hazards assumption. In fact, the proportional hazard assumption is only met if the covariates affect all transition intensities to the same degree. Equivalently if

$$Q(z) = Q^{(0)} \exp(\beta^T z). \tag{2.2}$$

This is illustrated in figure 2.6, which gives the estimated functions for log-hazard to death conditional on $X(0) = 1$ for the misclassification hidden Markov model for the CAV data. As can be seen the log-hazard functions do not stay parallel for varying values of the covariate. In the model, primary diagnosis and donor age only affect times of onset for CAV. The transition intensities to death are unaffected.

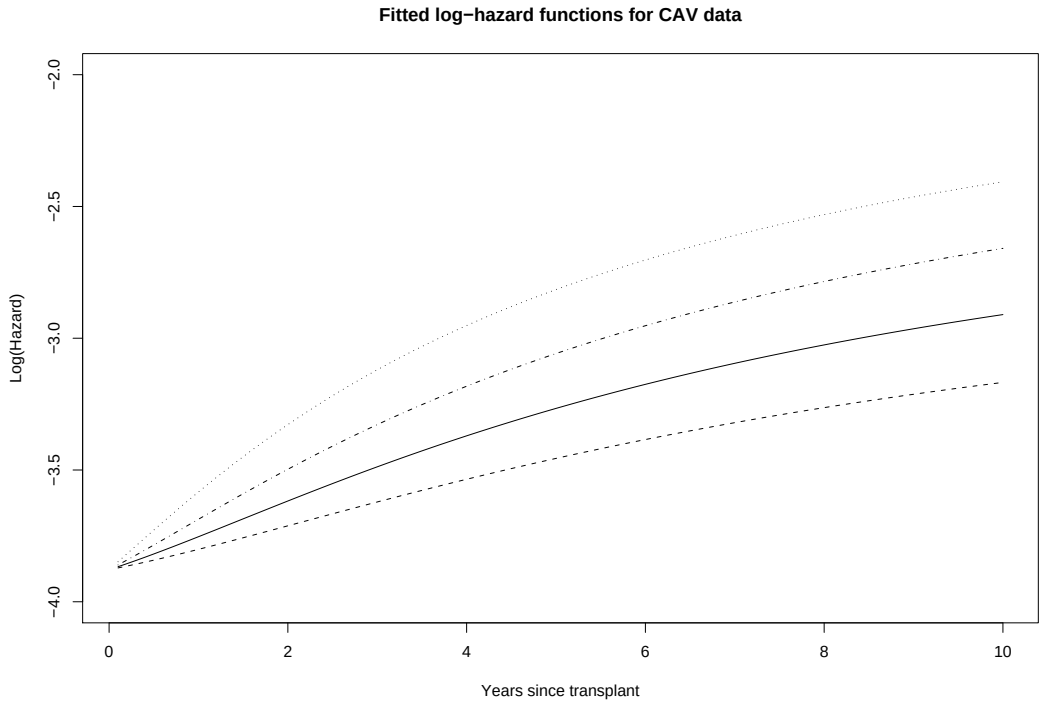
The consequence of this result is that it is entirely possible to have disagreement between the estimated survival curves for the Markov model and the Cox-proportional hazards model, even if the Markov model is correct. Thus such a test should only be employed when either the covariate model is such that equation (2.2) is satisfied or where the maximum likelihood estimates of the covariate effects turn out to be such that $\beta_{ij} \approx \beta_{kl}$ for all pairs $i, j$ and $k, l$ in $1, \ldots, R$ so that there is little evidence against an overall proportional hazard in the fitted Markov model.

In the absence of methods for fitting semi-parametric multi-state models to panel observed data, it is not clear what the correct approach to making empirical comparisons might be in the case of continuous covariates. Certainly other obvious standard survival modelling techniques such as *accelerated failure-time models* will suffer from the same problems as proportional hazards, since the Markov model will not, in general, satisfy the accelerated failure-time assumptions either.

### 2.2.2 The empirical hazard function

An alternative way of assessing the fit of a Markov model, when the times to absorption are known exactly, is via the empirical hazard function. The times from initiation to

Figure 2.6: Log-hazard functions for varying covariate values for the misclassification hidden Markov model fitted to the CAV data




absorption will, if the Markov assumption is correct, follow a *phase-type* distribution. Direct reference to the estimated empirical hazard has not been common in the Markov modelling literature, with the exception of Pérez-Ocón *et al* (2001) [105] who observed a hazard function which peaked at around 48 months and then declined. They then took 48 months as the change-point in a time inhomogeneous Markov model with two regions of piecewise constant hazard. Aalen [2, 4] encouraged the consideration of phase-type distributions to describe the shape of the hazard curve in general survival analysis.

Suppose we know with certainty the time of initiation of our process and the state at initiation. We might also be confident that the process is progressive, perhaps because the disease of interest is chronic and recovery is considered biologically infeasible. In a disease process situation, it would also be expected that higher states in the process - those corresponding to more advanced disease - lead to higher rates of mortality. If we were sure of these features of the process this limits the possible shapes that the phase-type

distribution can take. The hazard of entering the absorbing state can be written as

$$h_R(t) = \sum_{r=1}^{R-1} q_{rR}(t)\mathbb{P}(X(t) = r|X(0) = 1).$$

For a time homogeneous model where the $q_{rR}$ are constant, the hazard is driven entirely by $\mathbb{P}(X(t) = r|X(0) = 1)$ which are the occupation probabilities, conditional on not having reached the absorbing state. The *quasi-stationary distribution*, $(\tilde{p}_1, ..., \tilde{p}_{R-1})$, of the process is defined as the limit of these conditional occupation probabilities as $t$ tends to infinity. Since these conditional occupation probabilities tend to a limit, so does the hazard of absorption. Specifically

$$h_R(t) \to \sum_{r=1}^{R-1} q_{rR}(t)\tilde{p}_r \quad \text{as} \quad t \to \infty.$$

It is well established [75] that the quasi-stationary distribution is the normalised version of the left eigenvector corresponding to the dominant eigenvalue (the eigenvalue closest to zero) of the transition intensity matrix restricted to the transient state space. Note that this distribution depends not just on the transition intensities between transient states but also on the transition intensities to the absorbing state in each of those states. In particular the quasi-stationary distribution will only be

$$\tilde{p}_1 = ... = \tilde{p}_{R-2} = 0, \tilde{p}_{R-1} = 1$$

when the dominant eigenvector of the transient state space corresponds to state $R - 1$. For instance in a progressive disease model, this means that state $R - 1$ must have the longest mean sojourn time of any of the transient states.

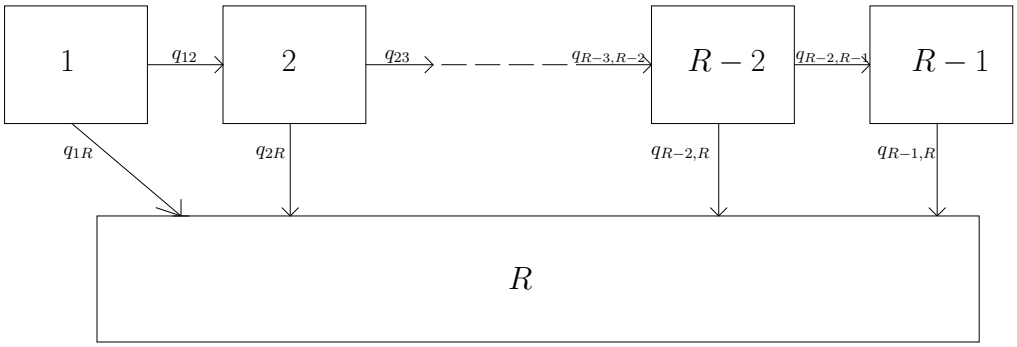
Aalen [2] noted that for a unidirectional process in which all subjects begin in state 1, there is an increasing hazard. It would be useful to be able to extend such a result to models with more practical applicability.

Let $X$ be continuous time Markov($Q$,$\mathcal{S}$), where $\mathcal{S} = \{1, ..., R\}$ is the set of states and ($Q$), the transition intensity matrix, is non-zero only along the diagonal, the first upper off-diagonal and the $R$ column, with constant entries $q_{r,r+1}$, $q_{r,R}$ $r = 1, ..., R - 1$. This implies that the time to absorption is governed by an $R - 1$ phase, Coxian phase-type distribution. The structure of the Markov process is shown in figure 2.7.

In addition, let

$$q_{1R} < q_{2R} < ... < q_{R-1,R}.$$

Figure 2.7: State diagram for Markov process $X(t)$




Then the hazard of entering the absorbing state

$$h_R(t) = \frac{\sum_{r=1}^{R-1} q_{rR} p_{1r}(t)}{\sum_{r=1}^{R-1} p_{1r}(t)}$$

is monotonically increasing in $t$.

This statement seems to be true as intuitively we would expect that as time progresses, the distribution of subjects yet to have been absorbed, will become more concentrated in higher states and hence the hazard will increase towards the hazard at the quasi-stationary distribution. However, a formal proof of the result, beyond the three-state case, seems non-trivial.

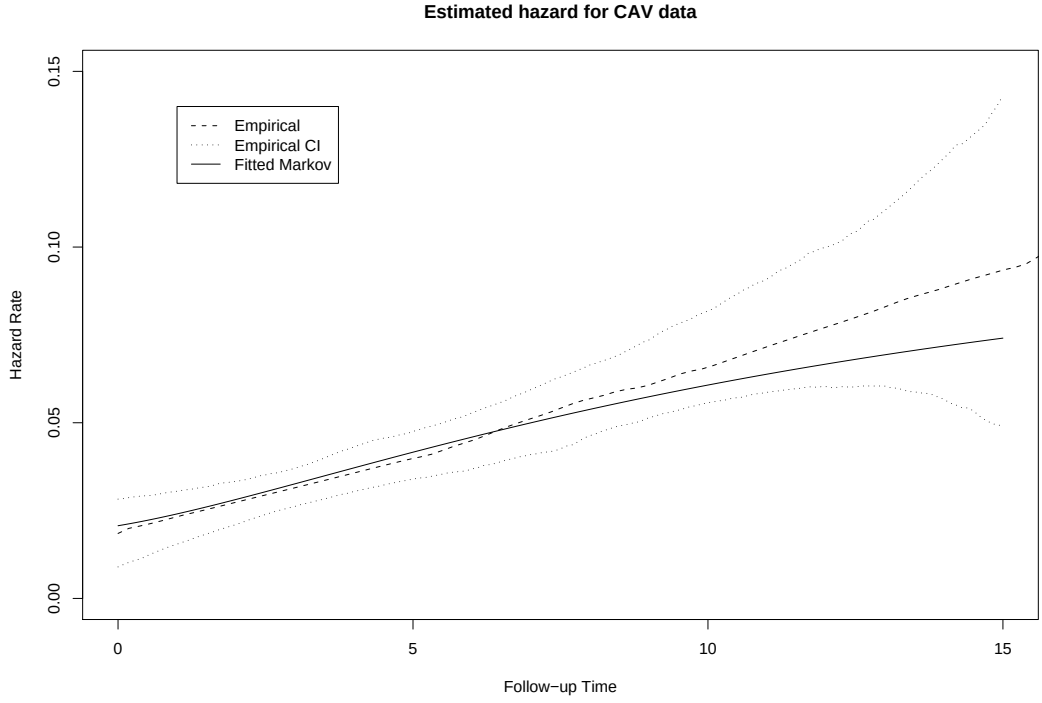
Assuming the result is true, it is useful if we have data that are assumed to be from a progressive disease model. Provided the hazard of absorption can be adequately estimated, if we do not see an increasing hazard we can be reasonably sure that one or more of the original assumptions (time homogeneity, subject homogeneity, disease free at transplant, progressive nature of process) is incorrect, although more investigation would be required to determine which assumption or assumptions are most likely to be incorrect. Thus inspection of the empirical hazard is a useful informal diagnostic to perform before fitting a Markov model in this context. The value of examining empirical hazards in more general models is uncertain.

The empirical hazard function for randomly right-censored data can be estimated using kernel density estimation [97]. The **muhaz** package in R performs such estimation. Pointwise 95% confidence intervals for the hazard estimates can be calculated by performing the kernel density estimation on 1000 bootstrap samples.

**Example: CAV and BOS data**

As outlined in section 2.1.1, the CAV data are assumed to follow a progressive disease model. Discarding for the moment, the known heterogeneity between subjects, we would otherwise expect an increasing hazard to be apparent in the data. Figure 2.8 shows that the empirical hazard function is indeed increasing. The same plot can also be used for verification after fitting a Markov model. Here, the hazard curve predicted by the fitted model stays well within the bootstrap 95% confidence intervals, although there is some divergence from the empirical hazard over time.

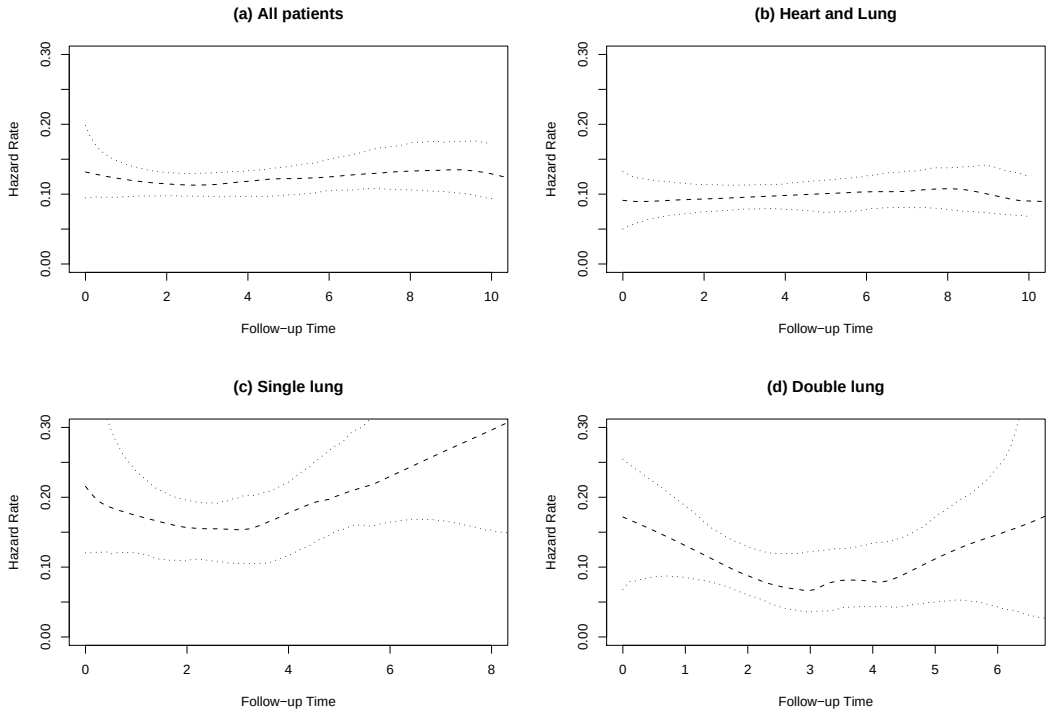
Figure 2.8: Empirical hazard function for the CAV data from kernel density estimation with bootstrap 95% confidence intervals. The bold line is the hazard for the fitted Markov model.




The BOS data are similarly assumed to follow a progressive disease model. As with the CAV data there are some known covariates, particularly the type of transplant. As these known covariates are categorical it is straightforward to consider the hazard for each subgroup. Without categorising by transplantation type, the estimated hazard function is relatively flat (figure 2.9 (a)). For the single and double lung transplantation subgroups

there is an insufficient number of events to estimate the shape of the hazard with any precision (figure 2.9 (c)-(d)). For heart and lung patients, who make up the largest single group, the hazard is markedly flat (figure 2.9 (b)). This alone is enough to tell us that a Markov disease model which makes the assumptions of time homogeneity or subject homogeneity, will not be a very good fit in terms of the survival process.

Figure 2.9: Empirical hazard function for the BOS data from kernel density estimation with bootstrap 95% confidence intervals.




There is some merit in considering the empirical hazard function, particularly for ruling out model assumptions. However, often there are insufficient data to adequately estimate the shape of the hazard function. Moreover, much of the analysis that can be performed with the empirical hazard is only applicable for homogeneous populations.

## 2.3 Prevalence counts

Comparisons with the survival or hazard curve of the absorbing state of the process test only part of the fit. Ideally, similar empirical curves for the occupancy within all the states

would be desirable. However, as explained in section 1.4.3, interval censoring makes these difficult to obtain. Prevalence counts provide a more informal empirical measure of state occupancy.

Prevalence counts involve comparing the observed state occupancies at a fixed set of times and comparing them with those expected by the fitted model [55]. This method attempts to overcome problems of irregular observation times. The method is also applicable in the case of exact death times.

A table of observed and expected state occupancies at a series of times is constructed. It is necessary to interpolate in some way because the precise state each subject occupies at the assigned times will not be known. The suggested method is to assume that a subject would still be in the state in which they were observed at the previous observation, effectively this means that jumps are assumed to occur immediately before an observation. This inevitably introduces some bias and inaccuracy, but it is suggested this might not be significant provided subjects are observed frequently. In the original paper, Gentleman *et al* were considering a disease model in which reverse transitions were possible. The bias in that case will probably be less severe than for a progressive model. For a progressive model, assuming no additional transition has occurred implies the estimated observed state for a subject is always an underestimate of their true state.

An alternative approach is to assume that any transitions occurred at the mid-point of any interval. This seems to be a better choice for progressive models. However, when observed transitions imply the passing through of a series of states, it is inevitable that intermediate states will be underestimated in the prevalence counts.

Mathieu *et al* [95] instead chose to only consider the subgroup of subjects who had an observation close to the time of interest. They chose those subjects who were observed within 44 days of 1 year for instance. Therefore interpolation is avoided. This method works well provided sampling is not only non-informative, but also avoids *doctor's care* [57], where ill patients have their next observation scheduled earlier. If doctor's care is present then ill patients are more likely to be observed near to the times of interest and cause over-representation in the sample. Mathieu *et al*'s approach is not possible either in the presence of exact death times, unless it was known when their sampling times would have been, had they survived.

The expected counts are calculated by summing the probability a subject is in the specified state given their initial state over all subjects who are under observation at the time of

interest. A subject is under observation until their last observation time, or, if they reach the absorbing state, until the time at which they would have been censored had they survived. This is crucial, if censored subjects are taken out at their censoring time but patients who die are left as under observation until the final follow-up time, then the observed prevalence in the death state will be systematically overestimated.

An indication of where the data deviate from the model is gained by comparing the observed count $O_{ri}$ with the expected count $E_{ri}$ for particular state $r$ and time $t_i$ through:

$$M_{ri} = \frac{(O_{ri} - E_{ri})^2}{E_{ri}}$$

where

$$E_{ri} = \sum_u p_{g_u,r}(t_i, z_u) \tag{2.3}$$

where $g_u$ is the (assumed known) initial state and $z_u$ the covariate vector, for subject $u$. A large value of $M_{ri}$ would indicate a poor fit. However, formal tests to determine whether the deviances observed are statistically significant are not possible. This is due to the *ad hoc* interpolation of observed states and also the dependence between the rows of the tables. Therefore prevalence counts can only be used as an informal measure of fit. Covariates within the model do not present a problem as each subject will merely contribute a different transition probability to the expected count.

### 2.3.1 Example: CAV data without misclassification

Using the CAV data with no misclassification, we construct a table of prevalence counts at times 1, 2, 4, 7, 10 and 15 years. Interpolation using the 'last observed state' method and the 'mid-point transition' methods is performed. At least for the chosen set of times, the 'last observed state' method performs quite badly (table 2.10), spuriously indicating a poor fit. This is because patients tended to be observed just after the year interval points. Hence, the observed state was badly under-estimated. The 'mid-point transition' method seems to perform better for these data (table 2.11). However, there are still some apparent problems especially for the counts at time 1, where the chi-squared style observed versus expected comparison of state 2 gives a value of 38.7, which would be a considerable cause for concern. However, the under representation of state 2 becomes entirely understandable given that the mean time at which patients who were assumed to have been observed at year 1 were actually observed was 0.20 years (in the majority of cases time zero was taken).

Table 2.10: Prevalence counts for CAV data without misclassification: using 'last observed state' interpolation. 'Obs time' refers to the mean time the state used in the observed table actually occurred.

| | Obs | Observed States | | | | Expected States | | | |
|---|---|---|---|---|---|---|---|---|---|
| Time | time | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | 0.03 | 581 | 0 | 0 | 15 | 527.0 | 50.0 | 5.0 | 13.9 |
| 2 | 0.23 | 562 | 6 | 2 | 20 | 462.2 | 81.9 | 16.8 | 29.1 |
| 4 | 2.51 | 415 | 47 | 29 | 50 | 335.0 | 103.1 | 43.2 | 59.6 |
| 7 | 5.63 | 239 | 82 | 44 | 78 | 195.6 | 85.5 | 64.5 | 97.3 |
| 10 | 8.34 | 116 | 60 | 45 | 120 | 110.1 | 55.4 | 60.5 | 115.0 |
| 15 | 13.1 | 25 | 15 | 15 | 85 | 30.0 | 14.9 | 23.8 | 71.4 |

The apparent over representation of deaths (23 observed compared to 13.9 expected) is because, if a death occurs within $(0, 1)$, the status at time 1 will be known precisely. Hence there is selection bias. We will return to this theme later (see Chapter 3).

It is clear that prevalence counts cannot be applied without some consideration of the choice of times. Given few subjects in the CAV data had angiography before time 2, it was inappropriate to have any counts before that time. Prevalence counts are more appropriate when each subject is observed regularly so that the potential bias in the observed counts is less important.

### 2.3.2 Graphical generalisation of prevalence counts

Rather than limit ourselves to a few times of interest, the estimated prevalences can be calculated at all times and plotted against the expected prevalences according to the fitted model. As before, this procedure is heavily influenced by the sampling scheme and interpolation procedure chosen. Figure 2.10 shows this method applied to the CAV data without misclassification, using the 'last observed state' interpolation method. There is higher than expected estimated state 1 prevalence but lower than expected for states 2 and 3. The sampling scheme is apparent in the large jumps in prevalences around year 2 and year 4, when the majority of subjects are observed. It is clear that a substantial proportion of the disparity between observed and expected is due to bias from interpolation.

Table 2.11: Prevalence counts for CAV data without misclassification: using 'mid-point transition' interpolation.

| | Obs | Observed States | | | | Expected States | | | |
|---|---|---|---|---|---|---|---|---|---|
| Time | time | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | 0.20 | 565 | 6 | 2 | 23 | 527.0 | 50.0 | 5.0 | 13.9 |
| 2 | 2.17 | 480 | 61 | 21 | 28 | 462.2 | 81.9 | 16.8 | 29.1 |
| 4 | 3.19 | 357 | 80 | 39 | 65 | 335.0 | 103.1 | 43.2 | 59.6 |
| 7 | 6.33 | 207 | 80 | 49 | 107 | 195.6 | 85.5 | 64.5 | 97.3 |
| 10 | 9.03 | 98 | 62 | 38 | 143 | 110.1 | 55.4 | 60.5 | 115.0 |
| 15 | 13.6 | 21 | 13 | 12 | 94 | 30.0 | 14.9 | 23.8 | 71.4 |

The estimate of state 4 prevalence differs from the Kaplan-Meier estimate. This is because a different estimation procedure has been used. Here, a *simple moment estimator* has been used which just considers the proportion of those still under observation in each state. This can result in upward jumps in the estimated prevalence for state 1, and downward jumps for state 4. A more sophisticated approach might be to employ the Aalen-Johansen estimator [1]. The Aalen-Johansen estimator is derived through the use of counting processes. The empirically estimated prevalences using Aalen-Johansen are defined by

$$\hat{p}_{1s}(t+\delta t) = \sum_{r=1}^{R} \hat{p}_{1r} \frac{dN_{rs}(t)}{Y_r(t)}$$

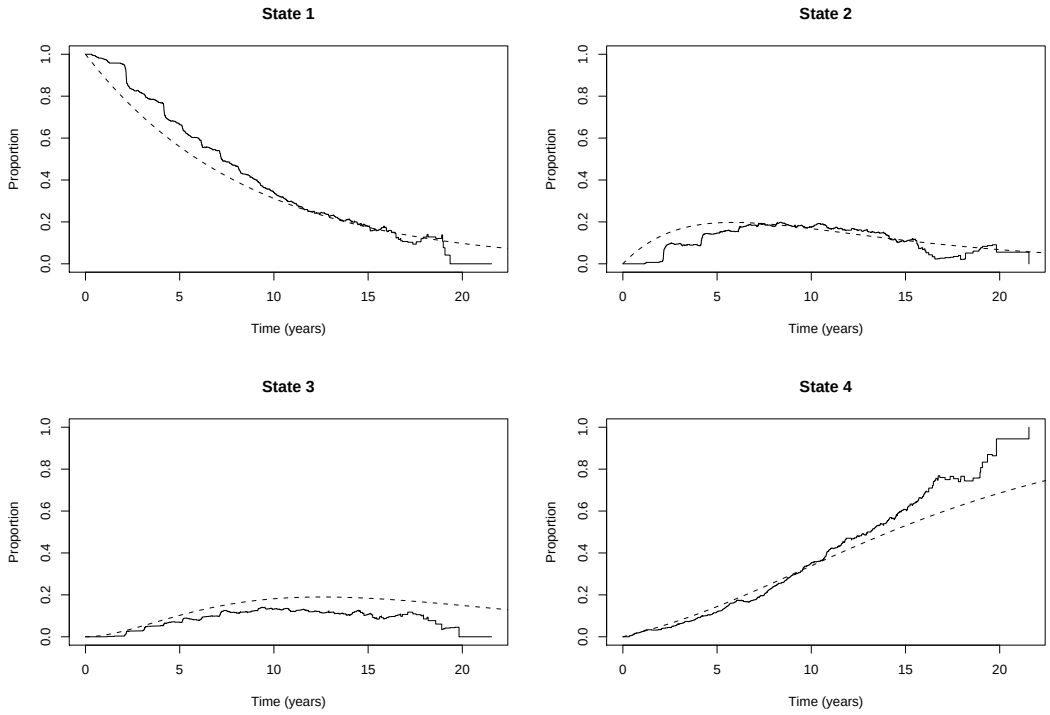
where $Y_r(t)$ represents the number of subjects under observation in state $r$ at time $t$ and $dN_{rs}(t)$ is the number of transitions from $r$ to $s$ in $(t, t+\delta t)$.

When transition times are known up to right censoring, the Aalen-Johansen estimator is more efficient than the simple moment estimator. However there doesn't seem much merit in employing it for the transient states here where transition times are interpolated.

### 2.3.3 Prevalence counts for misclassification models

Prevalence counts have not been used for assessing the fit of misclassification models. However the concept can be extended to incorporate misclassification. If the prevalence counts are constructed for times $t_1, ..., t_n$ and all subjects are to be observed at precisely

Figure 2.10: Graphical prevalence plots for the CAV data without misclassification using 'last observed state' interpolation. Bold line = observed prevalence. Dashed line = expected prevalence.




those times then the expected contribution for time $t_i$ and state $r$ is:

$$E_{ri} = \sum_u \sum_k p_{g_u k}(t_i; z_u) e_{kr}$$

where $z_u$ is the covariate value and $g_u$ is the assumed known initial state for patient $u$. As before, some interpolation is necessary to determine the actual observed counts. A new source of potential bias is that usually subjects are assumed to be observed at time zero without misclassification. One possible way around this problem is to give fractional weighting to the observed counts from time zero. This has the disadvantage of needing to use the estimated misclassification probabilities in order to construct the observed counts, but is otherwise effective. Specifically, given that all subjects are assumed to be in state 1 at time zero, we give a weighting of $e_{1r}$ to them having been observed in state $r$ at time zero.

The BOS data, which have fairly regular observations, are much more suited to the ap-

Table 2.12: Prevalence counts for BOS data: Fractional observed counts are due to the bias correction.

| | Observed States | | | | Expected States | | | |
|---|---|---|---|---|---|---|---|---|
| Time | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | 309.1 | 53.9 | 46 | 59 | 337.6 | 64.0 | 44.0 | 22.5 |
| 2 | 226.3 | 60.7 | 64 | 95 | 238.4 | 77.7 | 79.8 | 50.1 |
| 3 | 177.6 | 51.4 | 74 | 117 | 165.5 | 73.9 | 99.9 | 80.7 |
| 4 | 138.8 | 36.2 | 76 | 140 | 113.2 | 62.6 | 105.0 | 110.1 |
| 5 | 107.9 | 38.1 | 61 | 151 | 75.9 | 49.2 | 98.8 | 134.1 |
| 6 | 87.9 | 30.1 | 60 | 156 | 51.7 | 37.9 | 88.9 | 155.5 |

plication of prevalence counts than the CAV data. Here we compare the observed counts with the naive fitted model in which there are no covariates on the transition intensities or the misclassification probabilities (table 2.12).

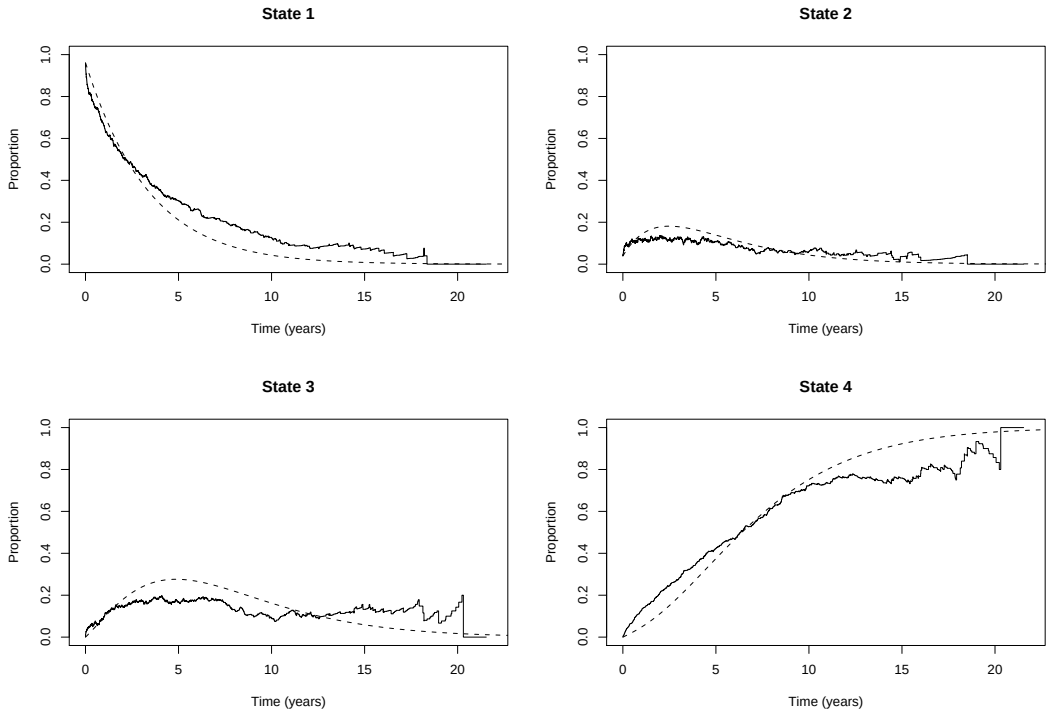
The graphical generalisation of the prevalence counts can also be applied when there is misclassification. Plotting the BOS data in this way is particularly illuminating (figure 2.11). There is a sharp decrease in state 1 prevalence and sharp increases in state 2 and state 3 prevalences immediately after time zero. This suggests there is a problem with either the assumption of universal state 1 occupancy at time zero or with the misclassification model. A patient's baseline level is established in the first 6 months after transplant. This lack of fit may be due to some patients taking longer to reach their baseline.

None of the states have a good overall fit, the HMM estimate for state 1 is an overestimate initially and a significant underestimate for later times, state 2 and state 3 have much flatter observed curves than predicted.

### 2.3.4 Conclusion

Prevalence counts have limited use for assessing model fit in the context of panel observed data. The table form of counts can be useful if all subjects are observed at (or close) to a set of time points. When the times of observation are more irregular the graphical form of the counts is preferable. However, unless observations are frequent, there will be considerable bias in the empirical prevalences making conclusions about the fit of the model difficult to arrive at. In most situations therefore, prevalence counts can only provide a crude

Figure 2.11: Graphical prevalence plots for the BOS data. Bold line = observed prevalence. Dashed line = expected prevalence.




measure of model fit. Accurate empirical prevalence estimates are needed, but currently no methodology exists to provide them for discretely sampled multi-state data.

## 2.4 Residual plots

### 2.4.1 Outlier identification

It is desirable to be able to identify subjects within our data that are in some respect outliers. They are likely to have a substantial influence on the parameter estimates. If many of the most significant outliers exhibit the same sort of behaviour, this might indicate areas of the model in which the fit is poor.

**Jackknife Residuals**

One way of determine the influence of a particular subject on the overall parameter estimates is to simply delete that subject and re-fit the model, then compare the distance between the estimates. Such an approach is standard in survival analysis [131]. If we have $n$ subjects and a parameter vector $\theta \in \Theta$, with maximum likelihood estimate based upon the whole data $\hat{\theta}$. Then let $\hat{\theta}_{(j)}$ represent the estimate with subject $j$ deleted. We will thus be interested in the quantities $\hat{\theta}_{(j)} - \hat{\theta}$ for $j = 1, ..., n$. The influences of each point on each parameter could be compared separately. If we want to get a measure of the overall influence of a particular subject we can take the scalar quantity

$$(\hat{\theta}_{(j)} - \hat{\theta})^T I(\hat{\theta})(\hat{\theta}_{(j)} - \hat{\theta})$$

where $I(\hat{\theta})$ is the observed Fisher information at the maximum likelihood estimates for the full data. Computing the jackknife estimates requires refitting the model $n$ times. This can be time consuming for large datasets, particularly those with a large number of subjects. An often used approximation to the full jackknife estimate is to consider instead the contribution to the score function of each subject evaluated at the maximum likelihood estimate for the full model. The score is just the first derivative of the likelihood. Subjects with a high influence will have a score with a high magnitude. Hence an analogous scalar measure:

$$U_j(\hat{\theta})^T I(\hat{\theta})^{-1} U_j(\hat{\theta})$$

can be used to identify outliers.

For time homogeneous Markov models, the primary difficulty in calculating the score function, $U(\theta) = \frac{\partial l}{\partial \theta}(\theta)$, is determining the derivative of the transition probability matrix $P(t)$ with respect to $\theta$. These derivatives were given by Kalbfleisch and Lawless [70]. As shown in section 1.2.2 we can write $P(t) = U \exp(tD)U^{-1}$ where $U$ is the $k \times k$ matrix of eigenvectors and $D$ is the diagonal matrix of eigenvalues of the intensity matrix $Q$. Then the first derivative with respect to the $u$th component of $\theta$ is

$$\frac{\partial P(t)}{\partial \theta_u} = U V_u U^{-1}$$

where $V_u$ is a $k \times k$ matrix with $(i, j)$ entry

$$\begin{cases} \frac{g_{ij}^{(u)}(\exp(d_i t) - \exp(d_j t))}{(d_i - d_j)} & i \neq j \\ g_{ii}^{(u)} t \exp(d_i t) & i = j \end{cases}$$

where $g_{ij}^{(u)}$ is the (i,j) entry of the matrix $G_u = U^{-1}(\frac{\partial Q}{\partial \theta_u})U$ and $d_1, \ldots, d_R$ are the eigenvalues of the intensity matrix $Q(\theta)$, which are assumed to be distinct. Derivatives can also be calculated in the presence of a singular matrix $U$, but this is unlikely to arise at the mle in practical applications [70].

For misclassification HMMs calculation of derivatives is more difficult. However, misclassification HMMs also tend to be more computationally intensive to fit, so it is this situation where the score formulation of the residuals is most useful. The derivatives could be calculated simply by applying the product rule to the matrix expression for the likelihood (equation 1.4). This gives an expression of the form

$$\frac{\partial L(\theta)}{\partial \theta_u} = \frac{\partial \pi_{\mathbf{0}}}{\partial \theta_u} M_1 M_2 \ldots M_N \mathbf{1} + \sum_{j=2}^{N} \pi_{\mathbf{1}} \tilde{M}_{2j} \tilde{M}_{3j} \ldots \tilde{M}_{N,j} \mathbf{1} \tag{2.4}$$

where

$$\tilde{M}_{ij} = \begin{cases} M_i & i \neq j \\ \frac{\partial M_i}{\partial \theta_u} & i = j \end{cases},$$

$\frac{\partial M_i}{\partial \theta_u}$ has $(r, s)$ entry

$$\frac{\partial e_{s,O_i}}{\partial \theta_u} p_{rs}(t_i - t_{i-1}) + e_{s,O_i} \frac{\partial}{\partial \theta_u} p_{rs}(t_i - t_{i-1})$$

and $\pi_{\mathbf{0}}$ is the vector of initial occupation probabilities.

While this method is reasonably straightforward, it is computationally expensive: to calculate the contribution to the score vector for a subject who was observed $N$ times in a model with $M$ parameters, requires roughly $NM$ times the computation required to evaluate their likelihood.

Lystig and Hughes [92], dealing with discrete-time HMMs, adapted the Forward algorithm to allow for recursive calculation of both the first and second derivatives. The algorithm is far more efficient than direct differentiation and is easily extended to the continuous-time case. The Forward algorithm recursively calculates the likelihood for a subject. For a subject with observed states $O_1, \ldots, O_n$ at times $t_1, \ldots, t_n$, forward weights $\alpha_k(j)$ for observation number $k = 1, \ldots, n$ and state $j = 1, \ldots, R$.

$$\alpha_1(j) = \mathbb{P}(O_1, X_1 = j) = \pi_{0j} e_{j,O_1},$$

where $\pi_{0j}$ is the $j$th entry of $\pi_{\mathbf{0}}$, and subsequent forward weights are calculated recursively:

$$\alpha_k(j) = \mathbb{P}(O_1, \ldots, O_k, X_k = j) = \sum_{i=1}^{R} \alpha_{k-1}(i) e_{j,O_k} p_{i,j}(t_k - t_{k-1}).$$

Then the likelihood for that subject is given by

$$\mathbb{P}(O_1, \ldots, O_n) = \sum_{i=1}^{R} \alpha_n(i).$$

If we further define

$$\phi_k(\theta_u, j) = \frac{\partial \alpha_k(j)}{\partial \theta_u} = \frac{\partial}{\partial \theta_u} \mathbb{P}(O_1, \ldots, O_k, X_k = j)$$

then this allows $\phi_k(\theta_u, j)$ to be calculated recursively as

$$\phi_k(\theta_u, j) =$$

$$\sum_{i=1}^{R} \left( \phi_{k-1}(\theta_u, i) e_{j,O_k} p_{ij}(t_k - t_{k-1}) + \alpha_{k-1}(i) \frac{\partial e_{j,O_k}}{\partial \theta_u} + \alpha_{k-1}(i) e_{j,O_k} \frac{\partial p_{ij}(t_k - t_{k-1})}{\partial \theta_u} \right).$$

Then the first derivative of the likelihood for the subject is given as

$$\frac{\partial \mathbb{P}(O_1, \ldots, O_n)}{\partial \theta_u} = \sum_{i=1}^{R} \phi_n(\theta_u, i).$$

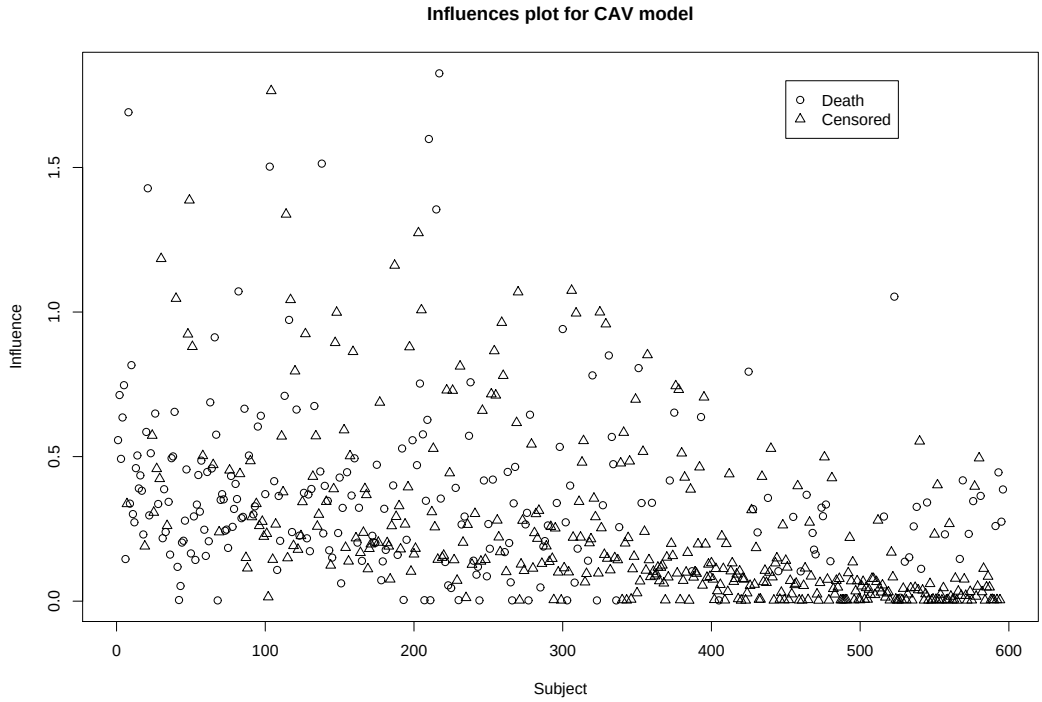
This is the basic idea of Lystig and Hughes' algorithm, though their version is modified to avoid *underflow*, meaning the problem of $\alpha_k(j)$ becoming exponentially small as $k$ increases. This is achieved by dividing through by $\sum_i P(O_{k-1} = i | O_1, \ldots, O_{k-2})$ at each stage $k$.

### Application to CAV and BOS data

For the CAV data without misclassification, refitting the model 596 times with a single patient deleted is feasible. As figure 2.12 shows, there are no subjects with particularly large influences. The pattern of influence generally bears a close resemblance to the amount of follow-up time, patients with higher numbers had transplants later and so were censored after transplantation sooner. Subjects for which a death was observed also tend to have higher influence.

Computing the 488 jackknife estimates for the BOS dataset would be considerably time consuming. Instead we can compute the individual score contributions. For the model with no misclassification covariates and misclassification permitted only to adjacent states, the influences plot identifies a clear set of outliers. In the bulk of cases, these anomalous subjects have an observation in state 3 followed by a series of observations in state 1. In

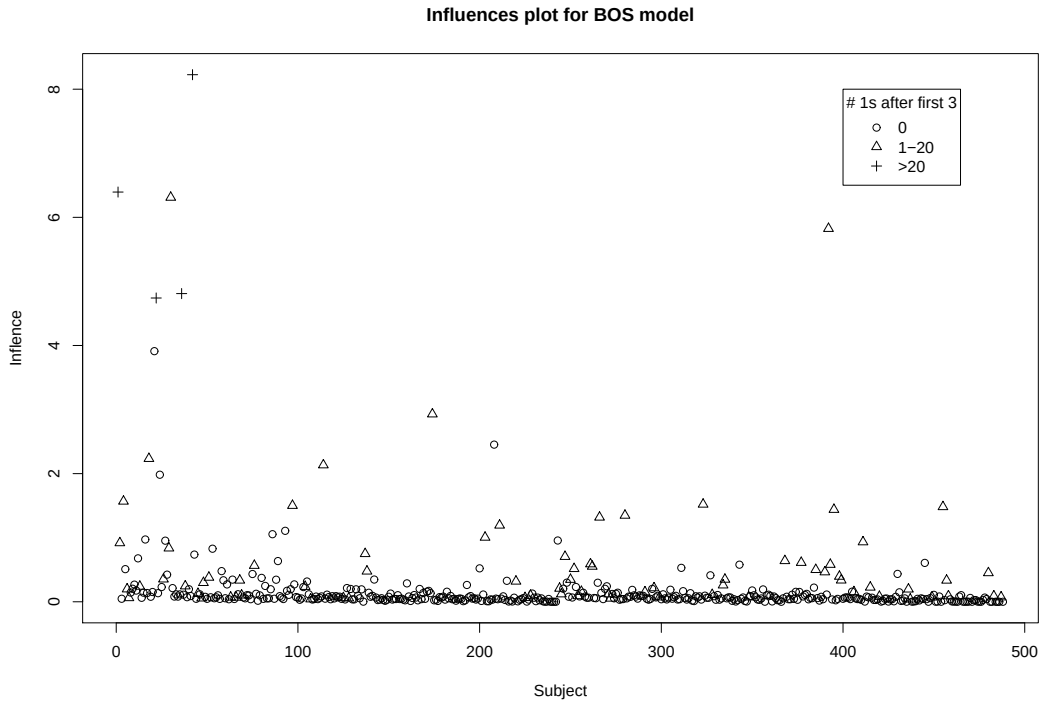
Figure 2.12: Plot of influence per subject for the CAV model on data without misclassification, calculated using jackknife estimates. Subjects are numbered in chronological order of transplant. Later patients contribute less influence because they are censored earlier.




a couple of cases a series of state 3 observations occur before a state 1. It is therefore the assumption that misclassification can only be to adjacent states which appears the most influential in determining the outliers. Figure 2.13 gives the influences plot. The point symbol for each subject is determined by the number of state 1 observations which occurred after the first state 3. As can be seen, many of the most extreme outliers fell into this category.

Aside from detecting outliers which are merely due to mistakes in the data, jackknife residuals also provide a way of identifying the subjects who may most contradict the assumptions of the model. If, as with the BOS model, there is a series of outliers who have similar histories, modifying the model assumptions to better accommodate these patients is likely to be an improvement. Jackknife residuals are not however without their limitations. They are only able to consider the influence of individual subjects. If the number of observations per subject is small, the potential influence for any particular

Figure 2.13: Plot of influence per subject for the BOS model on data with misclassification to adjacent states only, calculated using score contributions estimates.




subject is unlikely to be large. Systematic problems, affecting a group of subjects, may not be identifiable on an individual basis unless the correct method of grouping subjects in the plot is chosen.

### 2.4.2 Summary residuals

Kosorok and Chao (1996) [79] introduced a simple diagnostic for model appropriateness based upon *summary residuals*. They seek to construct random variables which, when the Markov model is correct, have zero mean and unit variance and are uncorrelated. Unlike jackknife residuals, the summary residuals refer to a particular observation of a patient rather than all the observations of a particular patient. This approach has some similarities with Pearson-type tests (discussed in section 1.4.4) as they also look at observed and expected quantities on the basis of individual pairs of observations.

Suppose a particular subject has observed states $x_0, x_1, ..., x_N$ at times $t_0, t_1, ..., t_N$.

Conditional on the last observed state, and assuming the sampling at time $t_{j+1}$ is independent of $X(t)$ between $t_j$ and $t_{j+1}$:

$$\mathbb{P}(X(t_{j+1}) = x_{j+1} | X(t_j) = x_j) = p_{x_j x_{j+1}}(t_{j+1} - t_j)$$

where $p_{r,s}(t)$ is the $(r, s)$ entry from the transition matrix $P(t; \theta)$ and $\theta$ are the parameters governing the Markov process.

Let $v = (1, 2, ..., R)^T$. Then define

$$r_j(\theta) = \frac{v^T \left( \mathbf{x}^*_{\mathbf{j+1}} - P(t_{j+1} - t_j; \theta) \right)}{\sigma_j(\theta)}$$

where $\mathbf{x}^*_{\mathbf{j+1}}$ is an $R$ dimensional vector with $k$th entry $\delta_{k,x_{j+1}}$, where $\delta_{ij} = 1$ if $i = j$ and is zero otherwise. Also,

$$\sigma_j^2(\theta) = v^T \left( \text{diag}(P(t_{j+1} - t_j; \theta)_{x_j}) \right) v - \left( v^T P(t_{j+1} - t_j; \theta)_{x_j} \right)^2$$

and $P(t; \theta)_r$ denotes the $r$th row of the transition matrix. If the true value of $\theta$, $\theta_0$, is known then the $r_j(\theta_0)$ will be independent, with mean zero and variance 1. Kosorok and Chao assert that the same properties will be approximately true when $\theta_0$ is replaced with $\hat{\theta}$, fitted from the data. Scatterplots of these residuals against quantities of interest within the model can be used to assess the appropriateness of assumptions within the model. In particular, the correctness of the functional form of the covariates can be tested. Typically it is assumed that there exists a log-linear relationship between a covariate and the transition intensity. Other diagnostics do not have the ability to test this assumption directly.

Summary residuals are not without their drawbacks however. Firstly, the resultant scatterplots can be very difficult to interpret. Typical output will be of the form of large clusters of data points in a series of rows, their number depending on the number of states in the model. From this it is necessary to judge whether there is any trend in the value of the residual. This is particularly difficult to achieve by eye. Kosorok and Chao computed a running mean-smoothed version of the summary residuals with an appropriately sized window width. However, except when the fit is either very good or very poor, it will still be difficult to tell whether the deviation of the running mean from the zero line is substantial.

Theoretically, if residuals really are uncorrelated, then crude pointwise confidence intervals for the running means can be provided by assuming convergence to normality. Lines with heights $\pm 1.96 n^{-0.5}$ where $n$ is the number of observations in the running average can then be taken as confidence bands.

A second problem is that the values of the residuals depend entirely on the labelling of states as they convert an ordinal state into a numerical value. The residuals are only coherent if the discrete states can be thought of as defining boundaries in some continuous measure, so that the distance between state $r$ and state $r+1$ is less than the distance between state $r$ and state $r+2$ by some well-defined and meaningful metric. For most disease models, provided we label our states $1, 2, 3, ...$ in correspondence with increasing severity of disease, we will usually have the required property. If however we chose some less natural labelling the residuals will have a less clear meaning. Moreover, for more general Markov models, where states cannot be interpreted as having well defined distance from one another, an appropriate labelling may not be possible.

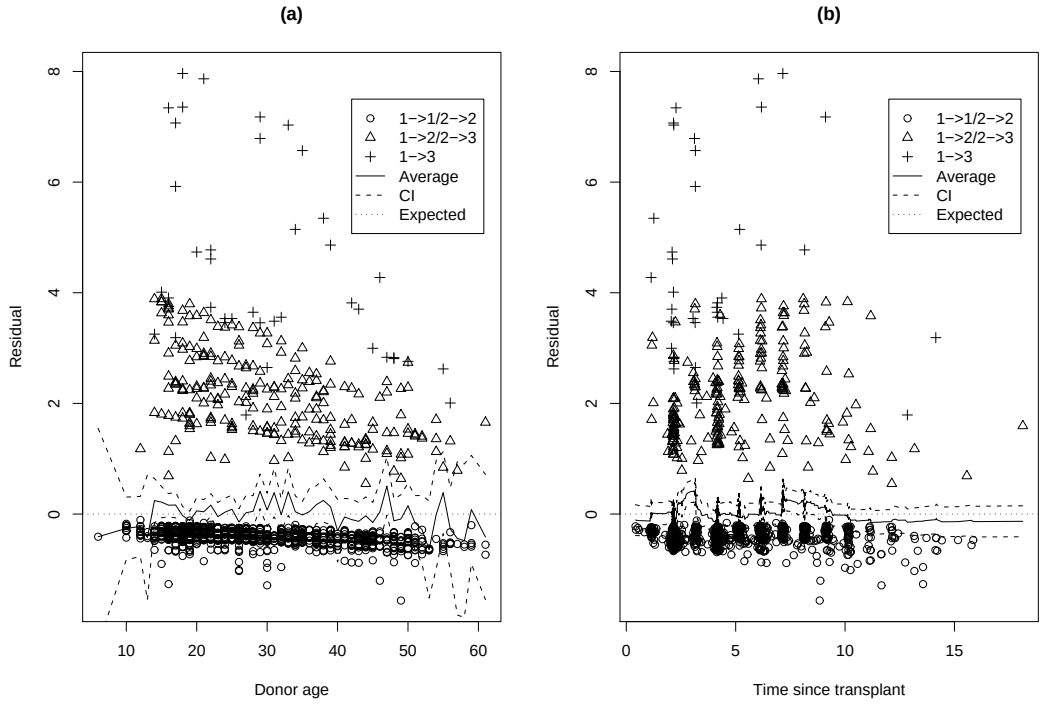
An additional drawback for the more general use of summary residuals is that they cannot be used for data with exact death times. To obtain $p_{rs}(t)$ it is necessary to know $t$. When a death was observed, $t$, the time an observation would have occurred had no death occurred, is not known.

**Example: CAV data**

Since the method is not applicable to data with exact deaths we cannot apply it directly to the CAV data. However, we can apply the method to a truncated version of the CAV dataset without misclassification, where state 3 is taken to be an absorbing state. This involves removing all transitions to death. Of primary interest is whether the assumption of log-linearity in the covariate effect of donor age is reasonable. This can be tested by plotting the summary residuals for each observation against donor age (figure 2.14(a)). This gives a plot with a series of horizontal clusters. The cluster just below zero corresponds to $1 \to 1$ or $2 \to 2$ transitions. The middle cluster corresponds to $1 \to 2$ or $2 \to 3$ transitions and most of the highest valued residuals are $1 \to 3$ transitions. A smoothed moving average is not really appropriate for the donor age covariate because there are only 51 unique values and some of them have over 100 observations. Instead the average value of the residuals for each distinct covariate value is plotted. As each residual should be approximately independent with mean zero and variance 1, these averages have approximate

mean zero and variance $n^{-1}$. The confidence intervals, assuming convergence to normality, contain 0 for all unique values of donor age except 29 years old. Moreover, there is no apparent trend in the value of the residuals. The plot therefore shows no evidence that a log-linear effect of donor age is inappropriate for the model.

Figure 2.14: Summary residuals for CAV data plotted (a) against donor age and (b) against time since transplant.




Time homogeneity can be assessed by plotting the residuals against the time at which the observation that the residual refers to took place. A running average can be plotted in this instance because most observation times have only 1 or 2 observations. Again there was no apparent trend in the residuals (figure 2.14(b)) which means there is no evidence against the assumption of time homogeneity in the model.

Summary residuals seem of limited use in assessing model fit. Having residual values which depend on the (arbitrary) state labelling could be problematic. The method's potential to allow assumptions about the functional form of the covariate effects to be tested is its main advantage. However, this might be better achieved by simply fitting a range of models with different functional forms for the covariate and using the log-likelihood to

choose between them. General goodness-of-fit can be better assessed using Pearson-type tests.

## 2.5 Tracking

Patient homogeneity, particularly conditional on known covariates, is a key assumption of a time homogeneous Markov model. In general 'random effects' models are beyond the scope of this thesis. To compute the likelihood for general random effects models it is necessary to either compute multi-dimensional integrals or make an approximation to a mixed Markov model [30]. However, an important special case is the 'tracking' model proposed by Satten [117]. 'Tracking' refers to the positive correlation between sojourn times in each state: under the model patients who progress rapidly through initial states are more likely to progress quickly through later states.

There is a baseline transition intensity matrix $Q_0$ for each patient, in addition each patient has a frailty parameter $Z_i$ which is an independent sample from some distribution $G(\phi)$ where $\phi$ is the parameter governing the variability of $G(\phi)$. Patient $i$'s transition intensity matrix is then $Q_0 Z_i$.

Clearly $G(.)$ must be a strictly positive distribution, and it is natural for $G(.)$ to have mean 1. The key factor in this model is that the individual frailties do not alter the eigen-decomposition of the transition intensity matrix except that the eigenvalues are each multiplied by $Z_i$. The transition probabilities of a time homogeneous Markov model with transition intensity matrix $Q_0$ can be written as

$$p_{rs}(t) = \sum_{k=1}^{R} u_{rk} u_{ks}^{-1} \exp(-d_k t)$$

where $\mathbf{d}$ is the vector of eigenvalues of $Q_0$ and $u_{rs}$ and $u_{rs}^{-1}$ are the $(r,s)$ entries to the matrix of eigenvectors $U$ and its inverse $U^{-1}$. The equivalent transition probability when the transition intensity matrix is multiplied by $z_i$ is

$$p_{rs}(t) = \sum_{k=1}^{R} u_{rk} u_{ks}^{-1} \exp(-d_k z_i t).$$

The likelihood contribution from a subject observed in states $x_0, x_1, \ldots, x_N$ at times $t_0, t_1, \ldots, t_N$ is

$$L_i^{\text{Markov}}(Q_0) = \prod_{l=1}^{N} p_{x_{l-1},x_l}(t_l - t_{l-1}) = \prod_{l=1}^{N} \left( \sum_{k=1}^{R} u_{x_{l-1},k} u_{k,x_l}^{-1} \exp(-d_k(t_l - t_{l-1})) \right).$$

This can be re-expressed as a multiple summation

$$L_i^{\text{Markov}}(Q_0) = \sum_{j_1=1}^{R}\sum_{j_2=1}^{R}\cdots\sum_{j_N=1}^{R}\left(\prod_{l=1}^{N} u_{x_{l-1},j_l}u_{j_l,x_l}^{-1}\right)\exp\left(-\sum_{l=1}^{N} d_{j_l}(t_l - t_{l-1})\right). \quad (2.5)$$

For bi-directional models, the number of terms in 2.5 increases exponentially with $N$, the number of observations. Keeping track of all the individual terms becomes computationally intractable. For progressive models, the eigenvalues of $Q_0$ are just its diagonal. Moreover, the transition probabilities can be written as

$$p_{rs}(t) = \begin{cases} \sum_{k=r}^{s} u_{rk}u_{ks}^{-1}\exp(-d_k t) & s \geq r \\ 0 & s < r \end{cases}$$

and so 2.5 becomes

$$L_i^{\text{Markov}}(Q_0) = \sum_{j_1=x_0}^{x_1}\sum_{j_2=x_1}^{x_2}\cdots\sum_{j_N=x_{N-1}}^{x_N}\left(\prod_{l=1}^{N} u_{x_{l-1},j_l}u_{j_l,x_l}^{-1}\right)\exp\left(-\sum_{l=1}^{N} d_{j_l}(t_l - t_{l-1})\right). \quad (2.6)$$

Hence the number of terms in 2.6 is manageable. The likelihood contribution in the tracking model is given by

$$L_i^{\text{Tracking}}(Q_0,\phi) = \int_{-\infty}^{\infty} L_i^{\text{Markov}}(Q_0 z)g(z,\phi)dz \quad (2.7)$$

where $g(z,\phi)$ is the probability density function for $G(\phi)$. As $L_i^{\text{Markov}}$ can be written as a sum of exponential terms, equation (2.7) is just a series of Laplace transforms. Hence we get

$$L_i^{\text{Tracking}}(Q_0,\phi) = \sum_{j_1=x_0}^{x_1}\sum_{j_2=x_1}^{x_2}\cdots\sum_{j_N=x_{N-1}}^{x_N}\left(\prod_{l=1}^{N} u_{x_{l-1},j_l}u_{j_l,x_l}^{-1}\right)h_\phi\left(\sum_{l=1}^{N} d_{j_l}(t_l - t_{l-1})\right) \quad (2.8)$$

where $h_\phi(x) = \int_{-\infty}^{\infty} g(z,\phi)\exp(-zx)dz$.

Satten suggested using an inverse Gaussian distribution for $G(.)$ so that

$$g(z,\phi) = \left(\frac{\phi}{2\pi z^3}\right)^{0.5}\exp\left(\phi - \frac{\phi(z+z^{-1})}{2}\right)$$

and

$$h_\phi(x) = \exp\left(\phi - [\phi(\phi+2x)]^{\frac{1}{2}}\right). \quad (2.9)$$

Under this parametrisation, $\phi$ refers to the precision of $G(.)$, so the time homogeneous Markov model is equivalent to the case $\phi = \infty$.

In the context of model diagnostics it will be of most interest to test whether the tracking model is significantly better at explaining the data than a time homogeneous Markov model. The tracking model can be fitted using maximum likelihood estimation. A likelihood ratio test can then be used to compare the tracking model with the Markov model. However, since the Markov model corresponds to $\phi = \infty$, this involves a test at the boundary point of the parameter space and so Wilks' theorem does not apply. Self and Liang [118] showed that the likelihood ratio statistic is a 50:50 mixture of chi-square distributions with 0 and 1 degrees of freedom in this situation.

### 2.5.1 Extension for exact death times

Satten only dealt with the case of unidirectional models for entirely interval censored data. In the previous section it was shown that tracking can be applied to all types of progressive model. Another important extension is the case of exact death times. For a progressive model the likelihood contribution of an exact death is

$$\sum_{k=r}^{R-1} p_{rk}(t) q_{kR}.$$

For a subject $i$ observed in states $x_0, x_1, \ldots, x_{N-1}, x_N = R$ at times $t_0, t_1, \ldots, t_N$, if we condition on the state $k$ from which the subject entered death we have that

$$L_i^{\text{Markov}}(Q_0) = \sum L^{(k)} q_{kR}$$

where $L^{(k)}$ is the likelihood of a sequence of states $x_0, x_1, \ldots, x_{N-1}, x_N = k$ and is given by equation (2.6). To calculate the likelihood under tracking it is therefore necessary to compute integrals of the form

$$\tilde{h}_\phi(x) = \int_{-\infty}^{\infty} z g(z, \phi) exp(-zx) dz.$$

When $g(z, \phi)$ is the p.d.f. of an inverse Gaussian distribution, we get

$$\tilde{h}_\phi(x) = (\frac{\phi}{(\phi - 2x)})^{\frac{1}{2}} \exp\left(\phi - [\phi(\phi + 2x)]^{\frac{1}{2}}\right).$$

A subject who is observed in states $x_0, \ldots, x_{N-1}$ at times $t_0, \ldots, t_{N-1}$ and dies at time $t_N$ therefore has a likelihood contribution of

$$\sum_{j_N=x_{N-1}}^{R-1} \left( \sum_{j_1=x_0}^{x_1} \sum_{j_2=x_1}^{x_2} \cdots \sum_{j_{N-1}=x_{N-2}}^{x_{N-1}} \left( \prod_{l=1}^{N-1} u_{x_{l-1},j_l} u_{j_l,x_l}^{-1} \right) q_{j_N,R} \tilde{h}_\phi \left( \sum_{l=1}^{N-1} d_{j_l}(t_l - t_{l-1}) \right) \right) \tag{2.10}$$

under the tracking model.

### 2.5.2 Other possible extensions

It would be desirable to allow the test for tracking to be extended to the cases of bi-directional models and misclassification HMMs. For misclassification HMMs, even when the underlying model is progressive, standard recursive methods of evaluating the likelihood such as the Forward algorithm cannot be applied. This means it would be necessary to sum over every sequence of underlying states directly. However, effective approximate likelihoods can be found by approximating the random effect distribution $g(z, \phi)$, by a discrete distribution. As mentioned in section 1.4.5, Cook *et al* [30] used this technique to allow the transition intensities to have separate, but potentially correlated, random effects distributions.

### 2.5.3 Similarity with time inhomogeneity

In section 1.4.2 a method of fitting a simple time inhomogeneous Markov model with transition intensity matrix $Q(t) = Q_0 g(t; \mu)$, for some non-negative scalar function $g(t; \mu)$ was outlined. The tracking model has some similarities to this. Satten noted that the tracking model may be sensitive to departures from Markov behaviour that may be different from tracking. Equally, Lancaster and Nickell [81] noted that the omission of covariates can lead to apparent time inhomogeneity in a time homogeneous process.

The effect of tracking on the population-wide survival curve compared to the survival curve of an analogous time homogeneous Markov model is for the most frail subjects to die quickly leaving less frail patients alive. Hence the transition intensities for the population as a whole appear to be decreasing with time. This is the same effect as occurs for a homogeneous patient group who are subject to a time inhomogeneous Markov process. The tracking model involves a subject's sojourn times in each state being positively correlated. However, this positive correlation will also be apparent in the time inhomogeneous model

with a monotonically decreasing intensity matrix. A subject who progresses through state 1 quickly, is more inclined to also progress quickly through state 2 because they have reached state 2 early while $q_{22}g(t;\mu)$ is of a greater magnitude.

The tracking and time inhomogeneous models can never be exactly the same. For instance, consider a three-state disease model, with fixed baseline rates $q_{12}$, $q_{13}$, $q_{23}$. Let $q_1 = q_{12} + q_{13}$. For a time homogeneous model, the survival estimate at time $t$ is given by

$$p_{13}(t) = \frac{q_{13} - q_{23}}{q_1 - q_{23}} \exp(-q_1 t) + \frac{q_{12}}{q_1 - q_{23}} \exp(-q_{23} t).$$

As explained in section 1.4.2, the transition probabilities for a time inhomogeneous model with $Q(t) = Q_0 g(t;\mu)$ can be found by changing the operational time. This gives a survival estimate at time $t$ given occupancy in state 1 at time zero of

$$p_{13}(0,t) = \frac{q_{13} - q_{23}}{q_1 - q_{23}} \exp\left(-q_1 \int_0^t g(s;\mu)ds\right) + \frac{q_{12}}{q_1 - q_{23}} \exp\left(-q_{23} \int_0^t g(s;\mu)ds\right).$$

In contrast, a tracking model, where for individual $i$, $Q_i = Q_0 z_i$ where $z_i$ is a sample from a frailty distribution $G(x;\phi)$ with Laplace transform $h_\phi(x)$, has a survival curve

$$\begin{aligned} p_{13}(0,t) &= \int_{-\infty}^{\infty} \left( \frac{q_{13} - q_{23}}{q_1 - q_{23}} \exp(-q_1 z t) + \frac{q_{12}}{q_1 - q_{23}} \exp(-q_{23} z t) \right) dG(z;\phi) \\ &= \frac{q_{13} - q_{23}}{q_1 - q_{23}} h_\phi(q_1 t) + \frac{q_{12}}{q_1 - q_{23}} h_\phi(q_{23} t). \end{aligned}$$

$\exp(-q_1 \int_0^t g(s;\mu)ds)$ is a function of $q_1$ and $t$ separately, not merely of $q_1 t$, hence the two models cannot coincide. Nevertheless the two models are sufficiently similar that if a significant improvement in likelihood against the time homogeneous Markov model is achieved through the tracking model, it is also likely to be achieved by the time inhomogeneous model. It will also be very difficult to correctly choose between the two models except when the sample size is very large.

From the perspective of merely identifying departures from a time homogeneous Markov model, testing for simple time inhomogeneity may be the more effective of the two tests. This is because time inhomogeneity can also accommodate the case where the intensities appear to be increasing with time, whereas tracking can only result in a decreasing hazard.

### Application to CAV data without misclassification

Using the extension for exact deaths, the model for the CAV data without misclassification (see section 2.1.2) can be tested for tracking. The tracking model was no improvement

on the Markov model (the Markov model, which is the limit as $\phi \to \infty$ optimised the likelihood). If we apply the simple test of time inhomogeneity to the same data, using $g(t;\mu) = \exp(t\mu)$, there is a non-significant improvement in likelihood (likelihood ratio statistic $T = 3.28$ compared to $\chi^2_1$, $p = 0.07$) but the point estimate gives mildly increasing transition intensities ($\hat{\mu} = 0.023$). Put into this context it was unlikely that there would be any support in the data for the tracking model.

## 2.6 Specific methods for misclassification HMMs

So far in this chapter we have dealt with techniques whose primary application is with Markov models. Whilst we have shown that in many cases they can be generalised to misclassification HMMs, there are also some diagnostic methods which have been proposed with misclassification HMMs in mind.

### 2.6.1 Prediction of future observations table

Satten and Longini [115] took the 'ability of the model to predict the next observation' as a criterion for model fit. As with prevalence counts, this method involves the comparison of observed and expected counts. However, rather than considering the counts at specific time points, they are instead averaged over a time period. This creates a *prediction of future observations* table. A subject observed $n$ times within an interval of interest with $n_1, \ldots, n_R$ observations in $1, \ldots, R$ respectively would have an observed contribution of $\frac{n_j}{n}$ in state $j$. The expected count is calculated by considering probabilities of the form $\mathbb{P}(O_i|O_1, \ldots, O_{i-1})$, meaning all previous observations are taken into consideration. These probabilities can be calculated using

$$\mathbb{P}(O_i|O_1, \ldots, O_{i-1}) = \sum_{r}^{R} \mathbb{P}(O_i|X_i = r)\mathbb{P}(X_i = r|O_1, \ldots, O_{i-1})$$

where $\mathbb{P}(X_i = r|O_1, \ldots, O_{i-1})$ can be obtained by taking the normalised vector

$$\boldsymbol{\pi_0} M_1 M_2 \ldots M_{i-1},$$

where $\pi_0$ and $M_j$ follow the notation of section 1.2.4. When a subject is observed $n$ times within the interval of interest, each of these probabilities is given a $\frac{1}{n}$ weighting. As with prevalence counts, discrepancy between the observed and expected counts can be taken as informal evidence of poor fit, but formal tests of fit cannot be carried out.

Table 2.13: Prediction of future observations table for BOS model.

| | | Time period | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| State | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | Obs | 303.2 | 213.7 | 138.3 | 107.1 | 74.3 | 57.2 | 38.8 | 28.4 | 21.7 | 15.3 |
| | Exp | 325.4 | 199.9 | 126.8 | 94.2 | 65.3 | 45.6 | 31.9 | 22.7 | 18.0 | 11.7 |
| 2 | Obs | 65.3 | 55.5 | 47.9 | 45.6 | 39.7 | 31.1 | 20.8 | 17.3 | 13.4 | 9.0 |
| | Exp | 54.9 | 60.2 | 51.0 | 46.0 | 35.4 | 32.9 | 21.5 | 15.4 | 11.5 | 8.2 |
| 3 | Obs | 52.5 | 69.8 | 75.7 | 75.2 | 65.0 | 58.7 | 51.4 | 35.3 | 22.9 | 17.8 |
| | Exp | 40.7 | 79.0 | 84.1 | 87.9 | 78.3 | 68.5 | 57.6 | 42.9 | 28.6 | 22.1 |

The advantage of this method over using prevalence counts is that it is more appropriate for irregularly spaced observation times. In addition, whereas prevalence counts primarily test the underlying Markov model, poor fit from a prediction of future observations table is more likely to be due to problems with the link between the underlying and observed processes. However, this approach may miss some systematic lack of fit because reference to the previous observed state in the categorisation is not made.

A further drawback is that the table cannot be constructed when death times are observed exactly for the same reasons as for summary residuals in section 2.4.2. This problem can be avoided by excluding deaths from the table and reweighting the expected probabilities so that they are correct, conditional on a death not having occurred.

### BOS data

The potential lack of power is apparent if we apply the method to a model for the BOS data. Using the model in which misclassification to adjacent states is permitted and the presence of acute events is a covariate on misclassification gives table 2.13. There is some disagreement particularly in the first year, where there are more state 2 and 3s observed than expected and fewer state 1s. Elsewhere, although there are some moderate areas of disagreement, there is weak evidence of poor fit. This contradicts other assessments of fit for this model, which suggest a poor or very poor fit.

Table 2.14: Example observation times and states for a subject in a 2 state model. The bottom two rows give the contributing times to the $1 \to 2$ and $2 \to 1$ plots respectively. The starred time is a censored contribution. $\times$ marks entry times into states.

| Time | 0 | 0.7 | 1 | 1.4 | 1.9 | 2.5 | 3 |
|---|---|---|---|---|---|---|---|
| State | 1 | 2 | 1 | 1 | 2 | 1 | 1 |
| $1 \to 2$ | $\times$ | 0.7 | $\times$ | | 0.9 | $\times$ | 0.5* |
| $2 \to 1$ | | $\times$ | 0.3 | | $\times$ | 0.6 | |

### 2.6.2 Bureau et al plots

Bureau *et al* [14] in the analysis of data from a two-state recurrent model, proposed calculating an empirical estimate of the waiting time between being first observed in one of the states to being observed in the other state. The plot for state 1 is constructed by taking the first time at which a subject is observed in state 1 as its initiation time. The event time is then the time elapsed when the subject is observed in state 2. The subject is censored if it is observed to remain in state 1. A particular subject may contribute more than one set of times if observed to exit and subsequently return to state 1, the time at which they are observed to return to state 1 is taken as a new initiation time. The distribution is then calculated by a Kaplan-Meier product-limit estimate.

For instance consider a subject with observation times and states as shown in table 2.14. The plot for 1→2 transitions will have contributions of two events at times 0.7 and 0.9 and a censored observation at 0.5. The plot for 2→1 will have two events at times 0.3 and 0.6.

Bureau *et al* only considered a two-state case. However, we can generalise the method to allow for multiple states. To do this a plot is constructed for every pair of states for which an observed transition is possible. The curves are constructed in a similar way as the two-state case. When the plot represents a forward transition an event is taken to have occurred if a state greater than or equal to the destination state of interest is observed. When the plot represents a backward transition an event is taken to have occurred if a state less than or equal to the destination state of interest is observed.

The shape of the curves reflect the underlying latent process, the observed process and the sampling scheme of the data. Curves can be compared to the curves predicted by

the fitted model. Since the shapes of the curves are dependent on the sampling scheme, these expected curves can only be determined by simulating many sets of states at the observation times for the existing data. Bureau *et al* considered the fit of the model by comparing the observed curve with a curve based on a single simulated dataset. However, we note that, if many simulated datasets are generated, approximate $100(1-\alpha)\%$ pointwise confidence intervals can be constructed by ordering the simulated curves and taking the $100(\frac{\alpha}{2})\%$ and $100(1-\frac{\alpha}{2})\%$ points. These confidence intervals will tend to have higher than $100(1-\alpha)\%$ pointwise coverage because the simulation does not take into account the fitting of the model to the data.

A problem may occur if the observation times are dependent on the previous state ("doctors care"). Unless this is incorporated into the simulated data, disagreement in the plots may be due to different sampling distributions rather than inaccuracy in the model.

Despite being quite *ad hoc* in their construction, Bureau *et al* plots do provide an effective informal method of assessing goodness-of-fit for misclassification models. However, the majority of their power is in detecting when misclassification of states is not time independent. Bureau *et al* plots can also be applied to Markov models which do not have misclassification, particularly if the model has reverse transitions.

### CAV data example

The result of applying these plots to the model for CAV data with misclassification is given in figure 2.15. Evidence of the sampling scheme is apparent in the plots, particularly in the steep gradients around 2 and 4 years, which is the typical time between observations. The curve for $1 \rightarrow 4$ corresponds entirely to the curve from section 2.2.1. The plots show generally good agreement as the observed curves stay within the 90% pointwise confidence limits, except for some minor deviations for $2 \rightarrow 1$ and $3 \rightarrow 2$ transitions. These provide some evidence against the assumption of independent misclassification. Given an observation in state 2, the probability of subsequently being observed in state 1 is lower than the fitted model would suggest.

### BOS data example

In contrast to the CAV data, the plots for the BOS dataset, with the model allowing misclassification only to adjacent states (figure 2.16), betray a clear lack of fit. None of the

plots suggest a good fit. In particular the $2 \rightarrow 1$ and $3 \rightarrow 1$ plots have the observed curve consistently higher than expected. This implies reverse transitions occur less frequently than they should do according to the fitted model and this provides stronger evidence against the assumption of independent misclassification.

### 2.6.3 Tests for independent misclassification

One of the key assumptions of a hidden Markov model is that

$$O_1|X_1, \ldots, O_N|X_N$$

are independent. Often, the observed process is subject to fluctuations. The validity of the HMM usually rests upon the assumption that these fluctuations occur on a shorter time scale than the frequency of observations. In this section, two simple tests for testing this assumption are developed.

For models where the underlying state is progressive, it is sometimes possible to identify a subset of the data for which a $\chi^2$ test can be applied to directly. For this to be feasible, there must be observed state patterns which can entirely determine the true state for a period of time. For instance, in a progressive disease model where misclassification is only possible to adjacent states we can be sure that, according to the model, if the first observation in state $(r+1)$ occurs before the last observation in state $(r-1)$, the true state is $r$ between these times. Therefore, the observed states strictly between these two observations should be independent and identically distributed multinomial.

#### Example: BOS data

We can apply this method to the BOS dataset for a model in which only misclassification to adjacent states is possible. If a subject is observed in state 1 after their first observation in state 3 then between the time of their first observation in state 3 and their last in state 1, the model assumes they are in state 2. For the BOS data, the last state 1 observation occurs after the first state 3 in 80 of 488 subjects and there were 1135 intermediate observations. For these 33.7% were observed in state 1, 29.4% in state 2 and 36.8% in state 3. This alone disagrees heavily with the estimated misclassification probabilities in state 2 which are 19.6% observed in state 1, 53.0% in state 2 and 27.4% in state 3. If observations are categorised according to current observed state versus previous observed state, a $\chi^2$ test can be performed. Given that the true state is 2, the distribution of the

Table 2.15: Contingency table for Chi-squared test on observed states when model assumes true state occupancy is 2, for BOS data.

| | Observed State | | | |
|---|---|---|---|---|
| Previous State | 1 | 2 | 3 | Total |
| 1 | 271 | 66 | 10 | 347 |
| 2 | 75 | 179 | 61 | 315 |
| 3 | 37 | 89 | 347 | 473 |
| | Expected State | | | |
| Previous State | 1 | 2 | 3 | Total |
| 1 | 117.1 | 102.1 | 127.8 | 347 |
| 2 | 106.3 | 92.7 | 116.0 | 315 |
| 3 | 159.6 | 139.2 | 174.2 | 473 |

observed state should be independent of the previous state. For this example therefore, we can construct a contingency table with three rows (corresponding to the three possible previous states). Table 2.15 gives the contingency table. Note that the expected counts are based only on the assumption of homogeneity between rows and the observed counts and not on the fitted misclassification probabilities. A $\chi^2$ test of homogeneity between rows gives a statistic of 722.9. Asymptotically this test has a $\chi^2_4$ distribution. Therefore there is a clear lack of fit.

The simplest explanation for this lack of fit would be that observations in state 3 from state 1, or state 1 from state 3 are possible.

For more general models, there will not be any pattern of observed states which ensure the true state is known. Moreover, the above test, while clearly effective for the BOS data, does not have good general power because it is restricted to data from a subset of subjects. A more effective test can be achieved by using the observed states as a misclassification covariate.

### Using observed states as a misclassification covariate

To test the conditional independence assumption in a more general way, we assume an alternative model where the misclassification probability at observation $j$ is affected by the observed state at observation $j-1$, taken to be a categorical covariate. Thus the

misclassification probabilities at time $t_i$ are defined as

$$\text{logit}(e_{rs}(t_i)) = \alpha_r + \beta_r \mathbb{1}\{\text{Acute Event}\} + \sum_{j=1}^{R-1} \gamma_{jr} \mathbb{1}\{O(t_{i-1}) = j\}$$

where $\gamma_{jr}$ determine the effect of the previous observed state $j$ given current occupancy in true state $r$. The previous state at the first observation time is assumed to be 1. If the assumption of independent misclassification is correct, the value of this covariate coefficient should be zero. Significance can be tested using a likelihood ratio test. A significant result might also occur if other aspects of the model are incorrect such as time dependent transition intensities or a non-Markov underlying process.

The model fitted does not, in itself, represent a reasonable model for dependent misclassification. We would expect correlation between observed states (conditional on the true underlying state) to decay over time, whereas in the previous observed state model, the dependency on the last state is the same regardless of the time elapsed since the previous observation.

**Example: BOS Data**

We can apply this test to the BOS data, in a model in which misclassification from state 1 to state 3 is permitted. The standard model with acute events as a misclassification covariate has -2 × LL = 12312.7. Allowing previous observed state to be a covariate for misclassification introduces 10 additional parameters. The resulting model has -2 × LL = 11236.9. This gives a likelihood ratio statistic of $T$=1075.8 compared to $\chi^2_{10}$. Clearly, therefore there is strong evidence against independent misclassification. Inspection of the fitted parameters suggests that there is a strong tendency for the current observed state to be the same as the previous state (table 2.16). For instance, the table shows that if at time $t_j$, $X(t_j) = 1$ and there is no acute event, then if $O(t_{j-1}) = 1$, there is only a probability of 0.028 that $O(t_j) = 2$. In contrast if $O(t_{j-1}) = 2$, meaning the previous state was misclassified, the probability that $O(t_j) = 2$ rises to 0.180. Acute events continue to have the effect of increasing the probability of misclassification to higher states and decreasing the probability of misclassification to lower states.

Table 2.16: Estimated misclassification probabilities for BOS dataset

| Parameter | Standard Model | Previous state model previous state | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| $\mathbb{P}(O = 2\|X = 1, \text{No Acute})$ | 0.032 | 0.028 | 0.180 | 0.105 |
| $\mathbb{P}(O = 3\|X = 1, \text{No Acute})$ | 0.005 | 0.006 | 0.014 | 0.212 |
| $\mathbb{P}(O = 1\|X = 2, \text{No Acute})$ | 0.108 | 0.471 | 0.089 | 0.038 |
| $\mathbb{P}(O = 3\|X = 2, \text{No Acute})$ | 0.277 | 0.070 | 0.181 | 0.721 |
| $\mathbb{P}(O = 2\|X = 3, \text{No Acute})$ | 0.011 | 0.568 | 0.278 | 0.001 |
| $\mathbb{P}(O = 2\|X = 1, \text{Acute})$ | 0.148 | 0.103 | 0.488 | 0.318 |
| $\mathbb{P}(O = 3\|X = 1, \text{Acute})$ | 0.022 | 0.018 | 0.045 | 0.463 |
| $\mathbb{P}(O = 1\|X = 2, \text{Acute})$ | 0.061 | 0.264 | 0.038 | 0.016 |
| $\mathbb{P}(O = 3\|X = 2, \text{Acute})$ | 0.475 | 0.180 | 0.389 | 0.882 |
| $\mathbb{P}(O = 2\|X = 3, \text{Acute})$ | 0.006 | 0.283 | 0.103 | 0.000 |

## 2.7 Conclusion

This chapter has presented a diverse range of methods for assessing model fit. Comparisons with empirical estimates, particularly of the survival curve, were shown to be a useful informal diagnostic measure of fit. However, the presence of covariates makes comparison more difficult. Also this approach does not directly consider components of the model relating to intermediate states. Prevalence counts attempt to compare the expected prevalence in the states with an empirical estimate of the observed prevalence. However, this empirical estimate is quite crude, meaning that, unless subjects' observations are frequent, misleading results may occur. This is particularly true of the tabular form. The graphical generalisation developed in this chapter makes the potential bias in estimated observed prevalence easier to see. Prevalence counts for misclassification HMMs are also shown to be useful, at least when observations are frequent.

Jackknife residuals, to determine the influence of a particular subject, can be used to identify outliers, and possibly identify assumptions of the model which do not seem to be true. When refitting the model with each subject removed is impractical, considering the weighted score contribution of a subject, provides a way of identifying subjects with extreme influence. Whilst the summary residuals method has potential usefulness in

identifying when assumptions about the functional form of covariates are not valid, the resulting plots are quite difficult to interpret. Moreover, the method involves converting the ordinal states into a numerical value, which is dependent on the arbitrary labelling of states.

The tracking model provides a simple frailty model to test the assumption of patient homogeneity (possibly conditional on covariates). The approach can be applied to progressive models, including datasets where exact death times are known. However, a tracking model closely resembles a time inhomogeneous Markov model with decreasing transition intensities. As a general diagnostic, it may therefore be more appropriate to test for time inhomogeneity, as this allows for both increasing and decreasing intensities.

The chapter also considered methods for hidden Markov models. The prediction of future observations method, which compares observed and expected states, averaged over a time period, was shown to be quite poor at detecting lack of fit. In contrast the plots proposed by Bureau *et al* [14], whilst *ad hoc* in construction, were quite effective at identifying poor fit. They can only however, give an informal assessment. They also require simulation of the process, including the sampling scheme. Misspecification of the sampling scheme could potentially cause a spurious deviation between observed and simulated plots. The chapter also presented a simple way of formally testing the assumption of independent misclassification in HMMs, through a likelihood ratio test. The alternative model assumes the current observed state depends both on the current true state and the previous observed state. Whilst this is quite an unrealistic model, the method was shown to be effective at identifying dependent misclassification in the BOS model.

Application of the diagnostics in this chapter to models fitted to the CAV data, have given few indications that the fit is inadequate. The overall Kaplan-Meier product-limit estimate of the survival curve was shown to be in good agreement with the survival curve from the fitted time homogeneous Markov model. There was also no evidence of tracking. Application of the plots suggested by Bureau *et al*, did give some indication that consecutive observed states were more likely to be the same than predicted by the HMM.

In contrast, the HMM for the BOS dataset was shown clearly to be a poor fit in virtually all the diagnostics applied. In particular, the test for dependent misclassification and the Bureau *et al* plots showed that consecutive observed states were significantly more likely to be the same than predicted by the HMM. The observed states in the BOS dataset are obtained through the discretisation of a continuous $FEV_1$ measure. For this particular

dataset it may be more appropriate to deal with the raw $FEV_1$ counts, rather than discrete states. Moreover, the general suitability of a HMM for these data is questionable.

While many of the methods are useful in aiding the assessment of fit, they all have the disadvantage of either being informal or only formally testing one specific aspect of the model. Such tests certainly have their place, however, there is also a clear need for more general formal tests of model fit. Such tests are the focus of chapter 3.

Figure 2.15: Bureau *et al* plots for the CAV data. Observed line in bold, 90% point-wise confidence intervals dashed.

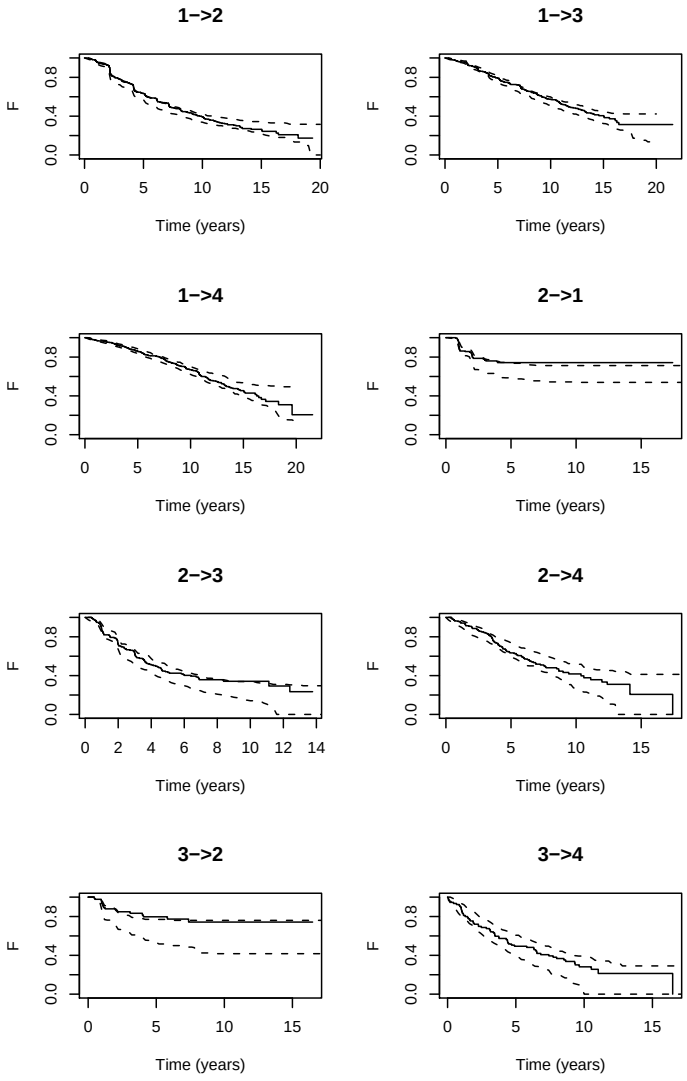
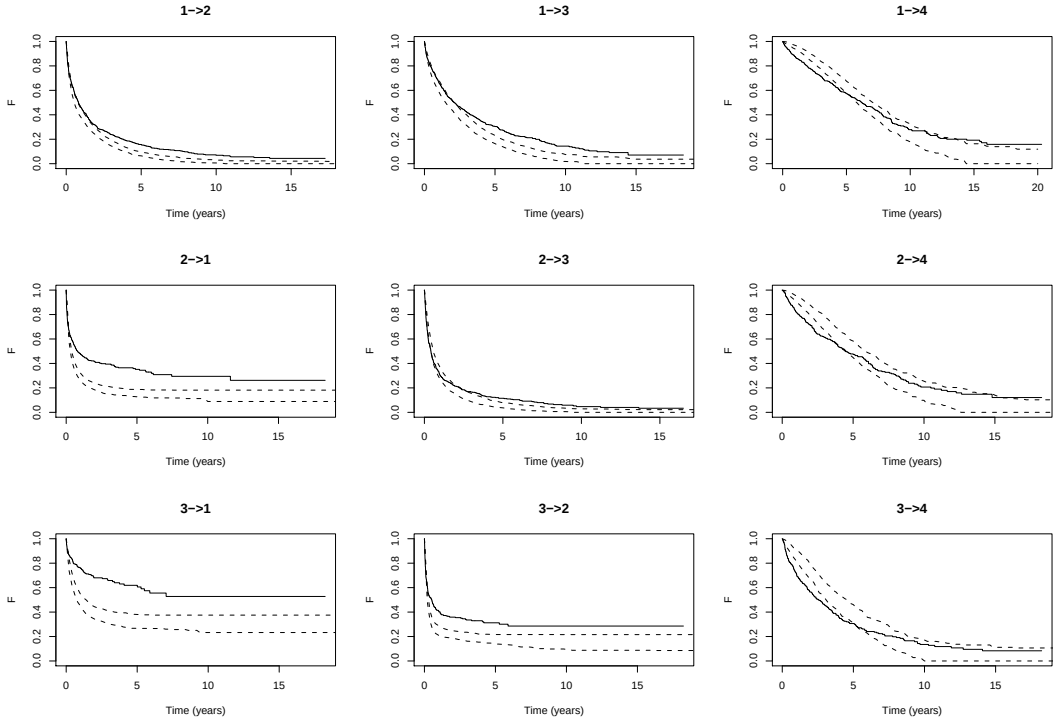


Figure 2.16: Bureau *et al* plots for the BOS data. Observed line in bold, 90% point-wise confidence intervals dashed.



# Chapter 3

# Pearson-type Goodness-of-fit tests

This chapter focuses on the construction of formal tests for general goodness-of-fit in Markov models for panel data. The existing methodology of Aguirre-Hernández and Farewell (AH/F) [6] for irregular sampling schemes and continuous covariates is extended. Firstly the null distribution of the AH/F statistic is explored and a method of getting a better asymptotic approximation than $\chi^2$ is provided. The method is also extended to the class of misclassification hidden Markov models. However, it is shown that in the common situation where the time into the absorbing state of the model (e.g. death) is known exactly, AH/F cannot be applied. The remainder of the chapter presents a way of allowing for exact death times by modelling the sampling time of observations and using a modified statistic based on this principle.

## 3.1 Pearson chi-squared tests for balanced observations

Formal approaches to a general goodness-of-fit test in Markov models have compared observed with expected transition counts based upon the model. Kalbfleisch and Lawless [70] and de Stavola [128] dealt with the case where all patients were observed at the same times, $t_0, t_1, \ldots, t_N$ (balanced observation), and there was a binary or categorical covariate. Let each individual, $i$, have process $X_i(t)$. In this situation, methods for hypotheses tests for discrete time Markov chains [10] can be easily extended. The transition counts can be grouped according to the observation number and the covariate value. A likelihood ratio test for the time homogeneous Markov model against a general alternative, where $j$ is the observation number (i.e. refers to the $j$th observation for a particular individual), $c$ the

covariate group, $r$ the observed state at the start of the interval and $s$ the state observed at the end of the interval, is given by

$$\Lambda = 2\sum_{j}\sum_{c}\sum_{r}^{R}\sum_{s}^{R} o_{jcrs}\log\left(\frac{o_{jcrs}}{e_{jcrs}}\right)$$

where

$$o_{jcrs} = \sum \mathbb{1}[X_i(t_j) = s, X_i(t_{j-1}) = r] \tag{3.1}$$

$$e_{jcrs} = \sum \mathbb{P}(X_i(t_j) = s | X_i(t_{j-1}) = r]\mathbb{1}[X_i(t_{j-1}) = r] \tag{3.2}$$

where $\mathbb{1}(A)$ is an indicator function for an event $A$ and the summation is over all individuals who have the covariate value $c$ and are observed at time $t_j$. In this case the expected counts, $e_{jcrs}$, can be calculated as just

$$e_{jcrs} = p_{rs}(t_j - t_{j-1}; \hat{\theta})n_{jcr}$$

where $n_{jcr}$ is the number of individuals with covariate value $c$ observed in state $r$ at time $t_{j-1}$, who have an observation at time $t_j$ and $\theta$ is the vector of model parameters. Moreover the counts $o_{jcrs}$ conditional on $n_{jcr}$ have a multinomial distribution. This likelihood ratio test therefore has an asymptotic null distribution which is $\chi^2$, with degrees of freedom given by $C - |\theta|$, where $C$ is the number of independent cells from the resultant contingency table and $|\theta|$ is the number of unknown parameters fitted from the data. It is also the case that the Pearson chi-squared statistic

$$X^2 = \sum_{j}\sum_{c}\sum_{r}^{R}\sum_{s}^{R}\frac{(o_{jcrs} - e_{jcrs})^2}{e_{jcrs}}$$

is asymptotically equivalent to $\Lambda$, having the same asymptotic null distribution [70].

Regular sampling times may arise in clinical trials or in some experimental studies with very strict sampling schemes. However, in many cases subjects are not observed at the same time points, either by design or due to clinical or subject constraints. Similarly covariates may be continuous, or take too many values for a grouping by distinct value to be meaningful. Coping with this scenario was the motivation for a test proposed by Aguirre-Hernández and Farewell.

## 3.2 The Aguirre-Hernández and Farewell test for irregular sampling schemes

Throughout the next section the notation relating to individual $i$ is suppressed. Aguirre-Hernández and Farewell [6] (AH/F) proposed a Pearson-type test that allows unique, irregular sampling times for each patient and also continuous covariates. Observations are categorised by observation number into observation categories, $h$, and, within each observation category, by time interval category, $l_h$. Additionally, observations are categorised by covariate category, $c$, according to quantiles of the estimated transition intensity $q_{rs}$. Then, for each transition type, $r \to s$ for a patient with observations at times $t_j$, $j = 1, \dots, n$, we calculate:

$$o_{hl_hrsc} = \sum \mathbb{1}[(X(t_{j+1}) = s, X(t_j) = r)] \tag{3.3}$$

$$e_{hl_hrsc} = \sum \mathbb{P}(X(t_{j+1}) = s | X(t_j) = r) \mathbb{1}[X(t_{j-1}) = r] \tag{3.4}$$

where the summation is over the set of observations:

$$\forall \text{patients}, j : t_{j+1} - t_j \in l_h, q(\mathbf{v}) \in c \tag{3.5}$$

where $\mathbf{v}$ is the vector of covariates for a patient.

Then the proposed statistic is given by:

$$T = \sum_h \sum_{l_h} \sum_r \sum_s \sum_c \frac{(o_{hl_hrsc} - e_{hl_hrsc})^2}{e_{hl_hrsc}}. \tag{3.6}$$

As shown in section 3.1, the analogous Pearson chi-square statistic has a null distribution which has degrees of freedom $C - |\theta|$. However, for the AH/F test, the null distribution is only approximately $\chi^2$. Aguirre-Hernández and Farewell showed through simulation that the $\chi^2$ approximation was adequate for models without covariates, but for models with fitted covariates, $T$ had a null distribution with a higher mean than the degrees of freedom. For a more accurate p-value they suggest it is necessary to bootstrap. The bootstrap algorithm involves simulating observed states from the fitted model at the existing observation times, refitting the model and calculating the statistic for each simulation. Since we are interested in the right hand tail of the distribution, at least 1000 bootstrap samples are recommended [44]. Thus, bootstrapping can be expensive computationally.

The lack of a known asymptotic null distribution for the Aguirre-Hernández/Farewell statistic causes problems for inference, particularly if the model being assessed has many

unknown parameters or if the dataset is large, meaning bootstrapping is prohibitively time consuming.

## 3.3 The null distribution of the AH/F statistic

In this section the null distribution of the AH/F statistic is investigated. A fast procedure for providing a good approximation of the asymptotic distribution of the statistic is outlined. This provides a practical alternative when the computation required to bootstrap is prohibitive.

### 3.3.1 Impact of non-identical multinomials for a fully specified test

The deviation of the null distribution of AH/F from a $\chi^2$ distribution is due to two main factors. Firstly, the counts are not identical multinomial, but rather the sum of non-identical multinomials. This is because individuals with distinct covariate values and time intervals between transitions are grouped together. Secondly, the maximum likelihood estimates for the data do not coincide with the minimum chi-squared estimate for the constructed contingency table.

The following result, which is arrived at by adaption of the standard derivation of the null distribution of Pearson chi-squared tests, assesses the impact that the non-identical counts have on the statistic.

**Lemma.** Let $\mathbf{X}_1, ..., \mathbf{X}_N$ be random variables with $\mathbf{X}_j \sim$ Multinomial$(1, \mathbf{p}_j)$ where $\mathbf{p}_1, ..., \mathbf{p}_N$ are known vectors of length $R$, s.t. $\sum_{r=1}^{R} p_{rj} = 1$ for all $j$ and let

$$T = \sum_{r}^{R} \frac{(\sum_j^N X_{rj} - \sum_j^N p_{rj})^2}{\sum_j^N p_{rj}}.$$

Then the limiting distribution of $T$ is not in general $\chi^2_{R-1}$.

The proof of this lemma is given in Appendix A. The proof shows that when observations comprising a group are not identically distributed, the limiting distribution of the resultant Pearson-type statistic, $T$, is given by a scalar product of a zero mean multivariate normal distribution with a non-diagonal covariance matrix. The disparity between the limiting distribution of $T$ and $\chi^2_{R-1}$ is dependent on the variability of the underlying probabilities within each group. In general however, the impact of non-identical multinomial counts is relatively small, with the statistic based on non-identical observations having only a

slightly lower mean and variance than $\chi^2_{R-1}$. When testing a fully-specified Markov model, the actual statistic would be of the form $\sum T_i$ where each $T_i$ is of the form given above. If each individual $T_i$ is not $\chi^2$ then neither is their sum.

### 3.3.2 Impact of unknown parameters

The AH/F statistic has a null distribution with an inflated mean compared to the naive degrees of freedom. This is due to the constructed contingency table not containing all the information in the data. As a result, the maximum likelihood estimate does not coincide with the minimum chi-squared estimate. This problem arises more generally in Pearson chi-squared tests, particularly for continuous data where the parameters of the model are fitted using the full data, but goodness-of-fit is assessed by considering the number of observations lying within discrete intervals.

The need for the estimator to be the minimum chi-squared estimator was first identified by Fisher [47]. Subsequently it has been established that a chi-squared test based on grouped data will result in a distribution which lies between $\chi^2_{d-p}$ and $\chi^2_d$ where $d$ is the number of independent cells in the contingency table and $p$ the number of unknown parameters fitted from the data (Kendall and Stuart) [76].

The asymptotic null distribution of the AH/F statistic can be derived, primarily by adaption of the methods used by Kendall and Stuart. This leads to the following theorem

**Theorem**: The asymptotic null distribution of AH/F, conditional on observation times, total group counts and the true parameter values, can be expressed as a scalar product of a multivariate normal distribution with zero mean vector and some known covariance matrix.

The proof, which is quite long, is given in appendix B. Although the result does not give an analytic distribution, the asymptotic distribution, conditional on the observation times, total group counts and true parameter values, can be obtained quickly, either by numerical integration or simplified simulation. The theorem establishes that AH/F is a quadratic form in normal random variables. Hence it can be expressed as a linear combination of independent $\chi^2$ random variables. Numerical algorithms for calculating the cumulative distribution function of such combinations exist [39]. Hence, one approach of determining a p-value for a particular observed value of the statistic, would be to evaluate the cdf. at that point. More generally however, the entire distribution can be simulated by simulating the

relevant multivariate normal distribution and taking its scalar product. 10000 or 100000 such samples can be generated for distributions of moderate dimension (e.g. 20-50) quite quickly.

The resulting distribution will still only be asymptotically accurate. The proof involves firstly finding the asymptotic joint distribution $p(\hat{\theta} - \theta, \mathbf{O})$ of $\hat{\theta} - \theta$ and $\mathbf{O}$, the observed counts in the AH/F contingency table. It then involves linearising each term of the AH/F statistic. The linearisation of the expression in equation B.1 to give B.6, in particular, may be a significant source of error for small or moderate sample sizes.

### 3.3.3 Example: CAV data without misclassification or deaths

To illustrate the method for calculating the null distribution we can apply it to the CAV data. However, AH/F can only be applied to data which are free from misclassification and do not contain exact death times. Hence we use the data with no misclassification assumed, where the observed state corresponds to the highest observed state up until that time. In addition, all transitions to death are excluded from the data. This involves censoring individuals who died at the last time they were observed to be alive. This method of removing deaths is not a legitimate way of testing goodness-of-fit in the presence of exact death times because such a censoring scheme will be informative. This makes the model a three state unidirectional process in which state 1 corresponds to 'CAV free', state 2 corresponds to 'mild CAV' and state 3 to 'severe CAV'. The resulting data consist of 1832 transitions. Note that this is a reduction from the 1972 angiograms in the full data because we assume that observation stops once a patient reaches 'severe CAV', which is now an absorbing state. We fit a time homogeneous Markov model to the data. The model includes two covariates affecting onset rates, donor age - which takes 51 unique values in the sample - and IHD which is binary. We wish to apply AH/F to this model. To avoid low counts we construct contingency tables that do not categorise by observation number. Instead we consider four different possible categorisations based upon time interval length and covariate value. These groupings are as follows:

1. Four groups corresponding to the quartiles of the time intervals. No grouping based on covariate values.

2. Categorised into groups depending on whether time interval is greater or less than the median, and whether donor age is greater or less than the median. No grouping

by IHD.

3. Categorised into groups depending on whether time interval is greater or less than the median, and whether IHD is absent or present. No grouping by donor age.

4. Categorised depending on whether IHD is absent or present and whether donor age is greater or less than the median. No time interval groupings.

Any observation may have occurred starting from either state 1 (with three possible outcomes) or state 2 (with two possible outcomes). This means each contingency table has 12 independent cells. However there are 4 unknown parameters estimated from the data, two transition intensities and two covariates effects on the $1 \to 2$ intensity. Hence a naive approximation to the null distribution of the statistic is $\chi^2_8$. This applies in all four cases. We would however expect the distributions to depend on the grouping.

The model fits badly. The primary reason for this is the informative censoring resulting from the exclusion of transitions to death. Subjects currently in state 2 are at greater risk of dying than those in state 1. As a result there are fewer than expected $2 \to 3$ transitions in long intervals and greater than expected $1 \to 2$ and $1 \to 3$ transitions in short intervals. The calculated goodness-of-fit statistics for the data under grouping schemes 1, 2, 3, 4 are 53.6, 51.4, 57.9 and 5.9 respectively. In the first three cases, the time interval groupings allow the poor fit to be recognised. For instance table 3.1 gives the case of four time quantile groups. In the final case where there are no separate time interval groups, the value of the statistic suggests a good fit (table 3.2). This is because the systematic biases for long and short intervals are cancelled out when the counts are joined. This result emphasises the need to choose appropriate groups for the statistic.

Using the methods of section 3.3.2, the approximate asymptotic null distributions have means of 9.88, 9.21, 8.88 and 8.16 for grouping methods 1, 2, 3 and 4 respectively. As we might expect, where there is no grouping by covariate value, the mean is higher because there is little correlation between the observed cell counts and the maximum likelihood estimates for the covariate effects. Grouping by IHD gives a lower mean than grouping by donor age. Grouping both by donor age and IHD gives a distribution with a mean not far from the expected 8. However, the variance is 15.5 which is significantly less than the 16 expected.

To assess the performance of the enhanced asymptotic null distributions, we can compare them to the approximate null distributions arrived at through bootstrapping. Table 3.3

Table 3.1: Contingency table for application of Aguirre-Hernández/Farewell statistic to CAV data with deaths removed, grouping by time interval length. TQ = Time quantile group

| TQ | | 1→1 | 1→2 | 1→3 | 2→2 | 2→3 |
|---|---|---|---|---|---|---|
| 1 | Obs | 288 | 37 | 7 | 100 | 27 |
| | Exp | 307.3 | 22.7 | 2.0 | 107.5 | 19.5 |
| | Dev | 1.3 | 5.6 | 3.6 | 0.6 | 2.1 |
| 2 | Obs | 319 | 42 | 7 | 72 | 18 |
| | Exp | 332.7 | 31.5 | 3.7 | 73.6 | 16.4 |
| | Dev | 0.6 | 2.6 | 1.5 | 0.0 | 0.1 |
| 3 | Obs | 375 | 48 | 10 | 27 | 4 |
| | Exp | 358.9 | 61.7 | 12.4 | 21.6 | 9.4 |
| | Dev | 0.7 | 3.9 | 0.6 | 1.1 | 7.2 |
| 4 | Obs | 353 | 58 | 16 | 21 | 3 |
| | Exp | 338.4 | 69.5 | 19.1 | 13.9 | 10.1 |
| | Dev | 0.6 | 2.3 | 0.6 | 2.4 | 16.9 |

Table 3.2: Contingency table for application of Aguirre-Hernández/Farewell statistic to CAV data with deaths removed, grouping by donor age (DA) and IHD.

| DA $\leq 25$ | IHD | | 1→1 | 1→2 | 1→3 | 2→2 | 2→3 |
|---|---|---|---|---|---|---|---|
| 0 | 0 | Obs | 328 | 45 | 7 | 50 | 10 |
| | | Exp | 324.1 | 46.4 | 9.5 | 47.2 | 12.8 |
| | | Dev | 0.0 | 0.0 | 0.7 | 0.2 | 0.6 |
| 0 | 1 | Obs | 265 | 65 | 16 | 66 | 21 |
| | | Exp | 270.6 | 62.0 | 13.4 | 68.8 | 18.2 |
| | | Dev | 0.1 | 0.1 | 0.5 | 0.1 | 0.4 |
| 1 | 0 | Obs | 414 | 35 | 10 | 43 | 10 |
| | | Exp | 418.5 | 34.1 | 6.4 | 42.1 | 10.9 |
| | | Dev | 0.0 | 0.0 | 2.0 | 0.0 | 0.1 |
| 1 | 1 | Obs | 328 | 40 | 7 | 61 | 11 |
| | | Exp | 324.1 | 42.9 | 7.9 | 58.5 | 13.5 |
| | | Dev | 0.0 | 0.2 | 0.1 | 0.1 | 0.5 |

Table 3.3: Comparison of summary statistics for the null distribution of AH/F under the different groupings using the four methods.

| Grouping | Method | Mean | Var | 25% | 50% | 75% | 95% | 99% |
|---|---|---|---|---|---|---|---|---|
| 1 | Bootstrap | 9.89 | 19.07 | 6.72 | 9.23 | 12.40 | 17.74 | 23.35 |
| | Linear | 9.88 | 19.34 | 6.67 | 9.23 | 12.38 | 18.06 | 22.85 |
| | Naive | 8.00 | 16.00 | 5.07 | 7.34 | 10.22 | 15.51 | 20.09 |
| 2 | Bootstrap | 9.24 | 17.41 | 6.20 | 8.61 | 11.56 | 16.98 | 21.65 |
| | Linear | 9.21 | 17.51 | 6.15 | 8.58 | 11.58 | 17.01 | 21.63 |
| | Naive | 8.00 | 16.00 | 5.07 | 7.34 | 10.22 | 15.51 | 20.09 |
| 3 | Bootstrap | 8.93 | 17.30 | 5.95 | 8.24 | 11.31 | 16.65 | 21.36 |
| | Linear | 8.88 | 17.43 | 5.83 | 8.22 | 11.22 | 16.65 | 21.39 |
| | Naive | 8.00 | 16.00 | 5.07 | 7.34 | 10.22 | 15.51 | 20.09 |
| 4 | Bootstrap | 8.21 | 15.46 | 5.38 | 7.59 | 10.35 | 15.61 | 20.20 |
| | Linear | 8.16 | 15.54 | 5.28 | 7.52 | 10.34 | 15.53 | 20.00 |
| | Naive | 8.00 | 16.00 | 5.07 | 7.34 | 10.22 | 15.51 | 20.09 |

gives summary statistics for the approximations to the null distribution arrived at via the four possible methods. The bootstrap is based on 5000 replications and the linear asymptotic approach on 100000 samples. 100000 samples using the linearised method takes around 30 seconds, in contrast 5000 replications via full bootstrapping takes around 15 hours.

The linear method does very well at replicating the means seen in the bootstrap distribution. Moreover, there is good overall agreement. In particular, the 95% points predicted by the asymptotic approximations are within 95% bounds from the Monte Carlo error of the bootstrap. Performing a two-sample Kolmogorov-Smirnov test between the bootstrap (5000 samples) and the linear distributions (100000 samples), is consistent with the null hypothesis, that they are from the same distribution, for all four cases. The naive method of using $\chi^2_8$ performs poorly, in terms of an overall approximation of the distribution, in all cases except when there is grouping by both covariates.

### 3.3.4 Conclusion

The above example demonstrates that bootstrapping is not always necessary to get a good approximation of the null distribution of the AH/F statistic. For the current example the disparity between the naive and the correct critical points for a 5% size test is fairly small

and would not have affected the conclusions drawn about the model. Even for the first grouping the model would only be incorrectly rejected if the value of the statistic lay between 15.5 and 18.0. The use of the improved asymptotic approximation is most useful for models with a larger number of unknown parameters, where the disparity between the true and the naive critical point point will be larger. It would also be useful if a series of covariate models are to be compared, because obtaining several null distributions through bootstrapping is likely to be time consuming and using the naive critical point will be unreliable. Using the linearised approximation may be preferable in a wide range of cases because the Monte Carlo error on the 95% point can be eliminated. The linearised approximation is however an asymptotic approximation, it will still therefore be necessary to bootstrap in cases where either the sample size or the cell counts are too small for the asymptotic results to be valid.

## 3.4 Modification for misclassification hidden Markov models

AH/F is only applicable to Markov models. It is however straightforward to extend the AH/F goodness-of-fit test to accommodate a misclassification hidden Markov model. Firstly note that the likelihood for an individual in such a model can be written as

$$\mathbb{P}(O_1, ..., O_n) = \mathbb{P}(O_1)\mathbb{P}(O_2|O_1)\mathbb{P}(O_3|O_1, O_2)\dots\mathbb{P}(O_n|O_1, \dots, O_{n-1})$$

where each of the $O_k|O_1, \dots, O_{k-1}$ are conditionally independent multinomial random variables. The contribution of the data from individual $i$ to the contingency table corresponds to the transition probabilities between observed states. Expected probabilities of obtaining the observed data are determined by first calculating the vector of true state occupancy probabilities at the start of the interval of interest, conditional on all observations up to that time and conditional on the current observed state, say $r$,

$$\hat{\xi}_{\mathbf{r}} = \frac{\pi_{\mathbf{0}} M_1 M_2 \dots M_{k-1}}{\pi_{\mathbf{0}} M_1 M_2 \dots M_{k-1} \mathbf{1}} \tag{3.7}$$

where $\mathbf{1}$ is a vector of length $R$ in which all the entries are 1, $\pi_{\mathbf{0}}$ is the vector of initial occupation probabilities and $M_i$ is an $R \times R$ matrix with (r,s) entry

$$e_{s,O_i} p_{rs}(t_i - t_{i-1}).$$

The probability that the next observed state is $s$ is then:

$$\tilde{p}_{rs}(t_k - t_{k-1}) = \sum_{j}^{R} \sum_{l}^{R} \hat{\xi}_{rj} p_{jl}(t_k - t_{k-1}) e_{ks}$$

where $\hat{\xi}_{rj}$ is the $j$th entry in the vector $\hat{\xi}_r$.

These filtered transition probabilities can then be substituted into the equations for expected transitions, and the statistic can be computed as in section 3.2.

As before, the statistic involves approximating a sum of $n$ non-identical multinomials of size 1 with a multinomial of size $n$. However, if we maintain the same type of grouping of observations as before we would group two observations with the same previous observed state together, regardless of their overall history. Clearly a subject with a pattern of observations like $1, 2, 2, 2, 2, 2, 1$ will have a considerably different distribution of their next observed state than one with $1, 1, 1, 1, 1, 1, 1$. The former having a much higher probability of actually being in state 2. In addition, because the grouping causes particular information loss with respect to the misclassification probabilities, the maximum likelihood estimate does not coincide with the value of $\theta$ which minimises the statistic. A good approximation of the null distribution can still be found by using the methods of section 3.3.2. The method remains applicable in this case, although calculation of the derivatives of the expected transition probabilities becomes more intricate.

Where sufficient data are available, these problems can be lessened by allowing grouping by the last two observed states rather than just the last state. Bureau *et al* [14] applied such a scheme to construct contingency tables. In their approach the usual comparison $\frac{(O-E)^2}{E}$ was used, but not compared to a known distribution. Grouping by more than the previous state should also allow greater power to test the assumptions in the model specific to misclassification, but requires a lot of data as there is the potential for many groups.

## 3.5 Exact death times

Often a Markov model for panel data has an absorbing state, such as death, for which the time of entry is known precisely. As noted in section 1.3.2, in this situation it is necessary to alter the likelihood calculation to accommodate this. The contribution of an observed

death is

$$\sum_{s}^{R-1} p_{rs}(t) q_{sR}$$

where $R$ is the absorbing state and $r$ was the last observed state. In chapter 2, it was noted that neither the summary residuals approach of section 2.4.2 nor the prediction of future observations table of section 2.6.1 could be applied in the case of exact absorption times.

For similar reasons, AH/F does not perform well in this situation. The test is valid when all transitions are interval censored since the likelihood contribution from each observation is proportional to that of a multinomial with cell probabilities determined by the time $t$, which is independent of the process, at least when there are no covariates. Hence there is a fully defined notion of the expected number of transitions for any particular interval. If a death occurs at a time $t$ since the previous observation, an observation only takes place at that time because of the death. Had a transition to the absorbing state not occurred the observation would have taken place later. In this way the sampling scheme is not independent of the process that has been modelled and a goodness-of-fit test performed on data that include exact death times using this incorrect way of calculating expected transitions will result in extreme values of the statistic even when the model is valid. This is due to large deviances in the cells relating to the deaths: patients are typically scheduled to have periodic observations, the actual between-visit intervals will vary, but there will be very few very short intervals. The quantile relating to the shortest time intervals will have a high proportion of observed deaths, which is not reflected in the expected number of deaths in such intervals.

### 3.5.1 A simulated illustrative example

In order to illustrate the problems caused by having exact times of entry into the absorbing state we simulate some data with an irregular sampling scheme. Each dataset contains 500 patients who are observed a series of times. Times between observations are independent and identically distributed. To imitate what might occur in an observational study with planned observation times at one year intervals, the distribution is taken to be a mixture of translated gammas, such that 50% of intervals are around 1 year, 25% are around 2 years, 20% are around 3 years and 5% have no further observation. Patients are each censored (meaning mortality follow-up ends) at a uniformly distributed time between 5

Table 3.4: Contingency table for naive application of Aguirre-Hernández/Farewell statistic to a simulated dataset including exact death times. TQ = Time quantile group

| TQ | | 1→1 | 1→2 | 1→3 | 1→4 | 2→2 | 2→3 | 2→4 | 3→ 3 | 3→4 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Obs | 454 | 25 | 4 | **47** | 55 | 9 | **18** | 56 | **30** |
| | Exp | 486.0 | 30.0 | 2.5 | **11.4** | 68.2 | 10.3 | **3.5** | 74.7 | **11.3** |
| | Dev | 2.11 | 0.83 | 0.90 | **111.2** | 2.55 | 0.16 | **60.1** | 4.68 | **30.9** |
| 2 | Obs | 440 | 44 | 6 | 20 | 53 | 18 | 6 | 41 | 6 |
| | Exp | 440.5 | 44.2 | 6.1 | 19.1 | 55.2 | 15.1 | 6.6 | 34.5 | 12.5 |
| | Dev | 0.00 | 0.00 | 0.00 | 0.04 | 0.09 | 0.56 | 0.05 | 1.22 | 3.38 |
| 3 | Obs | 437 | 71 | 25 | **13** | 43 | 10 | **3** | 17 | **2** |
| | Exp | 417.4 | 71.1 | 17.5 | **40.0** | 30.6 | 15.3 | **10.2** | 10.8 | **8.2** |
| | Dev | 0.92 | 0.00 | 3.21 | **18.2** | 5.02 | 1.84 | **5.08** | 3.56 | **4.69** |

and 10 years. The patients follow a 4-state disease Markov process in which all patients are in state 1 at time 0 and can either progress through states 2 and 3 or enter state 4 from any other state, in the same way as the CAV dataset introduced in chapter 2. The precise time of entry into state 4 is known, if a patient is censored this censoring time is known. Moreover, the time to which follow-up would have continued had the patient survived is also known. Such a scenario would be realistic if patients have distinct calendar time start points, and the final censoring time relates to a common calendar time.

The correct 4 state Markov model is fitted to the data, with the parameters estimated by maximum likelihood. A naive application of AH/F to this model, taking the death time as interval censored and ignoring censored observations, is applied. Three time quantile groupings are chosen. To limit the size of the resulting contingency table we do not group by observation number and there are no covariates in this simple illustration. We get a statistic of 309.5 which can be naively compared to $\chi^2_{13}$, thus the test suggests the simulated data are very unlikely to come from a Markov model. A similar result occurs if we also group by observation number.

Table 3.4 gives the resulting contingency table with cells contributing the most to the statistic given a bold font. The vast majority of the deviance is from transitions to death. In particular, there are greater than expected deaths in short intervals (time quantile 1) and fewer than expected in longer intervals (time quantile 3).

## 3.6 The modified goodness-of-fit test

A simple approach to the problem of exact death times is to remove transitions to death from the analysis and compute a reduced statistic. However, this does not assess the fit of the complete model structure. Instead in this section a modified goodness-of-fit statistic, incorporating exact death times, is sought.

### 3.6.1 Incorporating exact death times

For the $N$ individuals in the study, let $\mathcal{D}$ denote the set of times to death and $\mathcal{C}$ denote the set of maximum follow up times for mortality for all patients (including those who have died). We denote the set of sampling times by $\mathcal{S}$ and partition it as $\mathcal{S} = (\mathcal{Y}, \mathcal{Z})$ where $\mathcal{Y}$ is the set of observed sampling times for interval censored observations and $\mathcal{Z}$ the set of unobserved times of the next scheduled realisation of the process, that would have taken place had the individual survived. These may or may not be censoring times. In addition we denote $\mathcal{X}$ the set of observations of the process $X$ at the times in $\mathcal{S}$.

Initially, consider the case in which both the time of death $d_i$ and the time of the next scheduled realisation of the process $t^*$ for individual $i$, with $n$ observations, are known. Then if the penultimate observed state is $r$ at time $t_{n-1}$, then the expected transitions for the final interval would be given by $p_{rR}(t^* - t_{n-1})$ for transitions to death and $p_{rs}(t_n - t_{n-1})$ for non-death transitions and the AH/F statistic could be applied as described in section 3.2. The observed contribution to the contingency table for transitions to death is

$$\mathbb{1}\left(X(t^*) = s = R | X(t_{n-1}) = r\right)$$

and for non-death transitions is

$$\mathbb{1}\left(X(t_n) = s \neq R | X(t_{n-1}) = r\right)$$

where $X(t_{n-1})$ is the state at the previous observation. Transitions to death contribute to the cell for the time interval group that contains $t^* - t_{n-1}$. The statistic $T$ has a null distribution very similar to that of an analogous AH/F statistic for a model without exact deaths.

Hence if the set of next scheduled observation times $\mathcal{Z}$ were known precisely calculation of a goodness-of-fit statistic in the presence of exact deaths would be straightforward. Thus conditional on $\mathcal{S}$, $\mathcal{X}$, the value of $T$ is the AH/F statistic described in section

3.2, which is straightforward to calculate. However, in many observational studies of interest, the potential sampling times are not known or not recorded. Since $\mathcal{S}$ is only partially observed it is convenient to think of this as a missing data problem. Due to computational limitations we take a two stage approach to the estimation of $T$: first we estimate the distribution of sampling intervals from the empirical distribution and then we use Monte-Carlo simulation from this empirical distribution to calculate a sample from the distribution of $T$.

More formally, we want to obtain the distribution of

$$T|\mathcal{Y},\mathcal{X},\mathcal{D},\mathcal{C}.$$

To proceed we make the assumption that the sampling times $t \in \mathcal{S}$ are independent and identically distributed. Initially we can consider the case where the time intervals are an i.i.d. sample from an entirely specified sampling distribution $f(t)$ with associated survivor function $F(t)$. Then, to obtain the distribution of $T$ it is necessary to integrate over the missing observation times $\mathcal{Z}$,

$$\begin{aligned}\mathbb{P}(T|\mathcal{Y},\mathcal{X},\mathcal{D},\mathcal{C}) &= \int_{\mathcal{Z}} \mathbb{P}(T|\mathcal{Y},\mathcal{Z}=\mathbf{z},\mathcal{X})\mathbb{P}(\mathcal{Z}|\mathcal{D},\mathcal{C})d\mathbf{z} \\ &= \int_{\mathcal{Z}} \mathbb{P}(T|\mathcal{Y},\mathcal{Z}=\mathbf{z},\mathcal{X}) \prod_{i=1}^{N} \left( \left(\frac{f(z_i)}{F(d_i)}\right)^{\mathbb{1}(z_i<c_i)} \left(\frac{F(c_i)}{F(d_i)}\right)^{\mathbb{1}(z_i=c_i)} \right) d\mathbf{z}\end{aligned}$$

where $N$ is the total number of patients, $z_i$ represents the unobserved length of the interval in which patient $i$ died, $d_i$ the death time for patient $i$, and $c_i$ is the potential censoring time for patient $i$.

A more realistic situation is one in which $f(t)$ is not known. In this situation, the empirical estimate $\hat{f}(t)$ of $f(t)$, calculated using $\mathcal{Y}$, $\mathcal{D}$ and $\mathcal{C}$, provides a convenient approximation. Specifically we use a product-limit estimate of the time to next scheduled observation. The intervals between non-fatal observations are taken as the times to events. The time between the last non-fatal observation and death or (survival) censoring are taken as the times of censoring with respect to time to next scheduled event. This leads to a step-function estimate $\hat{F}(t)$, and a discrete distribution for $\hat{f}(t)$. Substituting $\hat{f}(t)$ for $f(t)$ gives

$$\begin{aligned}\hat{\mathbb{P}}(T|\mathcal{Y},\mathcal{X},\mathcal{D},\mathcal{C}) &= \int_{\mathcal{Z}} \mathbb{P}(T|\mathcal{Y},\mathcal{Z}=\mathbf{z},\mathcal{X})\hat{\mathbb{P}}(\mathcal{Z}|\mathcal{D},\mathcal{C},\mathcal{Y})d\mathbf{z} \\ &= \sum_{\mathcal{Z}} \mathbb{P}(T|\mathcal{Y},\mathcal{Z}=\mathbf{z},\mathcal{X}) \prod_{i=1}^{N} \left( \left( \frac{\hat{f}(z_i)}{\hat{F}(d_i)} \right)^{\mathbb{1}(z_i<c_i)} \left( \frac{\hat{F}(c_i)}{\hat{F}(d_i)} \right)^{\mathbb{1}(z_i=c_i)} \right)\end{aligned}$$

where the summation is over the set of unique values of $\mathbf{z} \in \mathcal{Z}$.

Given $\mathcal{Y}$, $\mathcal{Z}$ and $\mathcal{X}$, $T$ is deterministic, let $(T|\mathcal{Y},\mathcal{Z}=\mathbf{z},\mathcal{X}) = T_{\mathcal{Y},\mathcal{Z}=\mathbf{z},\mathcal{X}}$, and we can write

$$\hat{\mathbb{E}}(T|\mathcal{Y},\mathcal{X},\mathcal{D},\mathcal{C}) = \sum_{\mathcal{Z}} T_{\mathcal{Y},\mathcal{Z}=\mathbf{z},\mathcal{X}} \prod_{i=1}^{N} \left( \left( \frac{\hat{f}(z_i)}{\hat{F}(d_i)} \right)^{\mathbb{1}(z_i<c_i)} \left( \frac{\hat{F}(c_i)}{\hat{F}(d_i)} \right)^{\mathbb{1}(z_i=c_i)} \right),$$

where the summation is over the set of unique values of $\mathbf{z} \in \mathcal{Z}$.

Estimation of this expectation is problematic since the summation is over a large number of possible values for $\mathcal{Z}$. However, we can use Monte Carlo sampling to generate random vectors $\mathbf{z}^*$ from the estimated distribution of $\mathcal{Z}|\mathcal{D},\mathcal{C}$ and these can be used to calculate a sample $\mathbf{T}^*$ from an approximate distribution for $T|(\mathcal{Y},\mathcal{X},\mathcal{D},\mathcal{C})$. The mean of the sample provides a point estimate for $T$ and, in our examples, 100 realisations of $\mathbf{z}^*$ was sufficient to provide a robust estimate of the mean. Asymptotically, provided the sampling intervals are i.i.d., the mean of this random sample will have a null distribution, with a mean close to that of the equivalent AH/F statistic, but with a reduced variance. The variance is reduced because by taking the mean of $\mathbf{T}^*$ we remove the variability from the unknown $\mathcal{Z}$.

In common with AH/F the distribution of the statistic is complicated. Unfortunately, it does not seem possible to extend the methods of section 3.3.2 to the case of exact death times. Therefore to get an accurate p-value bootstrapping is required. As with the AH/F test, bootstrap samples of the null distribution are calculated by simulating data from the Markov model with true parameters taken to be the maximum likelihood estimates from the observed data, at the same observation times as in the original data. Model parameters are then fitted by maximum likelihood estimation and the statistic for the data is computed in the same way as the observed data. Hence the computation of 1000 bootstrap samples involves calculating 100 values of the statistic for each of the 1000 fitted models.

### 3.6.2 Similarity with Multiple Imputation

The proposed method of calculating $T$ is similar in many respects to multiple imputation (MI) [89, 112] which is the standard approach to missing data problems. However, for practical purposes we do not follow a principled MI method.

To perform MI one should, in addition to the above steps, also calculate the mle based on the imputed data, for each completed dataset. This involves assuming that rather than knowing that a death occurred at $d_i$, we only know the death occurred before $t_n^*$. For the $k$th completed dataset, we denote the mle based on this dataset as $\hat{\theta}_k^*$. The transition probabilities to be used to calculate the $k$th sample of the statistic would then need to be calculated on the basis of $\hat{\theta}_k^*$. This would ensure that the distribution of each MI sample would be the same as if the next observation times had been known (but exact times of death were not known).

However, recalculating both $\hat{\theta}_k^*$ and the expected transitions for each MI sample would greatly increase the computation necessary to calculate the statistic. With a full MI approach, it would potentially be possible to get a known distribution for $T_k^*$, but the distribution of the overall statistic $T = \frac{1}{N}\sum_{k=1}^{N} T_k^*$ would still be unknown. Therefore, bootstrapping would still be necessary. However, bootstrapping would be much more time consuming under a full MI approach.

Instead of using $\hat{\theta}_k^*$, we instead use the maximum likelihood estimate based on the exact death times $\hat{\theta}$ for all samples. The effect of this will be that the mean of $T$ under null conditions will tend to be slightly higher using $\hat{\theta}$ compared to $\hat{\theta}_k^*$. However, we would expect the difference to be quite small. We might also expect that $Var(T_k^*)$ may also be lower when using $\hat{\theta}$. But again, the effect is likely to be small.

Finally, using $\hat{\theta}_k^*$ means that effectively the additional information gained from knowing the exact death times is not used to assess the model fit. Disagreement between $\hat{\theta}_k^*$ and $\hat{\theta}$ is itself a sign of poor fit (provided the imputation model is correct). However, if any gain in power exists from using $\hat{\theta}$ it is likely to be small.

### 3.6.3 Incorporation of censoring

Often data with exact death times will also feature censored observations, at which time it is only known the patient has not entered the absorbing state. Potentially such a situation could also arise without exact death times. However, neither the likelihood ratio tests of

Kalbfleisch and Lawless [70] nor the Aguirre-Hernández/Farewell test, accommodate this type of observation. There are two *ad hoc* ways of dealing with censored observations.

Firstly one could remove the censored observations from consideration. However, to maintain consistency one would also be required to remove any observed death that came from intervals which would otherwise be censored and this would reduce the power of the analysis.

Alternatively one could treat intervals that may either end in censoring or death as a separate category in the contingency table. This would allow the whole data to be considered. The creation of a separate category implies that the alternative hypothesis includes the possibility of the transition probabilities being affected by whether or not the interval is subject to censoring. While there should be no harm in including this possibility, it does not seem particularly necessary. Also if the number of censored intervals is small, separate categories by censoring may not be viable. More generally, the number of overall categories is always limited by the total sample size. The inclusion of a censored category will impede on the number of other, potentially more relevant, categories that can be used.

It would be desirable to have a way of considering censored and non-censored observations together within the same categories. This can be achieved by using a likelihood ratio test approach.

### 3.6.4 Aguirre-Hernández/Farewell test as a likelihood ratio test

As explained in section 3.1 when observation times are the same for all patients and any covariates present are categorical, the Pearson chi-squared test is asymptotically equivalent to a likelihood ratio test of the Markov model (in which the $p_{rs}$ are specified as the relevant transition probabilities), against the alternative where $p_{rs}$ is unrestricted (except $\sum_s p_{rs} = 1$) within each of the multinomials. Wilks' theorem [135] ensures that asymptotically

$$2l(\hat{p}) - 2l(\hat{\theta}) \sim \chi^2_{|p|-|\theta|} \tag{3.8}$$

where $|x|$ denotes the dimension of the space $x$,

$$l(\theta) = \sum_c \sum_l \sum_r^R \sum_s^R n_{rsl}^{(c)} \log\left(p_{rs}(t_l; v_c; \theta)\right)$$

and $c$ represents observations categorised by covariate value, $l$ represents observations categorised by time interval between observations, $r$ and $s$ are the states at the start and

end of the interval, and $n_{rsl}^{(c)}$ represents the number of such transitions observed in the dataset.

AH/F allows for a wider range of sampling schemes and data. It is related to the likelihood ratio test via two approximations. A grouping technique analogous to the natural grouping in the likelihood ratio test is performed. Whereas previously the observation counts in each group were from a single multinomial of size $n_{rl}^{(c)}$, now they are formed from the sum of $n_{rl}^{(c)}$, independent but not identical multinomials of size 1. The likelihood function $l(\theta)$ can be thought of as being replaced by an approximation $\tilde{l}(\theta)$. Unless the grouping involved is particularly severe, properties of maximum likelihood estimates will be approximately maintained for the approximate likelihood. In particular

$$2(\tilde{l}(\hat{p}) - \tilde{l}(\theta^*)) \approx \chi^2_{|p|-|\theta|}$$

where $\hat{p}$ maximises $\tilde{l}$ under the alternative and $\theta^*$ maximises $\tilde{l}$ under the null. Note however, that $\theta^* \neq \hat{\theta}$, the latter being the maximum under the full likelihood. Provided the two estimates are close, the approximation will still be appropriate.

### 3.6.5 Efficient incorporation of censoring

Consider again the case where all patients are observed at the same times and any covariates are categorical, but in addition allow some observations to be censored. For instance it may only be known whether a subject is in state $R$ or not. Then the log-likelihood is:

$$l(\theta) = \sum_c \sum_l^L \sum_r^R \sum_s^R n_{rsl}^{(c)} \log\left(p_{rs}(t_l; v_c; \theta)\right) + \sum_c \sum_l \sum_r n_{rCl}^{(c)} \log\left(1 - p_{rR}(t_l; v_c; \theta)\right)$$

where $n_{rCl}^{(c)}$ denotes the number censored from state $r$ among intervals of length $t_l$ and covariate value $v_c$. The likelihood here can still be thought of in terms of a sample of multinomials, but a sample in which some of the observations are censored. We wish to perform a likelihood ratio test where the null is the fitted Markov model and where the transition probabilities are unrestricted (but unaffected by whether the interval is censored) in the alternative.

Maximising under the alternative model involves a straightforward application of Lagrangian multipliers. Each multinomial model, corresponding to a single covariate, time interval and initial state combination, can be maximised independently. Data for each multinomial is of the form of observed counts $(n_1, ..., n_R, n_C)$ and we seek to get estimates

for $(p_1, ..., p_R)$ from this. The estimate for the cell probability relating to the absorbing state is just

$$\hat{p}_R = \frac{n_R}{n} \tag{3.9}$$

while for the other probabilities, there is an additional reweighting term:

$$\hat{p}_r = \frac{n_r}{n}\left(\frac{\sum_{j\neq R} n_j}{\sum_{j\neq R,C} n_j}\right). \tag{3.10}$$

These maximum likelihood estimates, $\hat{p}$, can then be substituted into the usual likelihood ratio test formula (3.8), with the degrees of freedom staying the same.

A Pearson-type chi-square test can also be derived by an adaption of the standard derivation of it from the likelihood ratio test. We can write the statistic in the form:

$$2\sum_c \sum_{l=1}^{L}\sum_{r=1}^{R}\sum_{s\in V} n_{rsl}^{(c)} \log\frac{\hat{p}_{rslc}}{\tilde{p}_{rslc}} \tag{3.11}$$

where $V = \{1, ..., R\} \cup \{C\}$, $\hat{p}_{rslc}$ represents the fitted probabilities from the unrestricted model and $\tilde{p}_{rslc}$ represents the fitted probabilities from the Markov model, and $p_{rClc} = 1 - p_{rRlc}$. If we define $e_{rsl}^{(c)}$ such that it satisfies the equation

$$e_{rsl}^{(c)}\hat{p}_{rslc} = n_{rsl}^{(c)}\tilde{p}_{rslc} \tag{3.12}$$

then we can rewrite the expression in (3.11) as a sum of terms of the form

$$2n_{rsl}^{(c)} \log\left(1 + \frac{n_{rsl}^{(c)} - e_{rsl}^{(c)}}{e_{rsl}^{(c)}}\right).$$

Note that equation (3.12) only defines $e_{rsl}^{(c)}$ uniquely if $n_{rsl}^{(c)} > 0$. If $\tilde{p}_{rslc} = 0$ then $e_{rsl}^{(c)} = 0$. However, in the case in which $\tilde{p}_{rslc} > 0$ and $n_{rsl}^{(c)} = 0$ (something that would not occur in an asymptotic limit), to ensure $\sum_{r,s,l,c} n_{rsl}^{(c)} = \sum_{r,s,l,c} e_{rsl}^{(c)}$ we can choose $e_{rsl}^{(c)}$ so that $\sum_m n_{rml}^{(c)} = \sum_m e_{rml}^{(c)}$ for the particular values of $r$, $l$ and $c$.

We can follow the well known derivation [5], involving Taylor expanding the logarithm about $\frac{n_{rsl}^{(c)} - e_{rsl}^{(c)}}{e_{rsl}^{(c)}} = 0$ and discarding higher order terms to get an expression:

$$\frac{(n_{rsl}^{(c)} - e_{rsl}^{(c)})^2}{e_{rsl}^{(c)}} + (n_{rsl}^{(c)} - e_{rsl}^{(c)}).$$

At this point in the standard derivation, when we come to sum over $r$ we find that

$$\sum_s (n_{rsl}^{(c)} - e_{rsl}^{(c)}) = 0.$$

Table 3.5: Contingency table for the modified statistic applied to the simulated dataset including exact death times. TQ = Time quantile group. C = censored in either state 1,2 or 3

| TQ | | 1→1 | 1→2 | 1→3 | 1→4 | 1→C | 2→2 | 2→3 | 2→4 | 2→C | 3→ 3 | 3→4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Obs | 454 | 25 | 4 | 13.2 | 139 | 55 | 9 | 5.7 | 41 | 56 | 9.8 |
| | Exp | 451.7 | 29.0 | 2.5 | 13.0 | 139.0 | 55.6 | 9.3 | 4.3 | 41.5 | 55.8 | 10.0 |
| 2 | Obs | 440 | 44 | 6 | 23.2 | 58 | 53 | 18 | 8.2 | 24 | 41 | 15.2 |
| | Exp | 441.1 | 44.2 | 6.1 | 21.5 | 58.2 | 55.4 | 15.1 | 8.9 | 23.8 | 41.3 | 14.9 |
| 3 | Obs | 437 | 71 | 25 | 43.6 | 48 | 43 | 10 | 13.1 | 11 | 17 | 13.0 |
| | Exp | 437.6 | 74.9 | 18.5 | 45.9 | 47.8 | 34.8 | 17.4 | 14.1 | 10.8 | 16.9 | 13.0 |

This is also the case here, though less immediate. We can write

$$\sum_r (n_{rsl}^{(c)} - e_{rsl}^{(c)}) = \sum_r n_{rsl}^{(c)} (1 - \frac{\tilde{p}_{rslc}}{\hat{p}_{rslc}})$$

and then substituting the expressions for $\hat{p}_{rslc}$ given in equations (3.9) and (3.10) will give the required result. So

$$\sum_c \sum_l^L \sum_{r=1}^{R-1} \sum_{s \in V} \frac{(n_{rsl}^{(c)} - e_{rsl}^{(c)})^2}{e_{rsl}^{(c)}}$$

does have an asymptotic $\chi^2$ distribution.

The same arguments as section 3.6.4 regarding approximating the likelihood through grouping of non-identical intervals can be used to extend this statistic to the case of irregular time intervals and continuous covariates.

### 3.6.6 Results on the simulated dataset

Applying our modified test to the same simulated data from section 3.5.1, we get a value of 14.5 (standard deviation 3.1) for the Pearson-type statistic. We expect the mean of our null distribution to have a mean of around 13, so the test presents no evidence against a Markov model.

Table 3.5 gives the contingency table of observed and expected counts, averaged over the 100 imputed full datasets.

1000 bootstrap samples were taken. The null distribution was found to have a mean of 13.4 and variance of 16.0. Thus the null distribution does maintain the same mean as the

equivalent test without exact deaths or censoring, but with a reduced variance. The 95% point was at 20.3.

To assess the accuracy of the estimate of the distribution of sampling times, the distribution of individual realisations of $T^*$ can be examined. These had a variance of 25.1, which is reasonably close to 26, which is what would be expected from $\chi^2_{13}$. Therefore we would not expect the null distribution used in the test of fit to be affected by approximating the unknown sampling time distribution, $f$, by its estimate, $\hat{f}$.

### 3.6.7 Results for CAV example

No existing methodology is available to formally assess fit for these data due to the misclassification, censoring and exact death times. The modified goodness-of-fit test was applied to the fitted model for the CAV data, including misclassification and with donor age and IHD as covariates affecting onset. Three time quantile groups and two covariate groups were chosen. Ideally we would also wish to group by observation number and the pattern of previous states, but we are restricted by the moderate sample size.

Aguirre-Hernández and Farewell suggested that to avoid small cell counts, one could group rarer types of transitions together. Given the small numbers of $3 \rightarrow 1$ and $3 \rightarrow 2$ transitions in the CAV data, we choose to group these together with $3 \rightarrow 3$ transitions. This creates a column in the contingency table labelled $3 \rightarrow 3^*$. The expected counts in this column are found by adding the expected numbers of $3 \rightarrow 1$, $3 \rightarrow 2$ and $3 \rightarrow 3$ transitions together.

In this example, the covariate groups are determined by considering $q_{12}(\mathbf{z}, \hat{\theta})$ for each individual. If this onset intensity is above the median then the subject is categorised as a rapid progressor, otherwise they are a slow progressor. This allows two covariates, presence of IHD and donor age, to provide a single measure. The disadvantage of this is that the covariate groupings depend on the estimates of the unknown covariate effects. This could potentially affect the power of the statistic.

Overall there were 42 independent cells in the contingency table (see Table 3.6) and 11 parameters in the hidden Markov model, so the naive degrees of freedom were 31. The observed value of the statistic $T$ was 63.4. The most substantial deviances occur due to larger than expected counts of $1 \rightarrow 2$ and $1 \rightarrow 3$ transitions in short intervals. This indicates that transition rates may vary over time and a model with piecewise constant hazards in the underlying Markov model might provide a better fit. To a lesser extent,

$2 \to 1$ and $2 \to 3$ transitions in short intervals for patients in covariate group 2 (faster progressors) are under and over represented respectively.

1000 bootstrap samples gave a mean value of 36.5 to the statistic, with a variance of 47.8. The p-value was 0.001 as only one value in the bootstrap exceeded the observed value. Thus the model does not fit well.

The individual elements of the $T^*$ from the Monte Carlo simulations (1000 bootstrap $\times$ 100 Monte Carlo samples) had a variance of 70.4, which is close to being twice the mean, as we would expect from a $\chi^2$ distribution. As with the simulated example, estimation of the sampling distribution of observation times does not appear to have had much effect on the null distribution of the statistic.

## 3.7 Conclusion

Pearson-type goodness-of-fit tests can be used to assess the fit of both Markov and misclassification hidden Markov models. The methods of Aguirre-Hernández and Farewell, coupled with the new methods presented in this chapter, allow irregular observations, continuous covariates and exact death times.

An existing drawback of the AH/F statistic was the lack of a known null distribution, making bootstrapping necessary. Assumption of an asymptotic $\chi^2$ distribution is only appropriate in the case of a Markov model with regular observations and at most categorical covariates. However, at least in the absence of exact death times, the methods of this chapter allow a far better asymptotic approximation of the null distribution to be computed quickly.

Unfortunately, this method is not transferable to the case of exact death times. Here, bootstrapping is still necessary. Each bootstrap iteration involves simulating new data on the existing sampling frame using the fitted parameters from the original model, refitting the parameters and applying the statistic. The time it takes to refit the models is the most significant factor. If fitting the original model takes a non-trivial amount of time, then the bootstrapping will be time consuming. Misclassification models and those with continuous covariates or large datasets are most likely to be problematic.

For large models it may not be possible to get a precise p-value. However, we have established that the null distribution of the modified statistic has approximately the same mean as an analogous statistic performed on equivalent data without exact deaths. Moreover,

Table 3.6: Contingency table of observed and expected counts for the CAV dataset using the modified method for exact death times and censoring. TQ = Time quantile group. C= censored in state 1, 2 or 3.

| CG | TQ | | 1→1 | 1→2 | 1→3 | 1→4 | 1→ $C$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | Obs | 160 | 35 | 1 | 7.2 | 23 |
| | | Exp | 168.8 | 24.3 | 4.8 | 5.1 | 23.2 |
| | 2 | Obs | 217 | 41 | 6 | 16.1 | 18 |
| | | Exp | 216.6 | 39.8 | 10.8 | 12.7 | 18.2 |
| | 3 | Obs | 267 | 57 | 14 | 50.7 | 101 |
| | | Exp | 259.9 | 59.1 | 20.8 | 48.4 | 101.5 |
| 2 | 1 | Obs | 215 | 25 | 8 | 3.8 | 25 |
| | | Exp | 222.9 | 19.9 | 3.4 | 5.8 | 24.8 |
| | 2 | Obs | 251 | 20 | 6 | 7.0 | 15 |
| | | Exp | 241.3 | 25.9 | 6.2 | 10.8 | 14.8 |
| | 3 | Obs | 256 | 26 | 9 | 37.2 | 94 |
| | | Exp | 249.7 | 31.6 | 9.7 | 37.1 | 94.0 |
| CG | TQ | | 2→1 | 2→2 | 2→3 | 2→4 | 2→C |
| 1 | 1 | Obs | 19 | 54 | 20 | 2.8 | 7 |
| | | Exp | 19.5 | 50.8 | 19.9 | 5.9 | 6.8 |
| | 2 | Obs | 9 | 24 | 11 | 4.6 | 12 |
| | | Exp | 8.5 | 24.0 | 10.8 | 5.4 | 11.8 |
| | 3 | Obs | 1 | 10 | 3 | 22.6 | 31 |
| | | Exp | 2.3 | 6.8 | 5.2 | 21.3 | 31.6 |
| 2 | 1 | Obs | 7 | 28 | 16 | 3.2 | 3 |
| | | Exp | 14.4 | 27.1 | 9.7 | 3.0 | 3.0 |
| | 2 | Obs | 9 | 11 | 4 | 2.6 | 1 |
| | | Exp | 6.7 | 12.2 | 5.4 | 2.3 | 1.0 |
| | 3 | Obs | 1 | 7 | 0 | 12.1 | 15 |
| | | Exp | 2.3 | 3.4 | 2.7 | 11.3 | 15.5 |
| CG | TQ | | 3→ 3* | 3→4 | 3→ $C$ | | |
| 1 | 1 | Obs | 48 | 4.1 | 2 | | |
| | | Exp | 45.6 | 6.6 | 1.9 | | |
| | 2 | Obs | 24 | 3.3 | 0 | | |
| | | Exp | 23.1 | 4.2 | 0.0 | | |
| | 3 | Obs | 2 | 21.6 | 14 | | |
| | | Exp | 1.7 | 24.1 | 11.9 | | |
| 2 | 1 | Obs | 30 | 6.1 | 1 | | |
| | | Exp | 31.4 | 4.6 | 1.0 | | |
| | 2 | Obs | 18 | 4.8 | 0 | | |
| | | Exp | 19.2 | 3.6 | 0.0 | | |
| | 3 | Obs | 2 | 15.1 | 9 | | |
| | | Exp | 1.9 | 15.7 | 8.5 | | |

the variance will be reduced. Hence an upper bound on the 95% point can be obtained by taking the predicted 95% from the equivalent statistic without exact deaths using the approximation method of section 3.3.2.

The choice of groupings in Pearson-type goodness-of-fit tests can have a marked effect on both the power of the test and the test result in specific cases. This was demonstrated in section 3.3.3 where not grouping by time interval resulted in a very different value of the statistic. Specific guidelines on a general strategy for choosing groups are difficult to determine. Where possible it is desirable to have separate groups by time interval length, observation time or number, covariate values and (in the case of misclassification HMMs) previous pattern of observed states. However, it will not usually be possible to have separate groups for all these categories as the resulting contingency table would have many empty cells or low counts. The presence of multiple covariates is particularly problematic. The approach taken in this chapter is to use the estimated transition intensities $q_{rs}(z, \hat{\theta})$, to generate a partition of the space of covariates. This is straightforward if the covariates only affect a single transition intensity. However, if multiple transition intensities are affected, possibly in conflicting ways, a suitable partition is less clear. One possibility is to use the quantiles of $\sum_{r=1}^{R} -q_{rr}(z, \hat{\theta})$. This gives an indication of how prone to making transitions between states a subject is. However, especially for bi-directional models, there may be little correspondence between $\sum_{r=1}^{R} -q_{rr}(z, \hat{\theta})$ and the expected transitions $p_{rs}(t; z, \hat{\theta})$.

In this chapter we have only considered assessment of model fit in time homogeneous Markov models. Time inhomogeneous Markov models can however be assessed in exactly the same way just by substituting the appropriate expected transition probabilities, $p_{rs}(t_1, t_2)$ rather than $p_{rs}(t_2 - t_1)$ into the expressions. There should be some grouping by the start time of the interval (often observation number will suffice) in the construction of the contingency table to avoid information loss.

The application of the general goodness-of-fit to the CAV data showed that the fit of a time homogeneous Markov model is poor. The main areas of discrepancy are excess $1 \rightarrow 2$ and $1 \rightarrow 3$ transitions in short intervals. This implies that there is some time inhomogeneity in the CAV onset rates. The CAV example illustrates the advantages of a formal test as the problems of fit were not identified using the informal methods of chapter 2.

# Chapter 4

# The effect of model misspecification

The initial chapters of this thesis developed methods for assessing the fit of a multi-state model. However, the extent to which goodness-of-fit is important depends on the effect model misspecification has on the inferences that are drawn. Therefore, the effect of deviations from the standard time homogeneous Markov model is a worthwhile area of investigation. This chapter aims to analyse the probable bias and impact on inference of model misspecification in some fairly realistic examples and tries to make some comparison between the impact of sources of misspecification.

While simulation studies are avoided, even using asymptotic approximations, analytic expressions for bias and expected coverage of confidence intervals are not available. This makes it very difficult to make general conclusions about the direction and size of bias for particular sources of misspecification. It is also difficult to make direction comparisons between sources of misspecification. The potential importance of a particular type of misspecification will depend on the level of misspecification, for instance what level of time inhomogeneity there is relative to unobserved patient heterogeneity. This will depend on the particular application.

However, the chapter does develop a general framework, which could be adapted to investigate the potential for bias through model misspecification. The chapter will review the existing literature on the subject and then set out the methods to be used. Three main types of misspecification are considered; non-exponential sojourn distributions, unidentified patient heterogeneity and dependent misclassification in HMMs.

## 4.1 Previous investigations of misspecification

This section reviews the existing literature on the effect of model misspecification. In the general statistical literature, Cox [33] derived asymptotic results for the effect of assuming a random variable comes from a different family of distributions than the true family. White [134] gave a detailed general analysis of maximum likelihood estimation under a misspecified model.

There is a wide literature on model misspecification in right-censored survival and reliability models, which is to some degree related to multi-state models. The main focus is on the effect of departures from the proportional hazards assumption in Cox models [9, 126, 130].

Using simulation Li *et al* [86] considered the effect on tests of significance of covariate effects for a proportional hazards model of assuming one common parametric form (exponential, Weibull, log-logistic and log-normal) when the true underlying hazards were some other parametric form (Log-Laplace, Gamma, Gompertz or Exponential-power), specifically in the context of small sample sizes. They found that assuming an exponential model tended to produce tests of highly overestimated size. Weibull and log-logistic models had a reasonably accurate size in most cases.

In Weibull reliability models, due to the difficulty of estimating the shape parameter, a fixed known shape parameter is often assumed. Keats *et al* [74] investigated the effect of misspecification of the shape parameter in this context and found that relatively small deviations from the true shape parameter, could lead to very poor coverage of confidence intervals for the rate parameter.

There are few studies focusing on model misspecification for interval censored or panel observed data from multi-state models.

### 4.1.1 Grüger *et al*: The effect of dependent sampling

Grüger *et al* [57] in the context of Markov models considered the effect of dependent sampling on the validity of inferences. This is an important area because the construction of the likelihood for panel observed data depends on the assumption that the observation process can be ignored. In the paper they establish that inference is only valid when the sampling scheme is *non-informative*, which they define to be when either:

1. the probability of being in state $s_j$ at time $T_j$ given the history of both the disease process and the observational process up until time $T_{j-1} = t_{j-1}$ is independent of whether an examination is carried out at this time and the past examination times

2. the distribution of the $j$th examination time $T_j$, conditional on the history of both processes up to time $T_{j-1} = t_{j-1}$, is functionally independent of parameters governing the transition intensities of $X$, where $X$ is the disease process, the examination times are $T_1, ..., T_n$ and the number of observations $n$ is itself a random variable.

Feasible sampling schemes are considered against these criteria. Examination at regular intervals and random sampling of observation times (where the sampling and disease processes are independent) clearly meet the criteria for non-informativeness. It is also shown that 'doctor's care' - defined as a scheme in which the $j$th observation time is decided at the $(j-1)$th observation and may depend on the past history of the process up to and including time $t_{j-1}$ - is non-informative. However, patient self-selection, for instance when a patient who fells unwell is more inclined to seek an examination, is shown to violate non-informativeness. They provide a simulation study based on a 4 state disease model originally used by Kay [73] for liver cancer survival. States 1,2,3 represent disease free, mild disease and severe disease respectively. State 4 is death which can be reached from any other state. Transitions between adjacent disease states are possible in both directions (figure 1.6).

Patient self-selection in the simulation is defined by parameters $\theta_i$, $i = 1, 2, 3$ where

$$\mathbb{P}(\text{patient examined at } t | \text{patient in state } i \text{ at } t) = \theta_i$$

and where there is a candidate $t$ every 40 days.

The simulations concentrated on the case where the self-selection was such that a patient was more likely to be observed at a potential observation time if their disease state at that time was more advanced. In those conditions there is overestimation of the transition intensity between state 1 and state 2. In addition, transition intensities to death from state 1 are underestimated, while those from state 2 are overestimated.

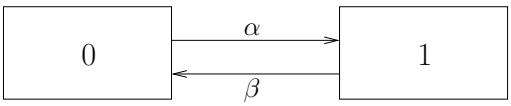
The level of bias, even for relatively modest patient self-selection and for a moderately large sample size, is more than sufficient to cause the confidence intervals to have very poor coverage. The assumption of non-informative sampling therefore has great potential for producing biased estimates. However, Grüger *et al* also point out that with the given data alone it is not possible to test the non-informativeness of an examination scheme.

### 4.1.2 Rosychuk and Thompson: Markov assumption when there is misclassification

Rosychuk and Thompson [110] considered the effect of assuming a Markov model for a binary process when realisations of the process were subject to misclassification.

This scenario is only realistic for models where the underlying Markov process admits reverse transitions, as misclassification will lead to some apparent reverse transitions in the data. The simplest case is a two-state model where patients alternate between states 0 and 1 (e.g. healthy and ill), this is also the case dealt with by Rosychuk and Thompson (figure 4.1).

Figure 4.1: Two-state model used by Rosychuk and Thompson




When patients are observed at equally spaced time intervals of fixed length $t$ and the process is assumed to be in its equilibrium distribution it is possible to parameterise using transition probabilities over a fixed interval $t$ rather than use transition intensities. The maximum likelihood estimates of each transition probability (i.e. $\alpha = p_{01}(t)$ and $\beta = p_{10}(t)$) under the Markov model are given simply by $\hat{\alpha} = \frac{N_{01}}{N_{01}+N_{00}}$ and $\hat{\beta} = \frac{N_{10}}{N_{10}+N_{11}}$ where $N_{ij}$ represents the number of $i \to j$ transitions observed. If in fact there is misclassification such that $e_{ij} = \mathbb{P}(O = j|X = i)$ then Rosychuk and Thompson show that the asymptotic bias of these estimators is given by

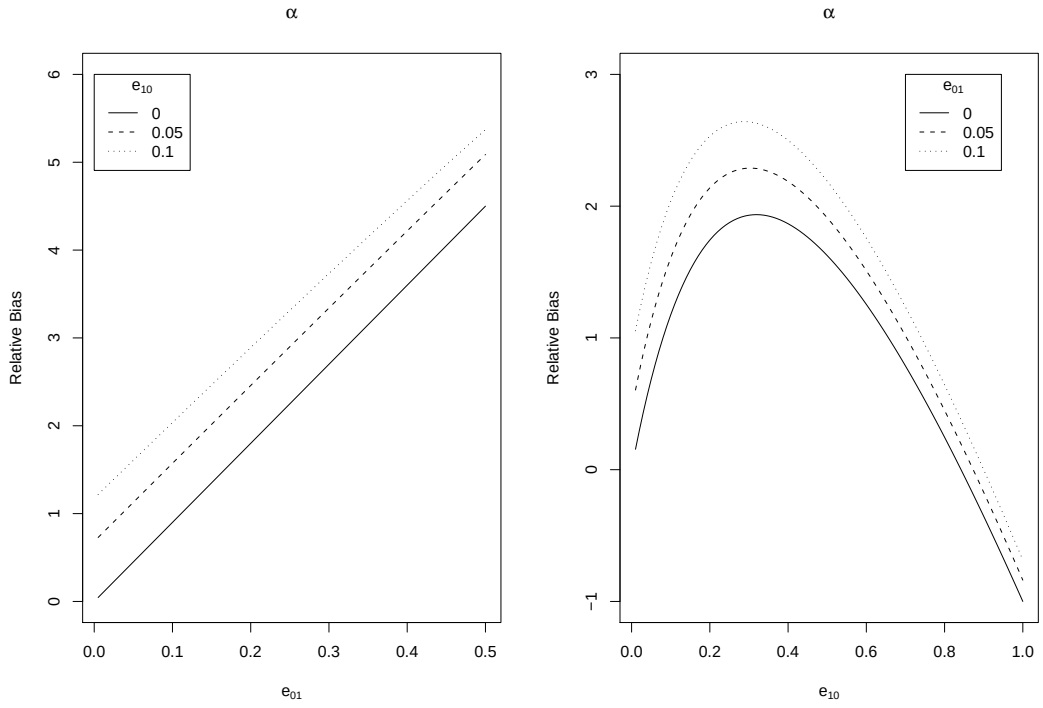
$$\frac{\mathbb{E}(N_{01})}{\mathbb{E}(N_{01}) + \mathbb{E}(N_{00})} - \alpha = \frac{(e_{01} - e_{11})^2\alpha\beta + e_{00}e_{01}\beta + e_{11}e_{10}\alpha}{e_{10}\alpha + e_{00}\beta} - \alpha$$

and

$$\frac{\mathbb{E}(N_{10})}{\mathbb{E}(N_{10}) + \mathbb{E}(N_{11})} - \beta = \frac{(e_{01} - e_{11})^2\alpha\beta + e_{00}e_{01}\beta + e_{11}e_{10}\alpha}{e_{11}\alpha + e_{01}\beta} - \beta.$$

When transition probabilities ($\alpha$ and $\beta$) are small, any misclassification leads to overestimation of both the transition probabilities. Variation in misclassification parameter $e_{ij}$ has a direct impact on the estimate of $p_{ij}$, an increase in $e_{ij}$ causes $p_{ij}$ to be overestimated to a greater extent. Its impact on $p_{ji}$, the opposing transition, is less direct, for small $e_{ij}$, $p_{ji}$ is overestimated. However for large enough $e_{ij}$, $p_{ji}$ will become underestimated because so many observed states will be $j$ (figure 4.2).

Figure 4.2: Relative bias in estimate of $\alpha$, for varying $e_{01}$ and $e_{02}$ when $\alpha$ and $\beta$ are small ($\alpha = 0.1$, $\beta = 0.05$, fixed)




For larger $\alpha$ or $\beta$, the estimates remain biased but there is a less clear pattern in the sign of the bias.

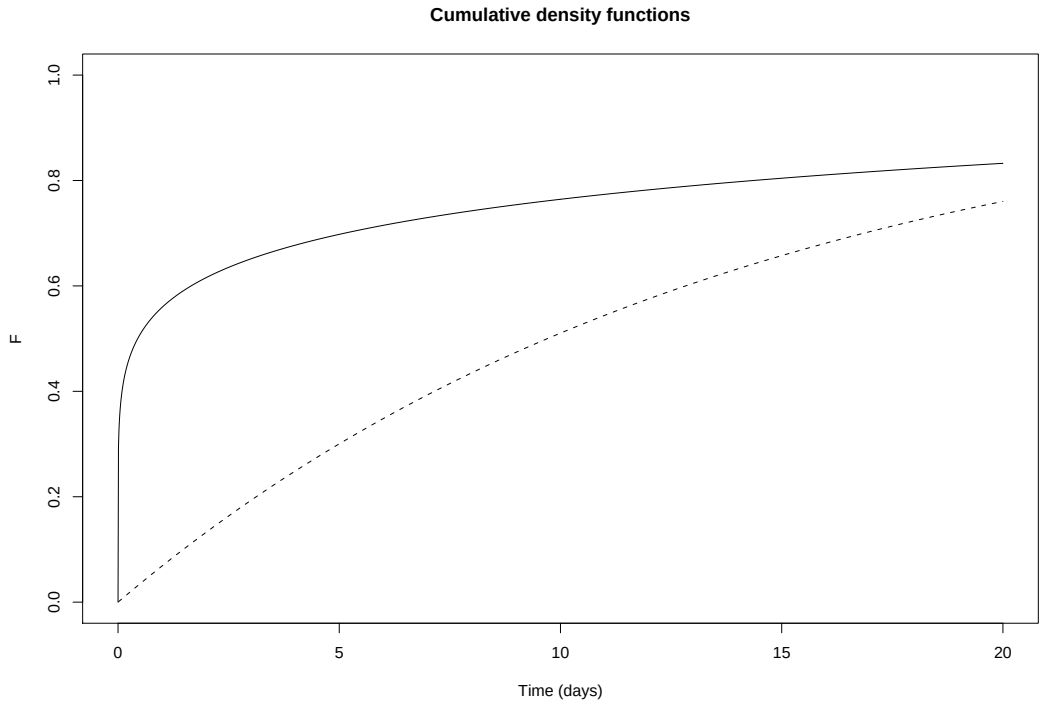
Misspecifying a misclassification HMM as a Markov model can clearly cause considerable problems in terms of inference. Fortunately, methods are available for fitting the HMM, so this type of misspecification can be identified, at least for models with a small number of states.

### 4.1.3 Rosychuk and Thompson: Time and subject heterogeneity

In a separate paper [111] which mainly focused on parameter identifiability, the same authors considered more general model misspecification. They again consider a two-state recurrent model with misclassified observed states. Realisations from Gamma distributed sojourn times were generated, with patient-specific rate parameters and common shape parameters for the Gamma distribution. Around 10% of subjects were defined as outliers

in the sense that they were observed either in state 0 at each sampling time or in state 1 at each sampling time. The data were assumed to come from a time homogeneous misclassification HMM in which the population was also homogeneous. The estimated mean sojourn times were found to be 453.9 days and 334.4 days for state 0 and state 1 respectively. The 'true' mean sojourn times for the generated data were around 14 days (ranging between 10 and 18 days) for each state. The effect of model misspecification in this case therefore seemed to be quite extreme. However, the choice of Gamma sojourn distribution, with a shape parameter between 0.1 and 0.18, was extreme. This degree of misspecification would be immediately spotted in practice.

Figure 4.3: Comparison of the cumulative distribution functions for $\Gamma(0.14, 0.01)$ (bold line) and Exp($\frac{1}{14}$) (dashed line) random variables




Subjects were observed at 7 day intervals. For the $\Gamma(0.14, 0.01)$ sojourn distribution chosen, there is both a non-negligible chance that multiple transitions occurred between observations (40% of sojourn times are less than 0.1 days), and that the subject stays in the same state throughout the observation period (5% of sojourn times are over 78 days). Figure 4.3 gives a comparison between the cdf of $\Gamma(0.14, 0.01)$, compared to an Exponential($\frac{1}{14}$)

which has the same mean sojourn time. A $\Gamma(0.14, 0.01)$ sojourn distribution is quite unrealistic and it is perhaps not surprising that assuming Exponential sojourn times for these data would result in significant bias.

## 4.2 Mathematical issues and methodology

In this chapter the investigation is done without using simulation. We will instead rely on asymptotic results, which will be used to provide approximations of bias and true coverage of 95% confidence intervals, for moderate sample sizes.

### 4.2.1 Asymptotic theories

In more general settings, results about the mean and variance of the maximum likelihood estimates under a misspecified model can be derived, following the arguments of Cox [33] and White [134]. Suppose the data are assumed to be from some probability model with parameters $\beta \in \mathcal{B}$, giving misspecified likelihood function $\tilde{l}(\beta; x)$, but in fact they are from some other probability model with parameters $\alpha \in \mathcal{A}$.

Then there exists a value $\beta_\alpha$ that satisfies

$$\hat{\beta} \xrightarrow{p} \beta_\alpha \tag{4.1}$$

where $\hat{\beta}$ is the maximum likelihood estimate of $\beta$ under the misspecified model and $\xrightarrow{p}$ denotes convergence in probability. Moreover, $\hat{\beta}$ can be shown to have asymptotic covariance matrix

$$\Sigma_\alpha = \mathbb{E}_\alpha(\tilde{I}(\beta_\alpha))^{-1} V_\alpha \mathbb{E}_\alpha(\tilde{I}(\beta_\alpha))^{-1} \tag{4.2}$$

where

$$V_\alpha = \mathbb{E}_\alpha(\tilde{U}(\beta_\alpha)\tilde{U}^T(\beta_\alpha))$$

and $\tilde{U}$ and $\tilde{I}$ are the score and Fisher information respectively under the misspecified likelihood. Overall

$$\hat{\beta} \xrightarrow{d} \mathrm{N}(\beta_\alpha, \Sigma_\alpha).$$

A derivation of this result is given in Appendix C.

Application of this result requires the calculation of the expectation of various functions of the likelihood or misspecified likelihood. To obtain these, the following general approach can be applied in many settings.

We suppose that an individual is observed at a series of observation times $t_1, \ldots, t_n$. At each of these observation times there is a response, $x_1, \ldots, x_n$, corresponding either to $x_i = X(t_i)$ in the case of a Markov model or $x_i = O(t_i)$ in the case of a misclassification HMM. Given that $X$ or $O$ can only take a finite number of values, corresponding to $|\mathcal{S}|$ - the dimension of the state space of the Markov model, there is only a finite number of possible response vectors $\mathbf{x} = (x_1, \ldots, x_n)$ that can arise. Hence, the expected likelihood and Fisher information functions can be obtained by considering,

$$\mathbb{E}(l(\theta)) = \quad \sum_{\mathbf{x}\in\mathcal{X}} l(\theta|\mathbf{x})\mathbb{P}(\mathbf{x})$$

$$\mathbb{E}(I(\theta)) = \quad \sum_{\mathbf{x}\in\mathcal{X}} -\frac{\partial^2 l(\theta|\mathbf{x})}{\partial\theta^T\partial\theta}\mathbb{P}(\mathbf{x})$$

where $\mathbb{P}(\mathbf{x})$ is the probability of a particular response $\mathbf{x} \in \mathcal{X}$. Other quantities of interest, such as $\mathbb{E}(U(\theta)U(\theta)^T)$, can be found in the same way. To find the asymptotic limit of $\hat{\theta}$, in cases where a closed form expression is not possible, we can numerically optimise $\mathbb{E}(l(\theta))$.

This approach for determining the expected Fisher information was used by Hwang and Brookmeyer [63] to derive approximate sample size calculations for panel studies with different sampling schemes. A similar approach was used previously by de Stavola [127] to determine sampling designs for short panel data.

### 4.2.2 Evaluation of the impact of model misspecification

In the limit as the number of subjects tends to infinity, the coverage of a 95% confidence interval based on a biased estimator, will be 0. However, if we assume that the asymptotic approximation is adequate for moderate to large sample sizes, we can get approximations for the coverage of a 95% confidence interval for the quantity of interest under a misspecified model for a specific sample size. For this we need the mean and variance of the estimator under the misspecified model. In addition, we need the expected Fisher information under the assumed likelihood, which we shall denote as $\tilde{I}$. Suppose we are interested in some quantity which under the misspecified model is given by $g(\beta)$. Following the results in section 4.2.1, we have that asymptotically

$$g(\hat{\beta}) \sim \mathrm{N}(g(\beta_\alpha), \psi_N^2)$$

where

$$\psi_N^2 = (\frac{\partial g}{\partial \beta})^T \Sigma_\alpha (\frac{\partial g}{\partial \beta}),$$

but under the misspecified model it is assumed to be

$$g(\hat{\beta}) \sim \mathrm{N}(g_0, \tau_N^2)$$

where

$$\tau_N^2 = (\frac{\partial g}{\partial \beta})^T \tilde{I}^{-1} (\frac{\partial g}{\partial \beta}),$$

$g_0$ is the true value of the quantity of interest and $N$ is the number of subjects. If $g$ is a scalar quantity then the 95% confidence interval under the assumed model has the form

$$g(\hat{\beta}) \pm 1.96\tau_N.$$

This confidence interval has approximate coverage given by

$$\Phi(\frac{g_0 - g(\beta_\alpha) + 1.96\tau_N}{\psi_N}) - \Phi(\frac{g_0 - g(\beta_\alpha) - 1.96\tau_N}{\psi_N}).$$

If $g_0 \neq g(\beta_\alpha)$, meaning the estimate is biased, the coverage of the confidence interval will tend to zero as $N \to \infty$. $\tau_N$ and $\psi_N$ depend on the overall sample size through their dependence on $\tilde{I}$ and $\Sigma_\alpha$.

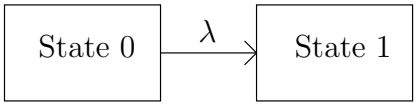
## 4.3 Misspecification of sojourn time distributions

This section considers the effect that non-exponential sojourn time distributions (producing semi-Markov processes) have on inference when a time-homogeneous Markov model is assumed. When there are only two states in the model and the initiation time of the process is known, the semi-Markov model is equivalent to a time-inhomogeneous Markov model.

When a time-homogeneous model is assumed, it is natural to parametrise in terms of the transition intensities. If the data arise from a process that does not have constant transition intensities there are no parameters in a time dependent model that are comparable. Therefore, this parameter is not useful for assessing the effect of model misspecification. If instead the parameter of interest is taken to be the mean sojourn time, then this has a clear interpretation for both Markov and semi-Markov models. Hence, throughout this section we take mean sojourn time as the quantity of interest.

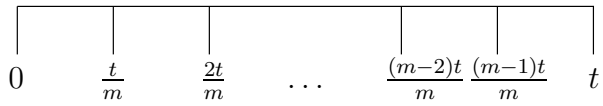
### 4.3.1 Two state model, repeated regular observations

Figure 4.4: Two state disease model




Consider firstly a two state model where patients all begin in state 0 at time 0 and can proceed only to state 1 which is an absorbing state and do so with constant transition intensity $\lambda$ (figure 4.4). Let the subjects be observed up to $m$ times at regular intervals of length $\frac{t}{m}$ up to a maximum time $t$ as shown in figure 4.5.

Figure 4.5: Sampling scheme for repeated observations




Under this model and observation scheme, there are only $m+1$ distinct sets of observations an individual can give. These correspond to the $m$ possible intervals within which the subject could make the transition to state 1, plus the case of no observed transition before time $t$. Hence the log-likelihood for a single individual is given by

$$l(\lambda) = \begin{cases} -\frac{(i-1)t}{m}\lambda + \log\left(1 - \exp\left(-\lambda\frac{t}{m}\right)\right) & i = 1, \ldots, m \\ -\lambda\frac{t}{m} & i = m+1 \end{cases} \tag{4.3}$$

for transition in interval $i = 1, \ldots, m$ and $i = m+1$ denotes censoring in state 0 at time $t$.

It follows that the expected likelihood contribution for an individual is given by

$$\mathbb{E}l(\lambda) = -\lambda\frac{t}{m}\sum_{i=1}^{m+1}(i-1)p_i + (\sum_{i=1}^{m} p_i)\log\left(1 - \exp\left(-\lambda\frac{t}{m}\right)\right), \tag{4.4}$$

where $p_i$ represents the probability of a $0 \to 1$ transition occurring in the $i$th observation interval under the true model and $p_{m+1}$ denotes the probability of being censored in state 0 at time $t$, so that $\sum_{i=1}^{m+1} p_i = 1$. When state 0 has a general sojourn distribution with

pdf $f(t)$,

$$p_i = \begin{cases} \int_{\frac{(i-1)t}{m}}^{\frac{it}{m}} f(t)dt & i = 1, \ldots, m \\ \int_t^\infty f(t)dt & i = m+1 \end{cases}.$$

Let $\nu = \sum_{i=1}^{m+1}(i-1)p_i$ and $\rho = (\sum_{i=1}^{m} p_i)$. The point $\lambda_\alpha$ to which the mle for $\lambda$ converges can be found by differentiating 4.4 and solving for $\frac{\partial \mathbb{E}l(\lambda)}{\partial \lambda} = 0$. This gives

$$\hat{\lambda} \xrightarrow{p} \frac{m}{t} \log\left(\frac{\nu+\rho}{\nu}\right). \tag{4.5}$$

Under the misspecified model, the observed Fisher information is

$$\tilde{I}(\lambda) = \frac{t^2 \exp(-\frac{\lambda t}{m})}{m^2(1-\exp(-\frac{\lambda t}{m}))^2},$$

if the subject goes to state 1 before time $t$, and is zero otherwise. Hence the expected Fisher information is

$$\mathbb{E}(\tilde{I}(\lambda)) = \rho \frac{t^2 \exp(-\frac{\lambda t}{m})}{m^2(1-\exp(-\frac{\lambda t}{m}))^2}.$$

In order to apply formula 4.2 we also need $\mathbb{E}U(\lambda)^2$.

$$U(\lambda) = \frac{t}{m}(i-1) + \left(\frac{t\exp(-\lambda\frac{t}{m})}{m(1-\exp(-\lambda\frac{t}{m}))}\right)^{\mathbb{1}(i<m+1)},$$

for $i = 1, \ldots, m$, so $\mathbb{E}(U(\lambda)^2)$ can be expressed as a double summation. An analytic expression for the asymptotic covariance of $\hat{\lambda}$ can therefore be found, although it is too complicated to provide any insight into its form.

So far we can see that the amount of bias will depend on the amount of misspecification, the time of censoring $t$ and the number of times the patient is observed $m$. The sign of the bias is also dependent on these factors. However, to learn what effect each factor may have, it is necessary to consider specific examples. A convenient and flexible family of distributions to consider for sojourn time distributions is the Gamma distribution represented by $\Gamma(\alpha, \beta)$ which has a probability density function:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}\beta^\alpha \exp(-\beta x)}{\Gamma(\alpha)}, x > 0$$

where $\alpha$ is the shape parameter and $\beta$ is the rate parameter and $\Gamma(.)$ is the gamma function.

The mean of this distribution is $\frac{\alpha}{\beta}$. When $\alpha = 1$ the distribution is Exponential, if $\alpha < 1$ then the hazard is decreasing with time and if $\alpha > 1$ then the hazard increases with time. The cumulative distribution function can be calculated numerically but does not exist in closed form. Alternatively, we could have chosen the Weibull distribution as it has similar properties to the Gamma distribution in terms of having Exponential as a special case and resulting in either increasing or decreasing hazards.

**Results**

Figure 4.6 gives a contour plot of the bias from assuming exponential sojourn times for varying right censoring time $t$ and shape parameter $\alpha$, with rate parameter $\beta = 1.5$. It shows that when $\alpha > 1$ (increasing hazards) the sojourn time will be overestimated if the right censoring occurs before a certain time dependent on the particular value of $\alpha$, and will be underestimated otherwise. If $\alpha < 1$ then the sojourn time will be underestimated if the right censoring is before a certain threshold time dependent on $\alpha$ and overestimated otherwise. The point at which the sign of the bias changes depends on the number of observations $m$, and the rate parameter $\beta$. As $m$ increases the time at which the sign change occurs also increases.

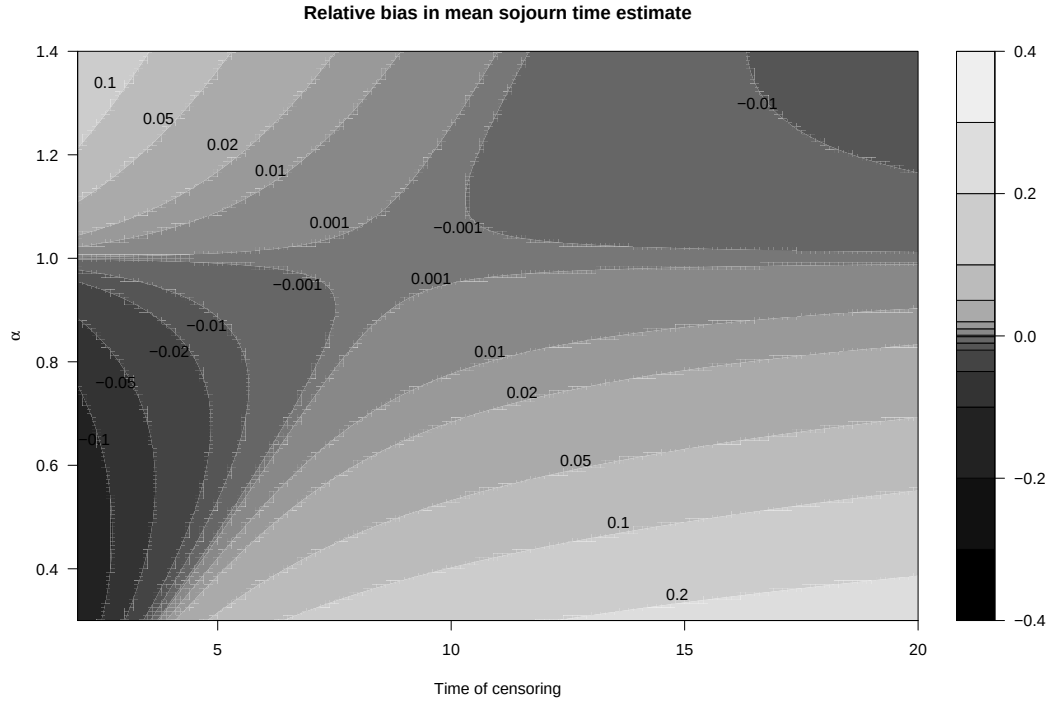
Note also that for small enough $\alpha$, the sojourn time will be overestimated regardless of the censoring time. The behaviour of the estimates for extreme values of the shape parameter, $\alpha$, is perhaps not of great importance either because such distributions are less likely to be seen in real data or because it would be unlikely that an exponential model would be chosen in such cases.

A particularly interesting result is that, for $\alpha$ reasonably close to 1, as the number of intermediate observations $m$ increases, the magnitude of bias also increases (figure 4.7).

Intuitively this can be explained by considering the shape of the pdf of a Gamma distribution. The largest differences from an Exponential distribution occur for small values of $t$. In particular, for $\alpha > 1$, if the observations are taken after the peak in the Gamma pdf, the data will appear close to an Exponential curve. In contrast, if observations are taken mostly before the peak, the bias will be much more pronounced. This can be seen from the shape of the pdfs of the respective distributions (figure 4.8).

Table 4.1 gives the approximate coverage of 95% confidence intervals constructed using the Fisher information from the misspecified likelihood. The estimates of mean sojourn

Figure 4.6: Contour plot of bias in mean sojourn time for two state model with a true gamma distribution for $\beta = 1.5$ and $m = 10$



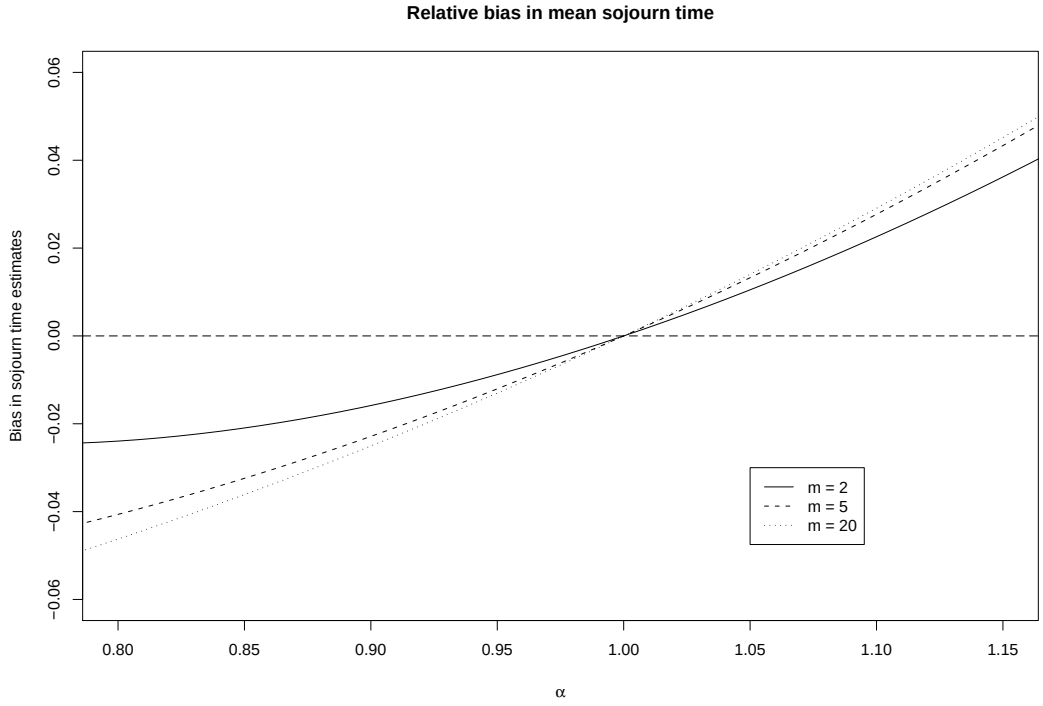

time are biased when $\alpha \neq 1$, so the coverage tends to zero as $N \to \infty$, where $N$ is the number of subjects. However, for $\alpha < 1$, the variance is slightly underestimated so the decay in coverage as $N$ increases is quite rapid. In contrast, as with the full information case, when $\alpha > 1$, the misspecified variance is an overestimate. Hence for small samples the coverage is greater than 95%, and the decay in coverage is much less marked.

### 4.3.2 More than two states

So far we have only considered interval censored data for a two-state model, which is only nominally multi-state. However, once there are more than two states, it is no longer possible to get closed form solutions to the maximum likelihood equations. Instead we maximise $\mathbb{E}(l(\lambda))$ numerically.

Consider a three state unidirectional model as shown in figure 4.9. Suppose that as before, subjects are observed at intervals of length $\frac{t}{m}$, up to a maximum time of $m$. Each

Figure 4.7: Bias in mean sojourn time for varied $\alpha$ and $m$, when $t = 3$ and $\beta = 2.5$. Positive values imply a mean sojourn time is overestimated.

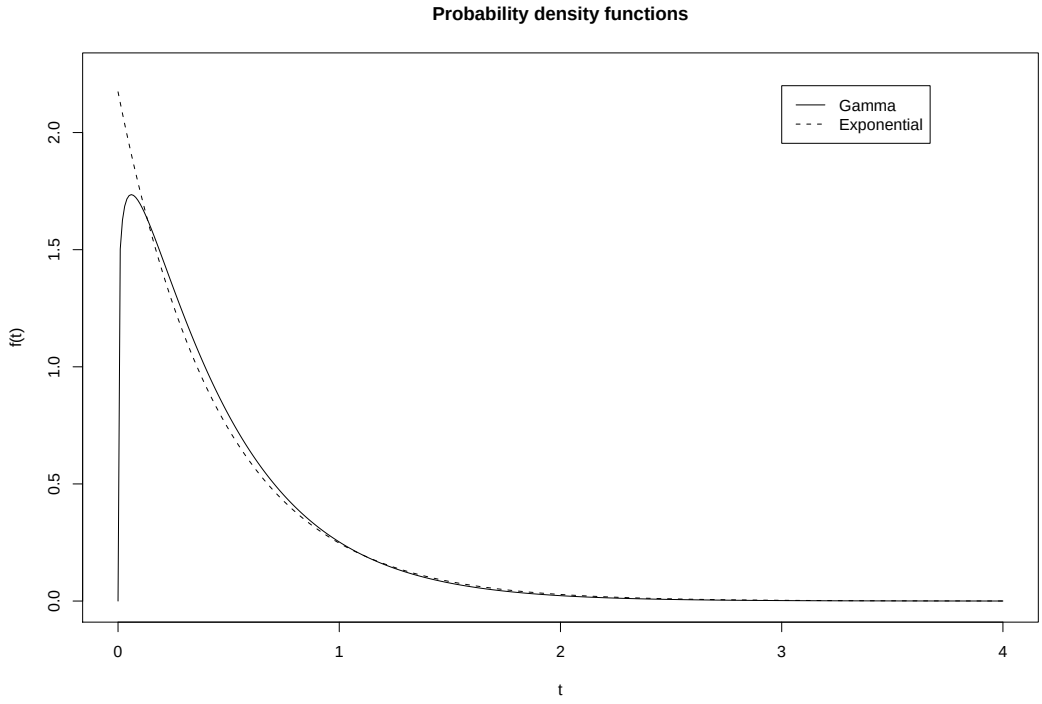



subject's history can be characterised by the time of the first state 2 and the first state 3 observations. Conditional on the first state 3 being on the $j$th observation, for $j > 1$, there are $j - 1$ possible positions for the first state 2, plus the possibility of no observed state 3. If no state 3 was observed there are $m$ possible positions for the first state 2, plus the possibility of not reaching state 2. The number of possible responses from an individual is therefore

$$\sum_{j=1}^{m+1} j = \frac{1}{2}(m+1)(m+2).$$

To proceed we need to consider the probability of observing each of the possible responses under the true model. If one or both the states have Gamma sojourn times, these probabilities are not available in closed form. However, if only state 1 has a Gamma distribution but state 2 is given an Exponential distribution then in certain cases, the required integrals can be expressed in terms of c.d.f's of Gamma distributions. For instance, if a response involves moving from state 1 to state 2 in the interval $(a_1, b_1)$ and state 2 to state 3 in

Figure 4.8: Comparison of the pdfs of a $\Gamma(1.15, 2.5)$ and a exponential with the same mean. The largest difference is near time zero.




interval $(a_2, b_2)$, where $a_1 < b_1 \leq a_2 < b_2$. The probability of this event can be written as

$$\int_{a_1}^{b_1} f_1(s) \int_{a_2-s}^{b_2-s} f_2(u) du ds \tag{4.6}$$

where

$$f_1(s) = \frac{\beta_1^{\alpha_1} s^{\alpha_1 - 1} \exp(-\beta_1 s)}{\Gamma(\alpha_1)}$$

and $f_2(u) = \lambda_2 \exp(-\lambda_2 u)$. We can rewrite this as

$$(\exp(-\lambda_2 a_2) - \exp(-\lambda_2 b_2)) \int_{a_1}^{b_1} \frac{\beta_1^{\alpha_1} s^{\alpha_1 - 1} \exp(-(\beta_1 - \lambda_2)s)}{\Gamma(\alpha_1)} ds. \tag{4.7}$$

In the case where $\beta_1 > \lambda_2$, we can rewrite this integral as

$$(\exp(-\lambda_2 a_2) - \exp(-\lambda_2 b_2))(\frac{\beta_1}{\beta_1 - \lambda_2})^{\alpha_1} (F(b_1; \alpha_1, \beta_1 - \lambda_2) - F(a_1; \alpha_1, \beta_1 - \lambda_2)$$

where $F(x; \alpha, \beta)$ is the cdf of a $\Gamma(\alpha, \beta)$. Thus when $\lambda_2 < \beta_1$, the probabilities are obtainable numerically. Moreover, the one-dimensional integral in (4.7), can be obtained via

Table 4.1: Approximate coverage of assumed 95% confidence intervals on mean sojourn time for varying $\alpha$, fixed $t = 5$, $m = 5$.

| | $N$ | | | |
|---|---|---|---|---|
| $\alpha$ | 100 | 500 | 1000 | 2000 |
| 0.6 | 0.881 | 0.736 | 0.571 | 0.322 |
| 0.8 | 0.932 | 0.914 | 0.891 | 0.846 |
| 0.9 | 0.943 | 0.940 | 0.936 | 0.927 |
| 1.1 | 0.955 | 0.953 | 0.951 | 0.947 |
| 1.2 | 0.959 | 0.954 | 0.947 | 0.933 |
| 1.4 | 0.965 | 0.952 | 0.936 | 0.900 |

Figure 4.9: Three state unidirectional model


numerical quadrature, for instance using the inbuilt function **integrate** in R. If $\lambda_2 > \beta_1$ the above method is not applicable: two-dimensional numerical quadrature could be applied to compute the probabilities in this case. However, singularities in the integrals make these results less reliable. We therefore only present results for $\lambda_2 < \beta_1$.

**Results for a semi-Markov initial state**

The most important result in this case is that the biases in the estimates of sojourn time in state 1 (the state with the misspecified sojourn distribution), closely resemble the bias observed in the two-state case. The resulting contour plot of bias in state 1 mean sojourn time closely resembles figure 4.6 for the 2 state case. Also the 95% confidence intervals have a similar pattern of coverage (table 4.2).

The effect on state 2 (with an Exponential sojourn distribution), is more complicated. Mostly, the bias in state 2 estimates are low. Figure 4.10 shows the bias in estimated state 2 sojourn times in the cases $\alpha = 0.8$ and $\alpha = 1.2$ for varying $t$ and $m$. In this example even when the observations are very spaced out, the relative bias doesn't exceed 2%. The main driver of bias is the proportion of observed $1 \to 3$ transitions. If the observations

are closely spaced, there are few $1 \rightarrow 3$ transitions observed and consequently the bias is small. When the observations are widely spaced there are a greater number of $1 \rightarrow 3$ transitions, and fewer $2 \rightarrow 2$ and $2 \rightarrow 3$ transitions, so the bias is larger.

Except in the case of extreme censoring (i.e. when the proportion of subjects reaching state 2 is negligible), the sign of the bias of the state 2 sojourn time corresponds to whether the hazard in state 1 is increasing or decreasing. If $\alpha < 1$, the sojourn time in state 2 is under estimated, but if $\alpha > 1$, it is an over estimated.

Figure 4.10: Bias in estimates of state 2 sojourn time in three-state model when state 1 has a Gamma distribution. Positive values imply the mean sojourn time is overestimated.

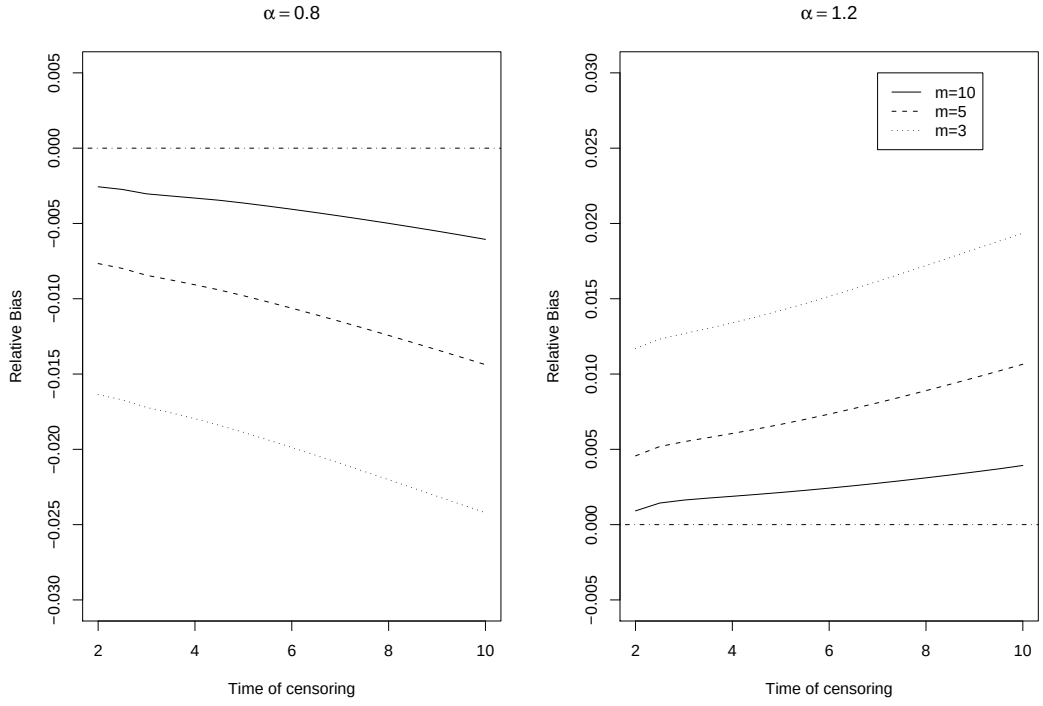



Table 4.2 gives the approximate coverage of 95% confidence intervals on mean sojourn time in states 1 and 2, for a range of values of $\alpha$. For state 1, the approximate coverages are virtually identical to those of table 4.1. For state 2 we see that the bias in the estimates is largely negligible. Even when $\alpha = 0.6$, coverage for $N = 2000$ is still around 87% when $m = 5$.

Table 4.2: Approximate coverage of assumed 95% confidence intervals on mean sojourn times for varying $\alpha$, fixed $t = 5$, $m = 5$ in a 3-state model with misspecified first state.

| | | $N$ | | | |
|---|---|---|---|---|---|
| State | $\alpha$ | 100 | 500 | 1000 | 2000 |
| 1 | 0.6 | 0.880 | 0.731 | 0.563 | 0.310 |
| | 0.8 | 0.932 | 0.913 | 0.890 | 0.842 |
| | 0.9 | 0.943 | 0.939 | 0.935 | 0.927 |
| | 1.1 | 0.955 | 0.953 | 0.951 | 0.946 |
| | 1.2 | 0.959 | 0.953 | 0.946 | 0.932 |
| | 1.4 | 0.965 | 0.952 | 0.935 | 0.898 |
| 2 | 0.6 | 0.944 | 0.929 | 0.910 | 0.869 |
| | 0.8 | 0.949 | 0.946 | 0.943 | 0.937 |
| | 0.9 | 0.950 | 0.949 | 0.948 | 0.947 |
| | 1.1 | 0.950 | 0.950 | 0.950 | 0.949 |
| | 1.2 | 0.950 | 0.949 | 0.948 | 0.946 |
| | 1.4 | 0.950 | 0.947 | 0.944 | 0.937 |

Table 4.3: Approximate coverage of assumed 95% confidence intervals on mean sojourn times for varying $\alpha$, fixed $t = 5$, $m = 5$ in a 3-state model with misspecified second state.

| | | $N$ | | | |
|---|---|---|---|---|---|
| State | $\alpha$ | 100 | 500 | 1000 | 2000 |
| 1 | 1.1 | 0.950 | 0.950 | 0.950 | 0.950 |
| | 1.2 | 0.950 | 0.950 | 0.950 | 0.950 |
| | 1.4 | 0.951 | 0.951 | 0.950 | 0.950 |
| | 1.6 | 0.951 | 0.950 | 0.950 | 0.948 |
| 2 | 1.1 | 0.949 | 0.931 | 0.909 | 0.863 |
| | 1.2 | 0.940 | 0.868 | 0.774 | 0.590 |
| | 1.4 | 0.916 | 0.713 | 0.474 | 0.171 |
| | 1.6 | 0.878 | 0.494 | 0.183 | 0.016 |

**Results for a semi-Markov second state**

If state 2 is allowed to have a Gamma sojourn distribution, it is not possible to reduce the two-dimensional integral in equation 4.6 to one that is one-dimensional. Provided $\alpha > 1$ in the Gamma sojourn distribution, adaptive numerical integration, such as that of the **adapt** package in R [56], can be used to compute the required double integrals. If $\alpha < 1$ however, the numerical integration fails to converge due to the singularity in the p.d.f. at $t = 0$. We shall therefore limit ourselves to cases in which the first state is exponential and the second state is Gamma with an increasing hazard (i.e. $\alpha > 1$).

In this case the effect of the misspecification in state 2 on the estimate of the mean sojourn time in state 1 is again negligible. Indeed, it is even less prominent than before. There is a small degree of over-estimation in the mean sojourn time in state 1, if state 2 has shape parameter $\alpha > 1$.

The pattern of bias in state 2 is similar to the 2-state case. When the time of right-censoring is small, the mean sojourn time is over estimated, when right-censoring is later, it is under estimated. Table 4.3 gives the approximate coverage of 95% confidence intervals for the case $t = 5$, $m = 5$.

The results in this section indicate that if only one state in a multi-state model is non-exponential, only parameters relating to that state will be significantly biased. Estimates

for other states will tend to have a small bias, except in extreme cases where the misspecification is very large or the observation scheme is very sporadic. The sensitivity of other states may be higher in more complicated models than a unidirectional model. In a two-state unidirectional model, a $1 \rightarrow 1$ transition is independent of state 2, whereas in a bi-directional model, $1 \rightarrow 1$ could imply multiple sojourns in state 2. Misspecification of one state is therefore more likely to affect the parameter estimates relating to other states in a bi-directional model.

## 4.4 Covariate effects

In many cases the transition intensities, and the mean sojourn times they imply, may be of secondary importance, and it is the effect of covariates that are of primary interest.

Consider again a 2-state model as in section 4.3.1 in which we focussed on estimates of the mean sojourn time. Here we consider the case of a binary covariate $y$, taking values 0 or 1, which has a multiplicative effect $\omega$ on the mean sojourn time. For simplicity, we shall assume an equal number of subjects for each covariate. When calculating the expected likelihood and Fisher information, it is therefore assumed subjects occur in pairs. Having a smaller proportion of 1s would increase the estimated standard error for the covariate effect, but relative biases would remain unaffected.

Under the assumption of exponential sojourn times the effect of the covariate on the transition intensity is: $\lambda_{y=0} = \lambda$ and $\lambda_{y=1} = \frac{\lambda}{\omega}$ If we parametrise the likelihood with respect to $\lambda_{y=0}$ and $\lambda_{y=1}$, then the log-likelihood is just the sum

$$l(\lambda_{y=0}, \lambda_{y=1}) = l_{y=0}(\lambda_{y=0}) + l_{y=1}(\lambda_{y=1}) \tag{4.8}$$

where $l_{y=i}$ denotes the likelihood contribution from patients for which $y = i$. Maximising this likelihood, or its expectation, with respect to $\lambda_{y=0}$ and $\lambda_{y=1}$ involves maximising each function separately. Therefore

$$\hat{\omega} = \frac{\hat{\lambda}_{y=0}}{\hat{\lambda}_{y=1}}$$

where $\hat{\lambda}_{y=0}$ and $\hat{\lambda}_{y=1}$ are the estimates of $\lambda$ for the patients with $y = 0$ and $y = 1$ respectively. Hence by applying equation 4.5 from section 4.3.1, we can get a general expression for the asymptotic mean of our estimate for $\omega$ when patients are observed up

to $m$ equally spaced times, up to a time $t$:

$$E(\hat{\omega}) \to \frac{\log\left(\frac{\nu_{y=0}-\rho_{y=0}}{\nu_{y=0}}\right)}{\log\left(\frac{\nu_{y=1}-\rho_{y=1}}{\nu_{y=1}}\right)}$$

where $\nu_{y=j}$ and $\rho_{y=j}$ for $j = 0, 1$ have the same definition as $\nu$ and $\rho$ in section 4.3.1, but for the cases $y = 0, 1$.

Similarly as a consequence of (4.8), the observed Fisher information can be written as a $(2 \times 2)$ matrix where the diagonal entries are of the form

$$I(\lambda_{y=j}) = \mathbb{1}(i_{y=j} \leq m)\frac{t^2 \exp\left(-\frac{\lambda_{y=i}t}{m}\right)}{m^2(1-\exp\left(-\frac{\lambda_{y=i}t}{m}\right))^2},$$

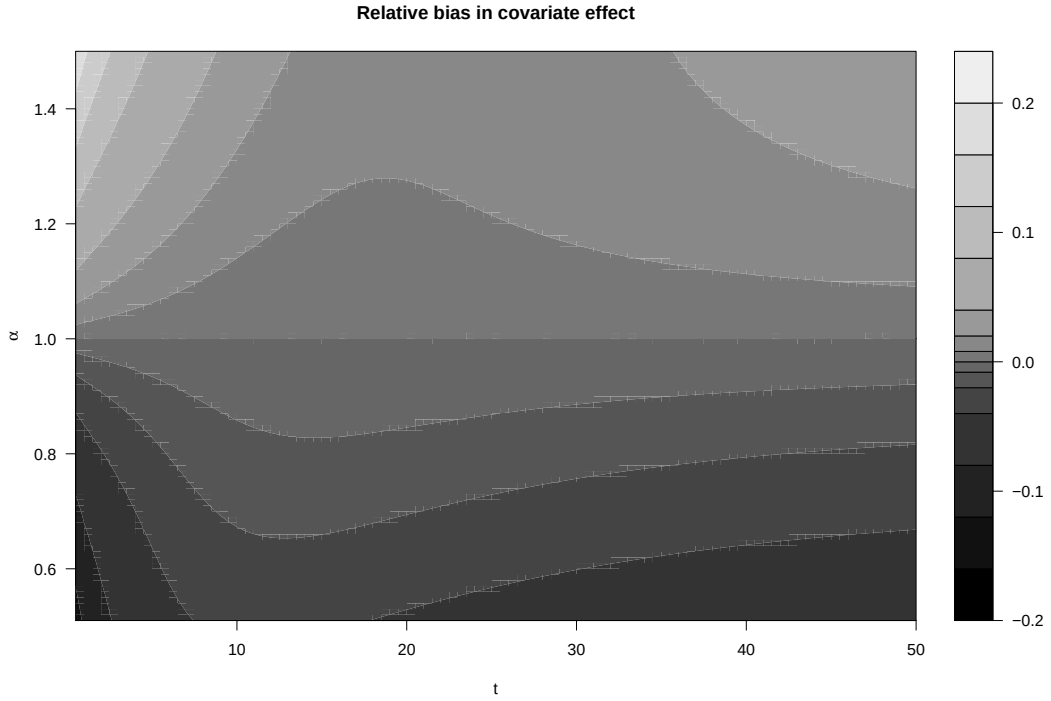
for $j = 1, 2$, where $i_{y=j}$ denotes which interval, for $i_{y=j} = 1, \ldots, m$, the transition $0 \to 1$ occurs, for an individual with $y = j$. The off-diagonal entries of the matrix are zero. The same techniques as in section 4.3.1 can be used to get the assumed confidence intervals.

If the sojourn times when $y = 0$ have a Gamma distribution with shape parameter $\alpha$ and rate parameter $\lambda$ then there is a choice of different parametrisations, all of which would maintain a multiplicative effect of $\omega$ on the mean sojourn time. The covariate could either affect the rate parameter exclusively, to give $\alpha$ and $\frac{\lambda}{\omega}$, or affect the shape parameter exclusively, to give $\alpha\omega$ and $\lambda$, or some mixture of the two, to give $\alpha\omega^{1-a}$ and $\frac{\lambda}{\omega^a}$, defined by an additional parameter $a$ which can take any real value. In this chapter we shall assume that the covariate only affects the rate parameter (i.e. $a = 1$). This means there is the same degree of misspecification from the assumption of Exponential sojourn times in each covariate group. However, the pattern of biases will not be the same for other values of $a$.

Assuming $a = 1$, if $t$ and $\alpha$ are varied then we get the straightforward result that when $\alpha < 1$ then $\omega$ is under estimated and when $\alpha > 1$ it is over estimated. There exists, for each value of $\alpha$, a value of $t$, lying strictly between 0 and $\infty$, for which the bias is minimised (figure 4.11). As the number of intermediate observations increases, the magnitude of bias decreases towards a limit (the bias from right censoring).

A key point to note is that, at least when $a = 1$, the mean estimate of the covariate effect cannot change sign as a result of model misspecification. If the true effect increases the mean sojourn time, the mean estimated covariate effect will also be positive. An increasing or decreasing hazard function simply increases or decreases the estimated effect. The estimates of variance again reflect the result that $\alpha > 1$ results in an overestimate of the

Figure 4.11: Contour plot of bias in estimate of covariate effect for varied $t$ and $\alpha$. $a = 1$, $m = 10, \lambda = 0.4, \omega = 1.5$




variance and $\alpha < 1$ an underestimate. For realistic sample sizes, the effect of the bias on the coverage of 95% confidence intervals is more moderate than for the equivalent effects on estimates of mean sojourn time itself. To some degree this is because, for equivalent sample sizes, there is much greater uncertainty about the difference between groups than about a particular group mean. Table 4.4 gives the approximate expected coverage of 95% confidence intervals when group 0 has mean sojourn time of 2.5 years, and group 1 a mean sojourn time of 3.75 years, i.e. a 50% increase.

For weaker covariate effects the same pattern of bias exists. However, the decay of confidence interval coverage is slower because there is less certainty about the true effect. The magnitude of the bias decreases as $m$ increases, although as with estimates of mean sojourn time, the influence of the value of $m$ is small.

Table 4.4: Approximate coverage of assumed 95% confidence intervals on the covariate effect on mean sojourn time in a 2-state model for varying $\alpha$, fixed $t = 5$, $m = 5$. $N$ refers to the sample in each group.

| | $N$ | | | |
|---|---|---|---|---|
| $\alpha$ | 100 | 500 | 1000 | 2000 |
| 0.6 | 0.899 | 0.781 | 0.702 | 0.511 |
| 0.8 | 0.934 | 0.917 | 0.900 | 0.852 |
| 0.9 | 0.943 | 0.940 | 0.935 | 0.926 |
| 1.1 | 0.954 | 0.951 | 0.948 | 0.941 |
| 1.2 | 0.957 | 0.946 | 0.933 | 0.907 |
| 1.4 | 0.958 | 0.925 | 0.880 | 0.785 |

### 4.4.1 More than 2 states

As in section 4.3.2, we can extend the analysis of estimated covariate effects to the three state unidirectional case. Now let there be two, equal sized, groups of patients and let the mean sojourn times for the second group be $\omega_1$ and $\omega_2$ times the mean sojourn times for the first group for states 1 and 2 respectively, where again we limit the form of the covariate effect to a factor on the rate parameter.

Mirroring results on mean sojourn times, results about the covariate effect on a misspecified first state continue to be true in the three-state case. The effect is over estimated when $\alpha > 1$ and under estimated if $\alpha < 1$. Moreover, when the sampling is reasonably frequent, the bias decreases as the time of right censoring increases. The magnitude of bias is also less when $m$ is large, i.e. when sampling is frequent.

It is also the case that the bias on the covariate effect for state 2 is very small in magnitude. When $\alpha < 1$, it is over estimated, whilst for $\alpha > 1$ it is under estimated. The magnitude of bias decreases for increasing $m$, but also increases as the time of right censoring, $t$, increases. This is the reverse of the pattern of biases in the mean sojourn times.

We can also consider the case in which the second state is misspecified whilst the first is correctly specified. As before, constraints caused by the effectiveness of numerical quadrature for densities with singularities mean we restrict ourselves to cases where the second state has mean sojourn time that is Gamma with shape parameter $\alpha > 1$. Mirroring

the results for mean sojourn time, there is virtually no bias in the covariate effect on transition intensities on state 1. The approximate coverage of 95% confidence intervals ranges from 94.1 - 95.5 % for the examples considered. The small bias that exists means the covariate effect is under estimated when the hazard in state 2 is increasing. The covariate effect on state 2 sojourn time is over estimated. The magnitude of bias decreases with increasing $m$ and $t$.

### 4.4.2 Implications

The pattern of bias from model misspecification on estimates of covariate effects corresponds quite closely with those for mean sojourn time. In particular, when there are multiple states, misspecification of one particular state does not seem to have an extensive impact on covariate effects on other states.

There is some indication that covariate effects on sojourn times are more robust to model misspecification than estimates of sojourn times themselves. It is certainly true, in this 2 state case, that the sign of the expected covariate effect (i.e. whether the parameter increases or decrease mean sojourn time) will match the true covariate effect in the presence of misspecification. However much of the apparent robustness may primarily be a reflection of the greater degree of uncertainty about the covariate effect for a particular sample size, meaning a greater sample size is needed for the bias in the estimate to dominate the variability.

## 4.5 Patient heterogeneity

Another source of model misspecification occurs when there is patient heterogeneity. Each subject may have their own unique set of transition intensities. This may be due to non-inclusion of covariates or through an unmeasurable individual frailty. In this section we shall assume the quantities of interest are the mean transition intensities among the population. This is to ensure the quantity estimated has an interpretation under both the assumed and true models.

### 4.5.1 Patient heterogeneity with interval censored observation

Consider again the simple two-state model illustrated in figure 4.4, but let $z_1, ..., z_n$ be a random sample from some distribution $Z$, and let the transition intensity for subject $j$ be some function $\lambda(.)$ of $z_j$. A common example is to allow $Z \sim \mathrm{N}(0, \sigma^2)$ and then let $\lambda(Z) = \lambda_0 \exp(Z)$, so that there is a multiplicative log-normal random effect. However, this distribution makes direct comparison of estimates difficult because $\mathbb{E}(\exp(Z)) = \exp(\frac{\sigma^2}{2}) \neq 1$ for $\sigma^2 > 0$. Moreover, determining the probability of each observable history $p_i$, is algebraically difficult in this case. Instead, we will use the inverse-Gaussian distribution with mean 1, as used in the tracking model of Satten [117], discussed in section 2.5, as this allows analytic solutions for $p_i$.

Using the regular sampling scheme as depicted in figure 4.5, the estimator under the assumed homogeneous Markov model is the same as in section 4.3.1. The mean and variance of $\hat{\lambda}$ depend on the probability, $p_i$, of each of the $m+1$ distinct observable histories. Conditional on a particular fixed value of $\lambda$, the probability of the $0 \rightarrow 1$ transition occurring in the $i$th time interval is

$$p_i = \begin{cases} \exp(-\lambda\frac{t(i-1)}{m}) - \exp(-\lambda\frac{ti}{m}) & i = 1, \ldots, m \\ 1 - \sum_{k=1}^{m} p_k & i = m+1 \end{cases}.$$

The probability, taking into account the random effect, is the expectation of this

$$p_i = \int \left( \exp(-\lambda(z)\frac{t(i-1)}{m}) - \exp(-\lambda(u)\frac{ti}{m}) \right) \phi(z)dz, \tag{4.9}$$

where $\phi(z)$ is the pdf for $Z$. Equation 4.5 then applies, with $\nu$ and $\rho$ calculated from the new $p_i$.

For an inverse-Gaussian multiplicative random effect with precision parameter $\phi$ and mean 1, as used in the tracking model discussed in section 2.5, the integral is tractable. The inverse-Gaussian distribution gives a regular pattern of bias, with a consistent underestimation of the true mean transition intensity (figure 4.12). When $m = 1$, equation 4.5 reduces to

$$\hat{\lambda} \xrightarrow{p} \frac{1}{t}\log(p_2).$$

By applying the Laplace transform from equation 2.9,

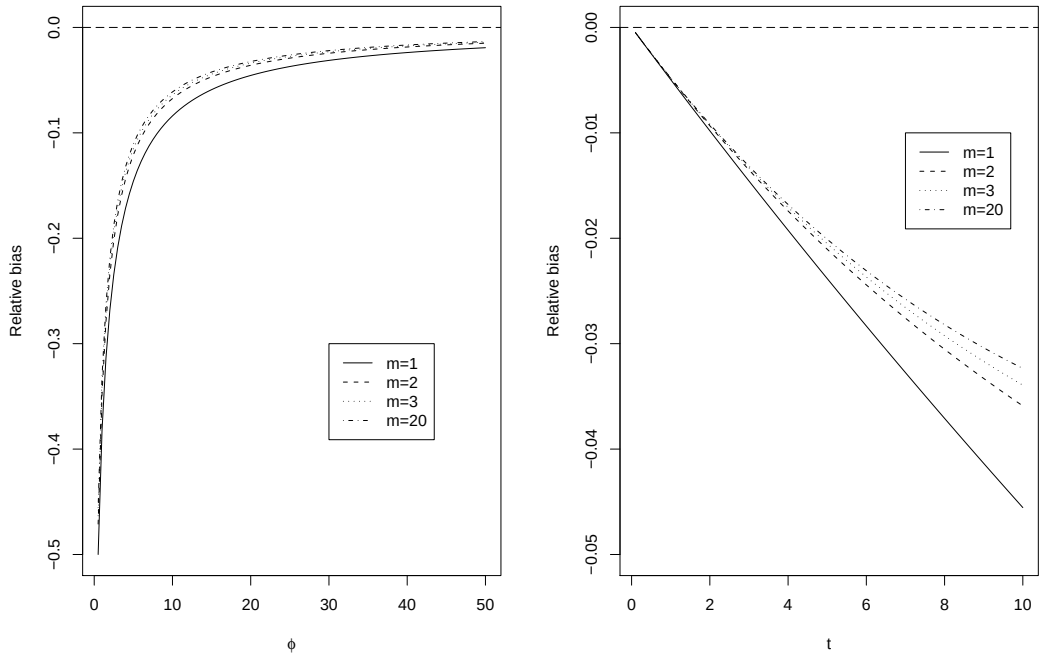
$$p_2 = \exp(\phi - (s^2 + 2\lambda\phi t)^{\frac{1}{2}})$$

and hence

$$\hat{\lambda} \xrightarrow{p} \frac{1}{t}\left((\phi^2 + 2\lambda\phi t)^{\frac{1}{2}} - \phi\right).$$

This is a monotonically increasing function in $\phi$, tending to $\lambda$ as $\phi \to \infty$. Hence, the mean transition intensity is always under estimated for this sampling scheme. For $m > 1$, the mean transition intensity is also under estimated, but the magnitude of bias decreases. The magnitude of bias increases as $t$ increases.

Figure 4.12: Asymptotic estimate of mean transition intensity with a inverse-Gaussian frailty when homogeneity is assumed for varied $t$ and $m$. Varied $\phi$ and $m$ (left) and fixed $\phi = 20$ (right) for $\lambda_0 = 0.2$.




### 4.5.2 Patient heterogeneity in a more complicated model

This section considers the effect of assuming a subject and time homogeneous Markov model for data in a three-state disease (illness-death) process, when the true process is that of the tracking model of Satten [117], discussed in section 2.5, where there exists a

frailty distribution $G(z;\phi)$, such that patient $i$'s transition intensity matrix is given by $z_i Q_0$, where $z_i$, for $i = 1, \ldots, n$ are independent samples from $G(z;\phi)$. As in section 4.5.1, we will take the frailty distribution to be inverse-Gaussian. If we again assume the sampling regime of equally spaced observations, we can use the methods of section 4.2.1, to numerically calculate the asymptotic bias in the estimates and approximate coverage of 95% confidence intervals for the mean transition intensities. Let $G(\ \phi)$ be inverse Gaussian, with mean 1 and precision parameter $\phi$, so that

$$g(z,\phi) = \left(\frac{\phi}{2\pi z^3}\right)^{0.5} \exp\left(\phi - \frac{\phi(z+z^{-1})}{2}\right).$$

This means that in the limit as $\phi \to \infty$, the subjects all have the same intensities and as $\phi$ decreases, the variability between intensities increases.

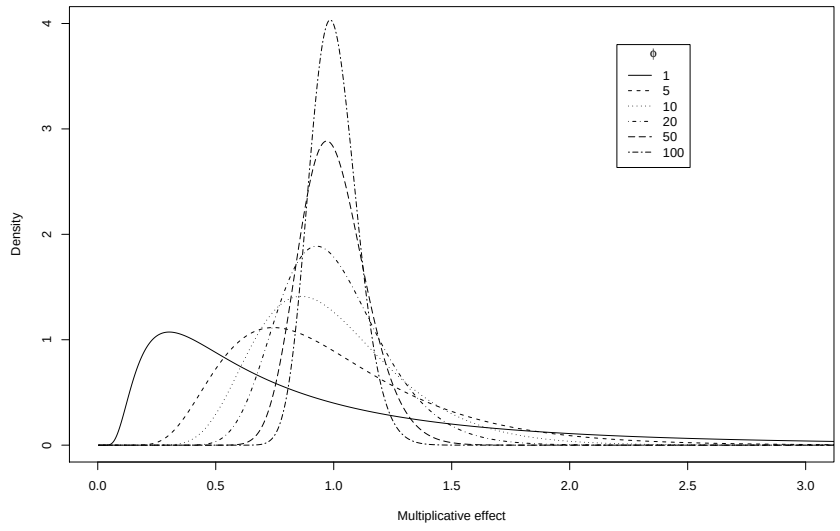
We consider the effect of different values of $\phi$ on parameter estimates of the mean transition intensities. Suppose the underlying process is such that the true mean transition intensities are $q_{12} = 0.4, q_{13} = 0.05, q_{23} = 0.1$. These parameters are chosen as they represent a realistic illness-death model scenario.

We shall consider 6 different values of $\phi$; 1, 5, 10, 20, 50, 100. The resulting densities for the frailty factors, $z_i$, are shown in figure 4.13. $\phi = 1$ or $\phi = 5$ are seen to involve quite extreme heterogeneity, being heavily positively skewed. $\phi = 10$ or $\phi = 20$ represent realistic levels of heterogeneity, with around 5% of subjects having rates 40% or more below the mean, and 5% of subjects around 50% above the mean. $\phi = 50$ or $\phi = 100$ may also be realistic but, representing weaker heterogeneity, are less likely to be detectable in realistic sample sizes.

As before we shall also vary the time of censoring, $t$, and the number of observations, $m$.

The pattern of bias is more complicated in this three-state disease model (figure 4.14). In terms of the mean rate of onset ($\mathbb{E}(q_{12})$), the pattern of bias is the same as the 2-state case: the intensity is underestimated, but to a fairly modest extent for realistic values of $\phi$. It is also true that the transition intensity $q_{13}$ is always slightly underestimated. However, the bias in the transition rate between 2 and 3, depends on the level of censoring. For small $t$, there is an over estimate, whilst for large $t$, it is underestimated. For small $t$, the estimate of $q_{23}$, is dominated by the subjects, typically with a high $z_i$, who reach state 3 quickly. However, for large $t$, the estimator is dominated by the subjects with low $z_i$, who, having reached state 2, then take a long time to reach state 3. As might be expected the magnitude of the bias depends directly on the value of $\phi$, with greater bias the smaller the value of $\phi$ (i.e. the greater the degree of model misspecification). Leaving aside considerations of

Figure 4.13: Densities of inverse-Gaussian frailty factor distributions for different values of $\phi$.



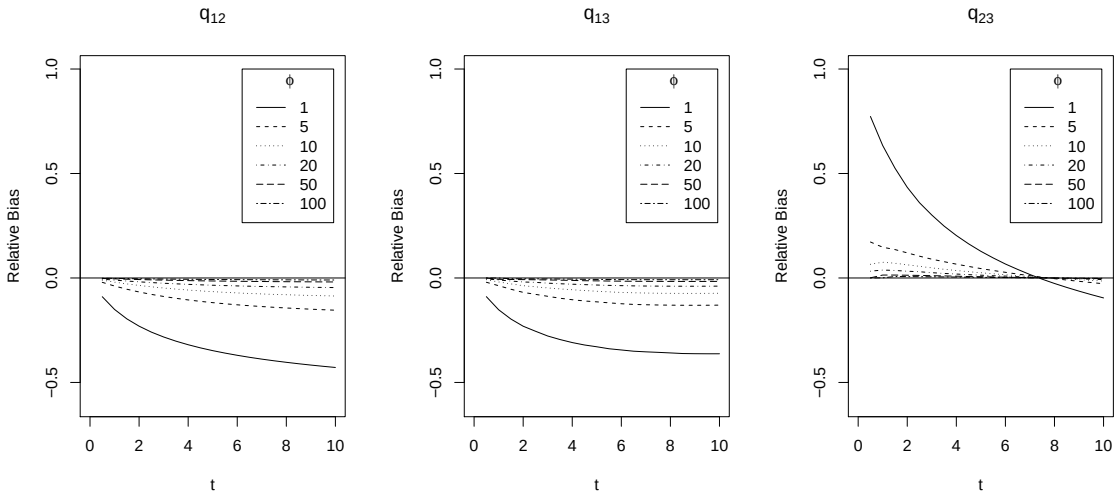

heterogeneity, in an illness-death model there is a complicated interaction between $t$ and $m$ in terms of the magnitude of bias (figure 4.15). The number of observations, $m$, has little impact on the degree of bias. The exception to this is for large $t$, where if $m$ is small, most subjects reach the absorbing state in the first time interval.

The bias in the variance of the estimators is closely linked to the bias in the estimate, being mostly underestimated except for the case of $q_{23}$, where for small $t$, the variance of the estimator is overestimated. However, the bias in the variance is a secondary factor to the coverage of confidence intervals. Even in the most extreme case considered ($t = 10, m = 10, \phi = 1$), $\widehat{Var}(\hat{q}_{12})$ is 76% of the true value. The error in the variance for $q_{13}$ and $q_{23}$ is negligible, never being more than 4% out. Table 4.5 gives the approximate coverage of 95% confidence intervals based on the biased estimators, for the case $m = 10$, $t = 5$, for varying $\phi$ and number of subjects $N$. Coverage of confidence intervals decays faster for $q_{12}$ than the other intensities. Mild heterogeneity causes little problems, for instance when $\phi = 50$, the confidence interval for $q_{12}$ for a sample size of 2000 subjects still has 91% coverage. Severe heterogeneity, however, leads to substantial bias, severely affecting coverage.

Table 4.5: Approximate coverage of assumed 95% confidence intervals for varying $\phi$. $t = 5$, $m = 10$.

| | | $N$ | | | |
|---|---|---|---|---|---|
| Intensity | $\phi$ | 100 | 500 | 1000 | 2000 |
| $q_{12}$ | 1 | 0.090 | 0.000 | 0.000 | 0.000 |
| | 5 | 0.802 | 0.336 | 0.084 | 0.003 |
| | 10 | 0.904 | 0.737 | 0.542 | 0.260 |
| | 20 | 0.936 | 0.891 | 0.833 | 0.718 |
| | 50 | 0.947 | 0.940 | 0.930 | 0.911 |
| | 100 | 0.949 | 0.947 | 0.945 | 0.940 |
| $q_{13}$ | 1 | 0.818 | 0.361 | 0.095 | 0.004 |
| | 5 | 0.936 | 0.882 | 0.813 | 0.676 |
| | 10 | 0.946 | 0.930 | 0.910 | 0.869 |
| | 20 | 0.949 | 0.944 | 0.939 | 0.928 |
| | 50 | 0.950 | 0.949 | 0.948 | 0.946 |
| | 100 | 0.950 | 0.950 | 0.949 | 0.949 |
| $q_{23}$ | 1 | 0.920 | 0.793 | 0.637 | 0.377 |
| | 5 | 0.946 | 0.928 | 0.905 | 0.859 |
| | 10 | 0.949 | 0.943 | 0.936 | 0.922 |
| | 20 | 0.950 | 0.948 | 0.946 | 0.942 |
| | 50 | 0.950 | 0.950 | 0.949 | 0.948 |
| | 100 | 0.950 | 0.950 | 0.950 | 0.950 |

Figure 4.14: Plot of asymptotic relative bias of mean transition intensity estimates for varying $t$ and $\phi$ and fixed $m = 10$.



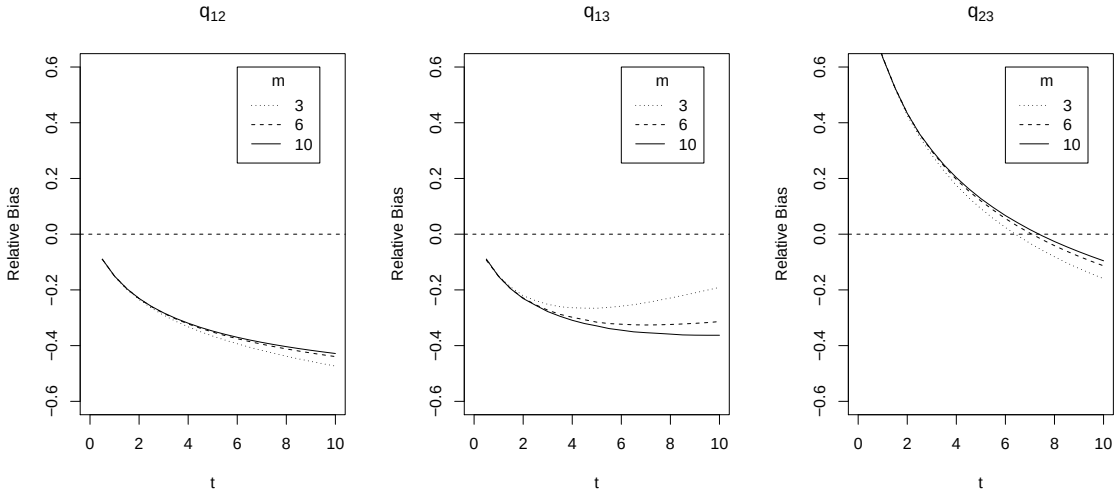

## 4.6 Misspecification in HMM

In section 2.6.3, it was noted that there is an inherent assumption in HMMs, that

$$O_1|X_1, \ldots, O_n|X_n$$

are independent. A simple test of this assumption was presented and it was noted that for the misclassification HMM for the BOS data, there was clear evidence that this assumption was incorrect. Any dataset that consists of a continuous or multi-valued measurement that is categorised into a few states could be subject to this type of model misspecification. Satten *et al* [115] recognised this potential in the case of modelling CD4 counts in HIV infected patients, but argued that provided short term fluctuations, which would be the driver of extra correlation between observed states, occur on a time scale shorter than the frequency of observations, there would not be a problem. In this section we investigate the possible effect of having more persistent time dependencies in HMMs.

Figure 4.15: Plot of asymptotic relative bias of mean transition intensity estimates for varying $t$ and $m$ and fixed $\phi = 1$.




### 4.6.1 Methods

To investigate the effect of time-dependent misclassification we firstly need an appropriate true model for such a situation. In section 2.6.3, the proposed test for detecting dependence in the misclassification was a likelihood ratio test based on an alternative model where the probability of misclassification at observation $i$ depended on the observed state at observation $i-1$. It was noted that this is an unrealistic model if the times between observations are irregular. For evenly spaced observations however, such a model is a reasonable first approximation. An advantage of using this model is that only a small change in the forward algorithm (see section 2.4.1), for calculating $\mathbb{P}(O_1, \dots, O_n)$ is needed. Specifically in the recursion

$$\alpha_k(j) = \mathbb{P}(O_1, \dots, O_k, X_k = j) = \sum_{i=1}^{R} \alpha_{k-1}(i) e_{j,O_k} p_{ij}(t_k - t_{k-1})$$

becomes

$$\alpha_k(j) = \mathbb{P}(O_1, \dots, O_k, X_k = j) = \sum_{i=1}^{R} \alpha_{k-1}(i) e^*_{j,O_k,O_{k-1}} p_{ij}(t_k - t_{k-1})$$

where $e^*_{j,O_k,O_{k-1}}$ denotes the probability of being observed in state $O_k$, given the true state is $j$ and the previous observed state was $O_{k-1}$. We can therefore use this algorithm to

get the probability, $\mathbf{p}^*$ of each possible observed response for a particular set of parameters. Taking these probabilities as fixed we can then optimise the expected likelihood function under the misspecified standard misclassification HMM. This gives the asymptotic estimates of the parameters under the misspecified model. By computing, at these parameter values, the Fisher information contribution and the squared score for each possible response, and weighting by $\mathbf{p}^*$, we can calculate the asymptotic assumed and true covariance matrices.

Ideally we would like to consider data with long sequences of observations for each individual. However, the method used requires enumeration of every possible sequence of observed states. For data subject to misclassification, the number of possible observed states goes up exponentially. Hence even for a three-state disease model, if an individual is observed up to $m$ times, the number of possible sequences is $2^{m+1} - 1$. We limit this study to a small number of observations, say $m = 10$ resulting in 2047 possible sequences. We shall again take the three-state disease model example used in section 4.5.2. Let the underlying process be homogeneous Markov with parameters $q_{12} = 0.40$, $q_{13} = 0.05$, $q_{23} = 0.10$. Three possible scenarios of dependent misclassification are considered. In all cases we shall assume that the observed state is more likely to follow the previous observed state than in a standard HMM. In each case, when the previous observed state was 1, the probability of being misclassified to state 2 from true state 1 is 0.02 and when the previous observed state was 2, the probability of being misclassified to state 1 from true state 2 is 0.05. We shall let the time between observations be 0.5 years, meaning around 48% of subjects reach the absorbing state by 5 years. The three scenarios attempt to encapsulate mild, moderate and strong dependence in the misclassification process. The misclassification probabilities are summarised in table 4.6. Note that although the strong scenario is quite extreme, the estimates for the misclassification probabilities found for the BOS data when applying the independence test in section 2.6.3 are similar.

### 4.6.2 Results

Table 4.7 gives the mean estimated parameters and the approximate coverage of 95% confidence intervals for the three scenarios. As well as the parameters of the Markov process, the bias on the log-hazard ratio of mortality between state 2 and state 1 is considered, and also the overall expected lifetime (expected time from initiation to absorption). In all three cases $q_{12}$ is under estimated whilst $q_{13}$ is overestimated. $q_{23}$ is almost entirely robust

Table 4.6: Scenarios of dependent misclassification

| Scenario | $\mathbb{P}(O_i = 2 \mid X_i = 1, O_{i-1} = 2)$ | $\mathbb{P}(O_i = 2 \mid X_i = 1, O_{i-1} = 1)$ |
|---|---|---|
| Mild | 0.05 | 0.02 |
| Moderate | 0.10 | 0.02 |
| Strong | 0.30 | 0.02 |
| Scenario | $\mathbb{P}(O_i = 1 \mid X_i = 2, O_{i-1} = 1)$ | $\mathbb{P}(O_i = 1 \mid X_i = 2, O_{i-1} = 2)$ |
| Mild | 0.10 | 0.05 |
| Moderate | 0.25 | 0.05 |
| Strong | 0.50 | 0.05 |

to the model misspecification. Similarly, the expected lifetime is robust, except that the variance of the estimator is overestimated leading to more than 95% coverage of confidence intervals. The hazard ratio is under estimated. Results for the 'mild' scenario suggest that small departures from independent misclassification are not problematic. However, when the dependency becomes more marked estimates of transition intensities for individual states can become heavily biased. Estimates of the overall pattern of mortality, both in terms of expected lifetime and the survival function remain virtually unbiased. It is reassuring that the dependency in misclassification doesn't impinge on the survival part of the model. The effect of dependent misclassification is to invalidate inferences made about the intermediate states in a misclassification HMM.

## 4.7 Conclusion

This chapter has attempted to investigate some of the effects of model misspecification in general cases, without recourse to simulation. The complexity of the models means that results cannot be found in generality. Instead results have only been provided in some relatively simple special cases. However, they do give some insight into the resulting size and the direction of biases and also the impact on coverage of assumed 95% confidence intervals, when there has been model misspecification.

In most of the examples considered, having more frequent observations, within a fixed follow-up time, leads to smaller expected bias. An exception to this is in the estimation of mean sojourn time when the true process is Gamma. Here the bias increased as the

Table 4.7: Bias in estimates and approximate coverage of 95% confidence intervals when assumption of independent misclassification is false. MLT = mean lifetime

| Mild correlation | | | $N$ | | | |
|---|---|---|---|---|---|---|
| | True | Estimated | 100 | 500 | 1000 | 2000 |
| $q_{12}$ | 0.40 | 0.394 | 0.949 | 0.940 | 0.929 | 0.906 |
| $q_{13}$ | 0.05 | 0.051 | 0.950 | 0.949 | 0.949 | 0.948 |
| $q_{23}$ | 0.10 | 0.100 | 0.950 | 0.950 | 0.950 | 0.950 |
| $\log(\frac{q_{23}}{q_{13}})$ | 0.693 | 0.681 | 0.950 | 0.950 | 0.949 | 0.949 |
| MLT | 11.111 | 11.112 | 0.959 | 0.965 | 0.969 | 0.971 |
| Moderate correlation | | | $N$ | | | |
| | True | Estimated | 100 | 500 | 1000 | 2000 |
| $q_{12}$ | 0.40 | 0.370 | 0.900 | 0.663 | 0.406 | 0.121 |
| $q_{13}$ | 0.05 | 0.053 | 0.948 | 0.939 | 0.927 | 0.903 |
| $q_{23}$ | 0.10 | 0.100 | 0.950 | 0.950 | 0.950 | 0.950 |
| $\log(\frac{q_{23}}{q_{13}})$ | 0.693 | 0.646 | 0.949 | 0.943 | 0.937 | 0.923 |
| MLT | 11.111 | 11.110 | 0.958 | 0.965 | 0.969 | 0.972 |
| Strong correlation | | | $N$ | | | |
| | True | Estimated | 100 | 500 | 1000 | 2000 |
| $q_{12}$ | 0.40 | 0.311 | 0.351 | 0.000 | 0.000 | 0.000 |
| $q_{13}$ | 0.05 | 0.057 | 0.931 | 0.850 | 0.745 | 0.548 |
| $q_{23}$ | 0.10 | 0.101 | 0.950 | 0.950 | 0.949 | 0.949 |
| $\log(\frac{q_{23}}{q_{13}})$ | 0.693 | 0.566 | 0.940 | 0.896 | 0.841 | 0.727 |
| MLT | 11.111 | 11.115 | 0.958 | 0.964 | 0.968 | 0.971 |

number of observations per patient increased. The Gamma distribution is most distinct from an exponential soon after entry into the state. This potentially has some implications for the design of studies. Observing patients soon after the initiation of the process could certainly allow more power to detect a non-exponential first state. The distance between the Gamma and exponential densities is greatest close to $t = 0$. It is less clear how to improve the power for other states.

Misspecification of the form of the baseline intensities, was shown to have a significant impact on estimates of covariate effects. Misspecification of the state misclassification process in HMMs, was shown to have the potential to lead to bias in estimates of the intensities of the underlying Markov model. It is therefore important to test that assumptions of independent misclassification are correct. The test developed in section 2.6.3 is available for this.

The diagnostic tests of chapter 2 and the general goodness-of-fit test developed in chapter 3 will often be useful in determining when the fit is poor and therefore when inferences are likely to be invalid. However, when the misspecification is less pronounced detection will require a test for a specific type of misspecification. It was shown that the effect of ignoring patient heterogeneity was similar to the effect of ignoring a time dependent transition intensity (with decreasing hazards). This mirrors results from section 2.5.3, which showed that a progressive process with patient heterogeneity, is close to a time inhomogeneous process with a decreasing hazard. Whilst random effects models may be useful in some circumstances, time dependent models would seem to be more flexible as they can accommodate an apparent increasing hazard as well. The development of more elaborate time dependent models would therefore seem a greater priority. Chapters 5 and 6 develop methods for fitting more complicated time dependent models. A likelihood ratio test provides a more powerful test of the assumption of time homogeneity or the Markov property.

# Chapter 5

# Methods for fitting time inhomogeneous Markov and semi-Markov models

## 5.1 Introduction

This chapter explores methods for fitting time inhomogeneous Markov models and semi-Markov models. Time inhomogeneous Markov models have transition intensities which depend only on the current state and current time (since initiation of the process). Semi-Markov models have transition intensities which depend on the current state and the time since entry into the state. The motivation for fitting these types of models is twofold. Firstly, a likelihood ratio test based upon a comparison between these alternative models and the time homogeneous model provides a specific test of time dependence that is more powerful at detecting time dependence than the general goodness-of-fit test of chapter 3. Secondly, if the time dependent model is more appropriate for the data then this model should provide better estimates of quantities of interest such as mean lifetime and covariate effects.

Three main approaches to fitting time inhomogeneous models exist.

The first approach is to allow the transition intensities to be piecewise constant. Since the likelihood is algebraically tractable, piecewise intensities are perhaps the most common method of fitting inhomogeneous Markov models. This chapter will show that the approach

can be extended to fit some types of semi-Markov models.

The second approach allows the transition intensities to have smooth parametric forms, for instance Weibull hazard functions. For Markov models, calculation of the likelihood requires solution of the Kolmogorov forward equations, which for time inhomogeneous models are a set of non-linear ordinary differential equations. Methods of numerically solving these equations are explored.

A third approach involves the use of non-parametric or semi-parametric techniques. However these methods are quite limited in their applicability to panel observed data. These limitations are discussed in section 1.4.3. This chapter concentrates only on fitting parametric models.

In principle, calculation of the likelihood for progressive time inhomogeneous Markov and semi-Markov models could be achieved through numerical integration. However, these integrals may be of multiple dimension, causing evaluation to be prohibitively slow. A Monte-Carlo Expectation Maximisation (MCEM) algorithm is developed which provides a more efficient way of optimising the likelihood for such models.

## 5.2 Piecewise constant transition intensities

Models involving piecewise constant transition intensities between states allow the assumption of time homogeneity to be relaxed, while at the same time retaining closed form algebraic expressions for the transition probabilities.

In the piecewise constant framework, a series of times $b_1, \ldots, b_M$ are chosen, which define the intervals of constant hazard. The transition intensity matrix is then defined to be:

$$Q(t) = \begin{cases} Q_0 & t < b_1 \\ Q_i & b_i \leq t < b_{i+1}, i = 1, \ldots, M-1 \\ Q_M & t \geq b_M \end{cases}$$

where $t$ is the time since the initiation of the process.

Using results for time homogeneous Markov chains, the transition probabilities between two times $t_a$ and $t_b$ where $b_k < t_a < t_b < b_{k+1}$, are given by

$$P(t_a, t_b) = \exp\left(Q_k(t_b - t_a)\right). \tag{5.1}$$

When $t_a$ and $t_b$ are in different hazard intervals, such that $b_i < t_a < b_{i+1}$ and $b_j < t_b < b_{j+1}$, $j > i$, by the Chapman-Kolmogorov equation, we can write

$$P(t_a, t_b) = P(t_a, b_{i+1})P(b_{i+1}, b_{i+2}) \dots P(b_{j-1}, b_j)P(b_j, t_b).$$

Each of the terms of this product give a transition probability matrix within a time interval of constant hazard and so can be computed using the matrix exponential as in equation (5.1). Computation of the likelihood for a time inhomogeneous Markov model with piecewise constant transition intensities is therefore only slightly more computationally demanding than a time homogeneous Markov model.

### 5.2.1 Piecewise constant Markov model for CAV data

To illustrate the method we fit a model with piecewise constant hazard intensities to the CAV data without misclassification. This is a four state disease model as shown in figure 2.1. Boundary points at times 3 and 6 years after transplant are chosen. This is essentially an arbitrary decision, ensuring roughly equal numbers of observations in each of the three regions of constant intensities. The intensities in period 1 (0-3 years) are given by $q_{rs}^{(1)}$ and for period $m = 2, 3$, the transition intensities are given by

$$q_{rs}^{(m)} = q_{rs}^{(1)} \exp\left(\tau_{rs}^{(m)}\right)$$

where $\tau_{rs}^{(m)}$ represents the effect of time in the $m$th period for the $r \to s$ transition intensity.

Initially a model allowing different effects for each time interval and for each transition intensity was fitted. However, many of the effects in different time periods were not significantly different from baseline according to a likelihood ratio test. A second model, allowing a non-zero value only to those time effects which were significant in the first model, is therefore applied. This model has $-2 \times \text{LL} = 3531.595$ representing an improvement of 21.4 from 3 additional parameters compared with the time homogeneous model. The parameter estimates are shown in table 5.1.

The main effect is a significant increase in onset ($1 \to 2$) transition intensity with time. It increases 43% after 3 years and a further 43% after 6 years. The rate of progression between state 2 and state 3 is estimated to decline after 6 years.

Note that the baseline values, $q_{12}^{(1)}$ and $q_{14}^{(1)}$ are not greatly altered by these time effects. This reflects the fact that the majority of observations informing the state 1 transition intensities occur within 3 years of transplantation.

Table 5.1: Comparison of parameter estimates and 95% confidence intervals for a time homogeneous Markov model and a time inhomogeneous Markov model with piecewise-constant intensities for the CAV data without misclassification

| | homogeneous model | | inhomogeneous model | |
|---|---|---|---|---|
| Parameter | Estimate | 95% CI | Estimate | 95% CI |
| $q_{12}^{(1)}$ | 0.094 | (0.082,0.107) | 0.078 | (0.066,0.092) |
| $q_{14}^{(1)}$ | 0.023 | (0.017,0.030) | 0.022 | (0.017,0.029) |
| $q_{23}^{(1)}$ | 0.200 | (0.162,0.246) | 0.230 | (0.180,0.293) |
| $q_{24}^{(1)}$ | 0.040 | (0.022,0.073) | 0.043 | (0.025,0.075) |
| $q_{34}^{(1)}$ | 0.146 | (0.116,0.184) | 0.145 | (0.115,0.183) |
| $\tau_{12}^{(2)}$ | | | 0.358 | (0.029,0.683) |
| $\tau_{12}^{(3)}$ | | | 0.715 | (0.385,1.046) |
| $\tau_{23}^{(3)}$ | | | -0.418 | (-0.851,0.013) |
| $-2\times$ LL | 3553.0 | | 3531.6 | |

**Hidden Markov model**

A time inhomogeneous hidden Markov model can be fitted to the CAV data with misclassified states. The transition probabilities for the underlying Markov process can be obtained in the same way as above. The likelihood for the observed states can then be obtained by applying the methods of section 1.2.4. As before, we choose transition intensities with three regions, with boundary points at 3 and 6 years. A preliminary model allowing different effects for each region and all transition intensities found that many were not significant. The same time effects as for the analogous model without state misclassification were retained. This second model represented a significant improvement in log-likelihood: $-2 \times \text{LL} = 3913.6$, which is an improvement of 19.7 from 3 additional parameters compared to the time homogeneous hidden Markov model. Table 5.2 gives the parameter estimates. The same notation for the misclassification probabilities, $e_{rs}$ is used as in chapter 2. The intensities for this model are given by

$$q_{rs}^{(m)} = q_{rs}^{(1)} \exp\left(\beta_{rs}^{(\text{IHD})} \times \text{IHD} + \beta_{rs}^{(\text{dage})} \times \text{dage} + \tau_{rs}^{(m)}\right).$$

The time effects on the transition intensities are similar to those for the Markov model. The time effects on the CAV onset rates are slightly higher in the HMM. The estimated effect

Table 5.2: Comparison of parameter estimates and 95% confidence intervals for a time homogeneous hidden Markov model and a time inhomogeneous hidden Markov model with piecewise-constant intensities for the CAV data with misclassification

| | homogeneous model | | inhomogeneous model | |
|---|---|---|---|---|
| Parameter | Estimate | 95% CI | Estimate | 95% CI |
| $q_{12}^{(1)}$ | 0.033 | (0.021,0.050) | 0.023 | (0.014, 0.037) |
| $q_{14}^{(1)}$ | 0.021 | (0.015,0.029) | 0.020 | (0.014, 0.028) |
| $q_{23}^{(1)}$ | 0.190 | (0.143,0.252) | 0.224 | (0.162, 0.310) |
| $q_{24}^{(1)}$ | 0.053 | (0.029,0.099) | 0.056 | (0.032, 0.099) |
| $q_{34}^{(1)}$ | 0.155 | (0.120,0.201) | 0.153 | (0.118, 0.198) |
| $\tau_{12}^{(2)}$ | | | 0.438 | (0.044, 0.832) |
| $\tau_{12}^{(3)}$ | | | 0.849 | (0.468, 1.229) |
| $\tau_{23}^{(3)}$ | | | -0.481 | (-1.073, 0.111) |
| $e_{12}$ | 0.025 | (0.015,0.042) | 0.026 | (0.007,0.043) |
| $e_{21}$ | 0.186 | (0.123,0.272) | 0.169 | (0.099,0.274) |
| $e_{23}$ | 0.065 | (0.038,0.108) | 0.068 | (0.040,0.114) |
| $e_{32}$ | 0.102 | (0.051,0.194) | 0.103 | (0.051,0.196) |
| $\beta_{12}^{(\mathrm{IHD})}$ | 0.520 | (0.234,0.807) | 0.538 | (0.251,0.825) |
| $\beta_{12}^{(\mathrm{dage})}$ | 0.025 | (0.013,0.037) | 0.029 | (0.016,0.041) |
| $-2\times$ LL | 3933.6 | | 3913.6 | |

of IHD and donor age on onset rates is largely unaffected between the time homogeneous and inhomogeneous models. This is in agreement with the findings of section 4.4 where it was found that the effect of misspecification of the baseline model, had less impact on covariate effects than on estimates of mean sojourn time. The result of an increasing hazard for state 1 is consistent with the result of the goodness-of-fit test in chapter 3 where the contingency table of observed versus expected counts showed the highest deviances from an excess of observed $1 \rightarrow 2$ and $1 \rightarrow 3$ transitions in the shortest time intervals.

The estimates of the misclassification probabilities remain similar between models. In particular $e_{12}$ remains very similar.

### 5.2.2 Piecewise constant intensities for semi-Markov models

Semi-Markov models, which were introduced in section 1.2.3, have transition intensities which depend on the time since entry into the current state. In some situations this can be a more appropriate assumption than to assume the time dependency is with respect to calendar time or the initiation of the process.

Computation of the likelihood is difficult for semi-Markov models, particularly when the model is bi-directional. When the model is progressive, the likelihood for a subject can be computed by integrating over the space of possible sojourn times and summing over the possible state paths, given the observed data. Conditional on a particular path, which, without loss of generality, we can label $1, 2, \ldots, m$, the likelihood is given by an integral of the form

$$\int_{l_1}^{u_1} \cdots \int_{l_{m-1}}^{u_{m-1}} f_{1,2}(t_1) f_{2,3}(t_2 - t_1) \ldots f_{m-1,m} \left( t_{m-1} - \sum_{i=1}^{m-2} t_i \right) dt_{m-1} \ldots dt_1 \qquad (5.2)$$

where $l_r$ and $u_r$ denote the lower and upper times the subject could have left state $r$ given the observed data and $f_{r,r+1}(t)$ is the likelihood contribution of leaving state $r$ for state $r+1$ after a period of time $t$ in state $r$.

For most choices of sojourn distribution, the integral in equation 5.2 will be intractable. But, as with time inhomogeneous models, piecewise constant transition intensities allow analytic solutions to equation 5.2. However, whilst going from a homogeneous to inhomogeneous Markov model using piecewise intensities is straightforward, going from homogeneous Markov to semi-Markov models is more complicated. In the following section the necessary calculations for the case of a progressive disease model where the intensities have one discontinuity point are outlined for illustration.

We define the transition intensities for each transition to be

$$q_{rs}(t) = \begin{cases} q_{1rs} & \text{if } 0 < t \leq b_r, \\ q_{2rs} & \text{if } t > b_r. \end{cases} \tag{5.3}$$

where $t$ now relates to the time since entry into the state and the boundary, $b_r$, is the same for all hazards affecting state $r$. The hazards affecting state $r$ are therefore constant between time 0 and $b_r$ since entry into the state, and from $b_r$ to $\infty$.

This gives the density of the jump time between state $r$ and $r+1$, where $r+1 \neq R$, as

$$f_{r,r+1}(t) = \begin{cases} q_{1,r,r+1} \exp(-q_{1,r,r+1}t - q_{1,r,R}t) & \text{if } 0 < t \leq b_r, \\ q_{2,r,r+1} \exp(-q_{1,r,r+1}b_r - q_{2,r,r+1}(t-b_r) - q_{1,r,R}b_r - q_{2,r,R}(t-b_r)) & \text{if } t > b_r. \end{cases} \tag{5.4}$$

Similar expressions can be found for jump times to the death state $R$.

The probability of remaining in state $r$ for time $t$ is

$$F_r(t) = \begin{cases} \exp(-q_{1,r,r+1}t - q_{1,r,R}t) & \text{if } 0 < t \leq b_r, \\ \exp(-q_{1,r,r+1}b_r - q_{2,r,r+1}(t-b_r) - q_{1,r,R}b_r - q_{2,r,R}(t-b_r)) & \text{if } t > b_r. \end{cases} \tag{5.5}$$

As in equation 5.2, computation of the likelihood involves integrating over the space of possible sojourn times given the observed data. For instance the likelihood for an individual observed in state 1 at time $a$, in state 2 at time $b$, and then observed in state 3 at time $c$ and subsequently censored in state 3 at time $T$, is an integral of the form

$$I = \int_a^b \int_{max(b-t,0)}^{c-t} f_{12}(t) f_{23}(u) F_3(T-t-u) du dt.$$

When $f_{rs}$ and $F_r$ are defined as above, the integral can be evaluated analytically by subdividing the integral region into boundaries for which the hazard is constant. Hence we take

$$I = \sum_{i,j,k} \int \int_{(t,u) \in V_{i,j,k}} f_{12}(t) f_{23}(u) F_3(T-t-u) du dt$$

where

$$V_{i,j,k} = \{(t,u) : q_{rs}(t) = q_{irs}, q_{rs}(u) = q_{jrs}, q_{rs}(T-t-u) = q_{krs}\}.$$

In this case where there are two regions for each hazard function, there are a total of 8 possible regions for the integrand. These regions can be shown to be expressible as the sum

of parallelogram, trapezium or rectangular regions of the $(t, u)$ plane. These integrands can be evaluated analytically.

The likelihood for an individual can be calculated by summarising their observations into whether a jump occurred and the minimum and maximum times for the jump, for each of the states. As with piecewise constant intensities for Markov models, the choice of the boundary points can be crucial to the fit of the model. However, this choice generally has to be arbitrary unless it can be informed by knowledge of the biological process being modelled. In the absence of any external knowledge it seems sensible to try to choose a boundary point such that there is roughly equal information in the data about times before and after the boundary point.

### 5.2.3 Illustrative example of necessary calculations

Consider a subject, in a disease model with the same structure as the CAV model (figure 2.1), who is assumed to begin in state 1 at time zero, is observed to be in state 1 at time 1 and is in state 3 at time 3. Suppose the boundary points for each sojourn time are 1.5 years for state 1, 0.5 years for state 2 and 0.5 year for state 3. Therefore the admissible region for the sojourn times is shown in figure 5.1.

The broken lines divide regions for which the transition intensities, at the times of transition, are the same. Hence to evaluate the integral

$$I = \int_1^3 \int_0^{3-t} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt$$

where the functions $f_{12}$, $f_{23}$ and $F_3$ are defined as in equations 5.4 and 5.5, with $b_1 = 1.5$, $b_2 = 0.5$, $b_3 = 0.5$, it is necessary to split into the following:

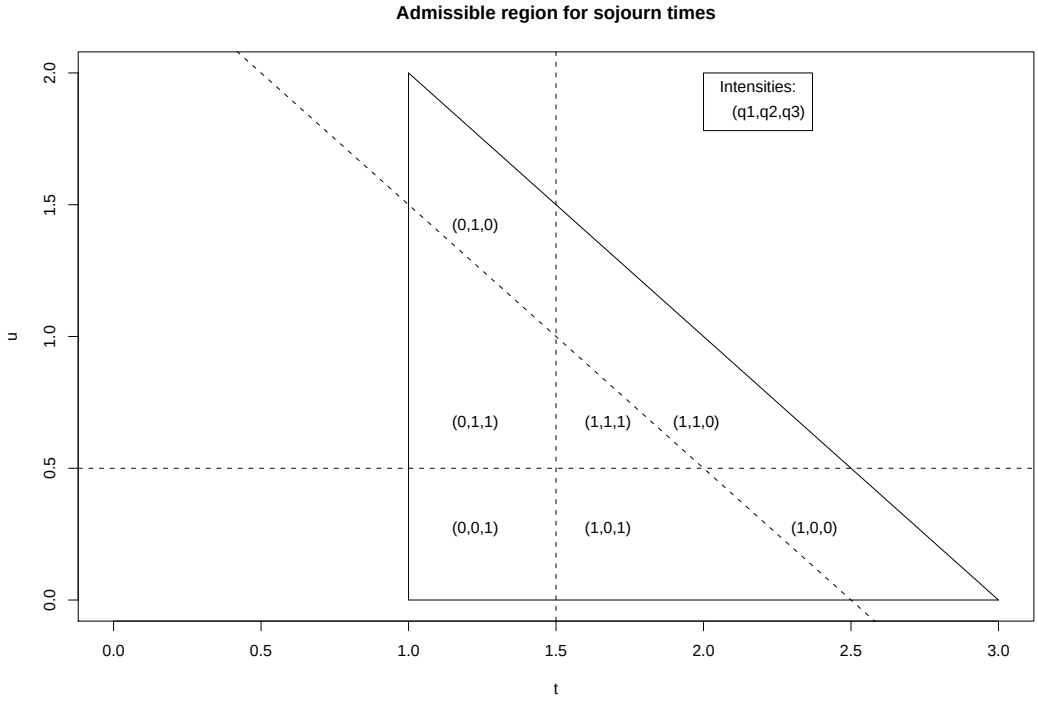
$$I_{(0,0,1)} = \int_1^{1.5} \int_0^{0.5} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt$$

$$I_{(0,1,1)} = \int_1^{1.5} \int_{0.5}^{2.5-t} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt$$

$$I_{(0,1,0)} = \int_1^{1.5} \int_{2.5-t}^{3-t} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt$$

$$I_{(1,0,1)} = \int_{1.5}^{2} \int_0^{0.5} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt + \int_2^{2.5} \int_0^{2.5-t} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt$$

Figure 5.1: Area of the $(t, u)$ plane to be integrated. $q_j = 0$ if the time is before the change point for intensity $j$ and $q_j = 1$ otherwise, e.g. (0,0,1) represents being before the change point for states 1 and 2 and after the change point for state 3.




$$I_{(1,1,1)} = \int_{1.5}^{2} \int_{0.5}^{2.5-t} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt$$

$$I_{(1,1,0)} = \int_{1.5}^{2} \int_{2.5-t}^{3-t} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt + \int_{2}^{2.5} \int_{0.5}^{3-t} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt$$

$$I_{(1,0,0)} = \int_{2}^{2.5} \int_{2.5-t}^{0.5} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt + \int_{2.5}^{3} \int_{0}^{3-t} f_{12}(t) f_{23}(u) F_3(3-t-u) du dt.$$

$f_{12}(t) f_{23}(u) F_3(3-t-u)$ is of the form $C \exp(-h_1 t - h_2 u - h_3(3-t-u))$ where $C, h_1, h_2, h_3$ are constants in the region of the integrand. Hence each integral can be evaluated analytically.

### 5.2.4 Extension for Misclassification

In the datasets considered in this thesis, the observed states are observed with error. To generalise the HMMs fitted to these data, we need to fit a hidden semi-Markov model

(HSMM). The presence of misclassification presents practical, rather than theoretical, problems. An individual, observed with misclassification in states $o_1, \ldots, o_N$ at a series of times $t_1, \ldots, t_N$ may be known to have been in state 1 before time $a$ and known to have been in at least state 2 by time $b$, and in state 3 by time $c$, and still in state 3 at time $T$, has likelihood given by

$$\int_a^T \int_{max(b-t,0)}^{c-t} f_{12}(t) f_{23}(u) F_3(T-t-u) \prod_{i=1}^{n} E(o_i, t, u) du dt \tag{5.6}$$

where $E(o_i, t, u)$ defines the contribution of misclassification probabilities to the likelihood from observation $o_i$ given sojourn times $t$ and $u$ and is defined as

$$E(o_i, t, u) = \begin{cases} e_{1o_i} & \text{if } t_i \leq t, \\ e_{2o_i} & \text{if } t < t_i \leq t+u, \\ e_{3o_i} & \text{if } t_i > t+u. \end{cases} \tag{5.7}$$

where $e_{rs}$ is the usual misclassification probability of being observed in state $s$ when the true state is $r$. Following the same approach as before, it is necessary to subdivide the integrand into regions where both the transition intensities and the $\prod_{i=1}^{n} E(o_i, t, u)$ term are constant. This occurs when the interval for each of the jump times is between consecutive observations. Hence it is necessary to determine all possible sets of true states at the observation times for each subject. When the misclassification is subject to restrictions, for instance if $e_{13} = e_{31} = 0$, there are fewer possible sets. Nevertheless, for datasets in which each individual is observed a large number of times, the number of possible sets of states and therefore the computation time will be large.

For the purposes of optimisation, it is possible to determine the admissible sets of states, which are independent of the parameter values, as a preliminary step. This information can then be stored and subsequent likelihood evaluations do not require the admissible paths to be recalculated.

### Illustrative example

Suppose a subject in a misclassification model with the same structure as the 4-state CAV model (figure 2.1) is observed 5 times (as in table 5.3).

It is assumed the subject begins in state 1 at time 0 and misclassification is possible only to adjacent states. The maximum time at which the 1→2 transition could have occurred is 4.2 (since misclassification from state 1 to 3 is assumed to be impossible.) The minimum

Table 5.3: Observations for an example patient

| Time | State |
|---|---|
| 1.2 | 1 |
| 2.4 | 1 |
| 3.5 | 2 |
| 4.2 | 3 |
| 5.1 | 4 |

time the transition could have occurred is 0. The subject may never have entered state 3, but if the 2→3 transition took place it must have been between times 2.4 and 5.1. The 15 distinct "paths" for the individual are given in table 5.4. Paths 4, 8, 12 and 15 refer to cases where the subject went straight from state 2 to state 4.

### 5.2.5 Application to CAV data

The methods of this section can be applied directly to the CAV data. A semi-Markov model might be more appropriate for the CAV data if, for instance, mortality given development of mild or severe disease, depended on the time since development of the disease rather than time since transplant. It is necessary to choose the location of the time points. A choice of 6 years for state 1, 4 years for state 2 and 2 years for state 3 seems a reasonable choice to ensure roughly equal data about each region of the sojourn distribution.

#### Data without misclassified states

Firstly, we apply the method to the CAV data without misclassified states and without covariates. A model with two time periods for each state and a different time effect for each state adds 5 additional parameters. The transition intensity for times before $b_r$ is given by $q_{rs}^{(1)}$ and for times after $b_r$ by

$$q_{rs}^{(2)} = q_{rs}^{(1)} \exp(\tau_r).$$

However, there is little evidence of a time effect in state 2. Similarly, the effect of the time change in state 1 for the $1 \rightarrow 2$ transition does not differ significantly from the effect on $1 \rightarrow 4$. Therefore a more parsimonious model with only two additional parameters is

Table 5.4: Possible paths for example patient

| | Interval transition occurred | |
|---|---|---|
| Path | 1→2 | 2→3 |
| 1 | 0-1.2 | 2.4-3.5 |
| 2 | 0-1.2 | 3.5-4.2 |
| 3 | 0-1.2 | 4.2-5.1 |
| 4 | 0-1.2 | NA |
| 5 | 1.2-2.4 | 2.4-3.5 |
| 6 | 1.2-2.4 | 3.5-4.2 |
| 7 | 1.2-2.4 | 4.2-5.1 |
| 8 | 1.2-2.4 | NA |
| 9 | 2.4-3.5 | 2.4-3.5 |
| 10 | 2.4-3.5 | 3.5-4.2 |
| 11 | 2.4-3.5 | 4.2-5.1 |
| 12 | 2.4-3.5 | NA |
| 13 | 3.5-4.2 | 3.5-4.2 |
| 14 | 3.5-4.2 | 4.2-5.1 |
| 15 | 3.5-4.2 | NA |

Table 5.5: Comparison of parameter estimates and 95% confidence intervals for a time homogeneous Markov model and a semi-Markov model with piecewise-constant intensities for the CAV data without misclassification

| | Markov model | | Semi-Markov model | |
|---|---|---|---|---|
| Parameter | Estimate | 95% CI | Estimate | 95% CI |
| $q_{12}^{(1)}$ | 0.094 | (0.082,0.107) | 0.085 | (0.074,0.099) |
| $q_{14}^{(1)}$ | 0.023 | (0.017,0.030) | 0.020 | (0.015,0.027) |
| $\tau_1$ | | | 0.452 | (0.188,0.717) |
| $q_{23}$ | 0.200 | (0.162,0.246) | 0.212 | (0.172,0.263) |
| $q_{24}$ | 0.040 | (0.022,0.073) | 0.024 | (0.008,0.076) |
| $q_{34}^{(1)}$ | 0.146 | (0.116,0.184) | 0.225 | (0.149,0.338) |
| $\tau_3$ | | | -0.618 | (-0.768,-0.468) |
| $-2\times$ LL | 3553.0 | | 3538.0 | |

fitted. The sojourn time in state 2 is assumed to be exponential, and the log-linear effect of having stayed in state 1 for more than 6 years is assumed the same for $q_{12}$ and $q_{14}$. Table 5.5 gives the parameter estimates for this model compared to the time homogeneous Markov model.

There is a clear improvement in the likelihood, giving a likelihood ratio statistic of 15.0 on 2 df (p=0.001). The model estimates that the hazard of onset or death from state 1 increases after 6 years in the state. However, there is also a significant decrease in the hazard of death after a subject has spent two years in state 3. The hazard of death from state 2 is lower under this model although the uncertainty about this parameter is considerably greater than before.

**Data with misclassified states**

The method can also be applied to the full dataset including misclassified states. Using the technique described in section 5.2.4, it is necessary to determine all possible true observed states at the observation times. For the CAV data, this is reasonable because no subject is observed more than 15 times and misclassification is assumed to only be possible to adjacent states. Hence the total number of sets of observed states is 6943 for these data. Evaluating the likelihood for the misclassification model therefore takes around 12 times

longer than for the equivalent model without misclassification. We assume again that state 2 has no time dependencies and that the time dependency for state 1 is the same for each transition intensity. In addition, it is assumed that the covariates IHD and donor age only affect disease onset rate ($1 \rightarrow 2$ transitions), and moreover, the log-linear covariate effect is the same before and after the 6 year time boundary. This creates a model with 13 parameters.

Parameter estimates are shown in table 5.6. As with the piecewise constant hidden Markov model, the baseline estimate $q_{12}^{(1)}$ is not significantly altered by the time effect. This reflects the fact that most of the observations occur within 6 years of transplantation. The covariate effects $\beta_{12}^{(\text{IHD})}$ and $\beta_{12}^{(\text{dage})}$ affect the onset rate such that

$$q_{12}^{(m)} = q_{12}^{(1)} \exp\left(\beta_{12}^{(\text{IHD})} \times \text{IHD} + \beta_{12}^{(\text{dage})} \times \text{dage} + \tau_1 \times \mathbb{1}\{m = 2\}\right),$$

for time periods $m = 1, 2$.

As with the model without misclassification, removing the Markov assumption gives a significant improvement. The likelihood ratio statistic is 15.0 on 2 df so we again have a significant improvement in fit compared to the Markov model. We also find an increasing hazard for state 1 and a decreasing hazard for state 3. The estimates of covariate effects stay virtually unchanged between models. This again shows that the covariate effects seem robust to moderate changes in the baseline intensities.

### 5.2.6 Conclusion

Models with piecewise constant transition intensities have the advantage of allowing analytically tractable expressions for the likelihood. For Markov and hidden Markov models, a time inhomogeneous model with piecewise constant intensities is only slightly more difficult to fit than a time homogeneous model. As shown in this section, the principle of piecewise constant hazards can be applied to the semi-Markov case. Whilst it is still true that the likelihood from such models is analytically tractable, the difficulty of fitting the model is greater.

In this chapter we have only presented a semi-Markov model in which transition intensities have two intervals of constant hazard. If the number of intervals is increased, the computation time required to enumerate all the regions of constant hazard grows exponentially. Also, whilst a 4 state model where absorption times are known (such as the CAV data) produces two dimensional distributions of sojourn times, the dimension increases as

Table 5.6: Comparison of parameter estimates and 95% confidence intervals for a time homogeneous hidden Markov model and a hidden semi-Markov model with piecewise-constant intensities for the CAV data with misclassification

| | Markov model | | semi-Markov model | |
|---|---|---|---|---|
| Parameter | Estimate | 95% CI | Estimate | 95% CI |
| $q_{12}^{(1)}$ | 0.033 | (0.021,0.050) | 0.027 | (0.017,0.042) |
| $q_{14}^{(1)}$ | 0.021 | (0.015,0.029) | 0.020 | (0.013,0.025) |
| $\tau_1$ | | | 0.495 | (0.210,0.790) |
| $q_{23}$ | 0.190 | (0.143,0.252) | 0.210 | (0.157,0.282) |
| $q_{24}$ | 0.053 | (0.029,0.099) | 0.024 | (0.006,0.120) |
| $q_{34}^{(1)}$ | 0.155 | (0.120,0.201) | 0.264 | (0.165,0.420) |
| $\tau_3$ | | | -0.740 | (-1.363,-0.116) |
| $e_{12}$ | 0.025 | (0.015,0.042) | 0.025 | (0.015,0.043) |
| $e_{21}$ | 0.186 | (0.123,0.272) | 0.177 | (0.116,0.262) |
| $e_{23}$ | 0.066 | (0.038,0.108) | 0.066 | (0.039,0.109) |
| $e_{32}$ | 0.102 | (0.051,0.194) | 0.096 | (0.047,0.184) |
| $\beta_{12}^{(\mathrm{IHD})}$ | 0.520 | (0.234,0.807) | 0.529 | (0.239,0.819) |
| $\beta_{12}^{(\mathrm{dage})}$ | 0.025 | (0.013,0.037) | 0.027 | (0.015,0.039) |
| $-2\times$ LL | 3933.6 | | 3918.6 | |

the number of states increases. Computation of the likelihood requires that the shape of each region of equal hazard is determined. This becomes harder to program and is more computationally intensive as the dimension increases.

Whilst for the CAV data incorporating misclassification of states is still feasible, it is less feasible for the BOS data. The BOS data contain some subjects who have over 100 observations, meaning the number of possible paths of true states is considerably higher. Applying the same method would lead to 62114 distinct sets of states if misclassification only to adjacent states is assumed and considerably more if misclassification from state 1 to state 3 is allowed. Computation of the likelihood in this case becomes too time consuming to be practical. The range of applications for which a hidden semi-Markov model can be fitted using piecewise constant intensities is therefore limited.

In piecewise constant intensities models it is also necessary to specify the location of the boundary points and in most cases the choice will be arbitrary. Moreover, it may also be considered unsatisfactory, for instance in terms of biological feasibility, for the transition intensities to contain discontinuities. A more satisfactory approach may be to define continuous, non-constant intensities. This is the focus of the next section.

## 5.3 Numerical solutions to Kolmogorov forward equations

The popularity of the piecewise constant hazard approach to fitting inhomogeneous Markov models stems partly from the lack of alternatives. If we allow the transition intensity matrix $Q(t)$ to be an arbitrary function, then the Kolmogorov forward equations,

$$\frac{dP(t_1,t)}{dt} = P(t_1,t)Q(t), \tag{5.8}$$

define a system of first order non-linear differential equations which cannot be solved analytically except in special cases. In the case of progressive models, direct numerical integration is a possibility. Some authors [19, 62] have used this technique, but were restricted to models in which only one state had time varying transition intensities. If more states have non-constant intensities, the integration becomes multi-dimensional and is generally slow to compute.

Anisimov *et al* [11] considered a model with time varying hazards using the Euler method for solving differential equations. Previously, Ocaña-Riola [100] suggested using the approach as a simpler alternative to computing a matrix exponential when calculating the

transition probabilities for a time homogeneous Markov model. These methods are reviewed here.

### 5.3.1 The Euler method

The Euler method is the simplest iterative numerical approach to solving initial boundary problems for ordinary differential equations (ODEs). Given a differential equation of the form

$$\frac{dy(t)}{dt} = f(y, t)$$

and initial value $y(0)$, the value of $y(h)$, for some sufficiently small $h$ can be approximated by

$$y(h) \approx y(0) + hf(y(0), 0).$$

The value of $y$ at subsequent times can be found by repeatedly applying the formula

$$y(mh) \approx y((m-1)h) + hf(y((m-1)h, (m-1)h).$$

For a time inhomogeneous Markov model in which the transition intensity matrix is given by $Q(t)$, such that the transition probability matrices are found by solving equation 5.8, we can start with the initial transition probability matrix $P(t_0, t_0) = I$, and we can write

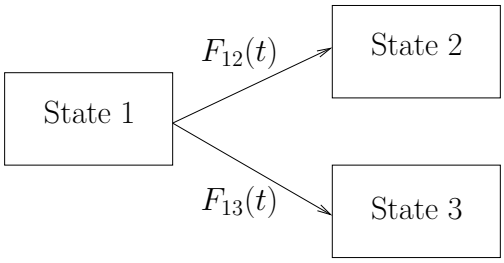
$$P(t_0, t_0 + h) \approx P(t_0, t_0) + hP(t_0, t_0)Q(t_0)$$

and subsequent points can be found using

$$P(t_0, t_0 + mh) = P(t_0, t_0 + (m-1)h) + hP(t_0, t_0 + (m-1)h)Q(t_0 + (m-1)h). \quad (5.9)$$

However, the Euler method is only a first-order approximation. In all cases, $h$ has to be chosen to be sufficiently small to avoid significant errors in the estimates. However, there are systems of equations for which the Euler method will perform poorly even for very small $h$. Such equations are often called *stiff* differential equations [121], although no formal definition of a stiff equation exists. Given a known initial value $y(0)$, any approximate step will result in the estimated value $\tilde{y}(h)$ being slightly different from the true value $y(h)$. For stiff problems, this error, which will be small at time $h$, grows to a significant error at later times.

Figure 5.2: Weibull competing risks model



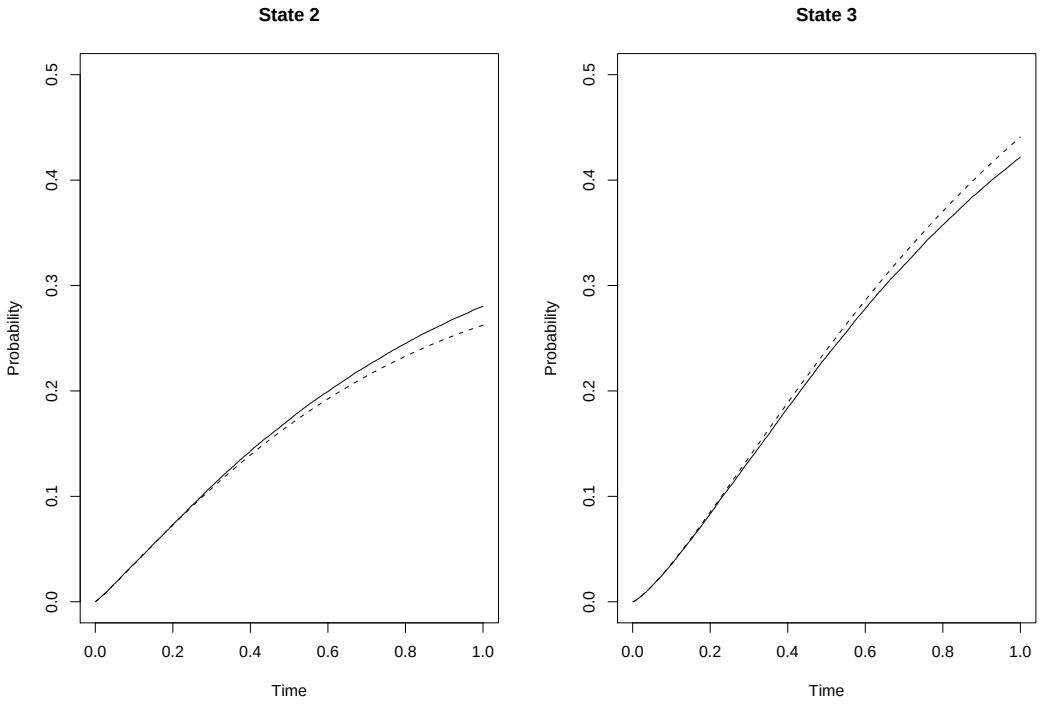

**Example: Competing Weibull risks model**

Consider a model with an initial state and two absorbing states, as shown in figure 5.2. Let the transition intensities from the initial state to each absorbing state have Weibull intensities with differing shape and rate parameters. This is a simple competing risk model, for instance the initial state could be 'alive' and the two absorbing states might represent different causes of death. The Euler method performs poorly at estimating the transition probabilities given occupancy in state 1 at time 0. Figure 5.3 shows the true solution (arrived at using numerical quadrature) and the approximate solution using the Euler method for the case where the time to enter state 2 is $F_{12} \sim$ Weibull(1.1,0.5) hazard and the time to enter state 3 is $F_{13} \sim$ Weibull(1.3,0.8) for a step size of $h = 0.0001$. There is a systematic overestimation of the occupancy in state 3 and a systematic underestimation for state 2. This pattern of bias occurs at each step because, for all $t > 0$, the hazard of entry into state 2 is increasing faster than the hazard of entry into state 3. State 1 occupancy remains reasonably well approximated in this case. Clearly, if maximum likelihood estimation was attempted using this approximation to the transition probabilities we would get an inaccurate estimate of the mle.

### 5.3.2 Advanced methods for solving differential equations

As shown, the Euler method can lead to inaccurate estimates of transition probabilities, even for quite simple models. A more reliable procedure must assess whether the system of ODEs to be solved are stiff or non-stiff, and must keep track of the error bounds for the system and choose appropriate methods and step-size to ensure the required accuracy. Many computer programs and packages have such checks built in so that the user does not have to determine for themselves the difficulty of the problem. For instance, the LSODA

Figure 5.3: Comparison of state occupancy estimates for competing Weibull risks three state model for numerical quadrature (bold line) and Euler method (dashed line).




solver [108], automatically switches between methods suitable for non-stiff ODEs and to more computationally intensive methods for stiff ODEs, when necessary. This routine, originally programmed in FORTRAN is available in the **odesolve** package in R [119]. The user specifies a system of differential equations to be solved, defines initial conditions and specifies the time points at which a solution is desired. For a time homogeneous Markov model we can specify the system as each component in $P(0,t)$, with derivative

$$\frac{dP(0,t)}{dt} = P(0,t)Q(t)$$

and set the times to be the observation times of the process, $t_1, \ldots, t_N$. The function gives us the output $P(0,t_1), \ldots, P(0,t_N)$. However, to evaluate the likelihood we require transition probability matrices $P(t_i, t_{i+1})$. These however, can easily be retrieved because the Chapman-Kolmogorov equation

$$P(0,t) = P(0,u)P(u,t)$$

implies that

$$P(u,t) = P(0,u)^{-1}P(0,t).$$

Hence the transition probabilities between any two times $t$ and $u$ can be expressed as a function of transition probabilities with respect to time 0. For many models, calculation of the inverted transition probability matrices, $P(0,u)^{-1}$, is the most computationally intensive part of the likelihood calculation. The LSODA package has no problems determining the correct transition probabilities for the competing Weibull risks model in section 5.3.1. Standard Nelder-Mead or BFGS approaches to numerical optimisation can be used to maximise the likelihood.

Methods based on solving differential equations cannot be applied to calculate the likelihood for semi-Markov models because the likelihood is neither expressible as a simple product of transition probabilities nor are these transition probabilities defined by differential equations. For progressive models, the likelihoods for semi-Markov models can be expressed as an integral over the space of sojourn times. These integrals are generally intractable making standard numerical optimisation techniques too time consuming. Instead a Monte-Carlo EM algorithm can be applied. This method is also applicable to time inhomogeneous Markov models and comparison with directly solving the ODEs numerically is made in the next section.

## 5.4 Monte-Carlo Expectation-Maximisation algorithm

In this section, an outline of the Monte-Carlo EM algorithm is given, along with details of how it can be used to fit certain types of time inhomogeneous Markov models or semi-Markov models, for which existing methods are too time consuming. Deltour *et al* [41] used an MCEM algorithm for data modelled as being from a discrete-time Markov chain with intermittent missing observations. Cook *et al* [30] mentioned the possibility of using an MCEM algorithm to fit random-effects continuous-time Markov models. However, there does not seem to have been any previous use of MCEM algorithms to fit multi-state models to panel observed data.

### 5.4.1 Expectation-Maximisation algorithm

The Expectation-Maximisation (EM) algorithm is a well established method of maximising likelihood functions in the presence of missing data [42]. The data are partitioned as $(\mathbf{y}, \mathbf{z})$ where $\mathbf{y}$ is observed and $\mathbf{z}$ unobserved. The algorithm is as follows

1. Choose initial values $\theta_0$ for the model parameters $\theta$

2. *E-Step* At the $r$th iteration, the conditional expectation of the complete-data log-likelihood is computed given the observed data $\mathbf{y}$ and the current values of the parameters, $\theta_r$:

$$Q(\theta|\theta_r) = \mathbb{E}_{\theta_r}(l(\theta, \mathbf{x})|\mathbf{y}).$$

3. *M-Step* The new values $\theta_{r+1}$ of the parameters are chosen to maximise $Q(\theta|\theta_r)$ with respect to $\theta$.
4. Repeat steps 2 and 3 until convergence.

At each iteration of the algorithm, the observed likelihood is guaranteed to improve, i.e. $l(\theta_r, \mathbf{y}) \geq l(\theta_{r-1}, \mathbf{y})$, this is called the *ascent property*.

The EM algorithm is useful when $f(\mathbf{y}, \mathbf{z}|\theta)$, for $\mathbf{z}$ known, is easy to maximise and where the E-step is tractable. As discussed in section 1.2.4, the Baum-Welch or Forward-Backward algorithm [12] for maximising the likelihood for hidden Markov models, is one example of an EM algorithm.

### 5.4.2 Application to multi-state modelling

The incomplete nature of panel observed multi-state data suggests an EM-type algorithm may be applicable. The observed data take the form of a series of observed states at the sampling times. Since the likelihood for data subject only to right-censoring of transition times is generally easy to compute, this motivates 'completing' the data so that it is right censored.

For example for a inhomogeneous Markov process, if the state $x_0$ at time $t_0$ is known and the times of all transitions are also known, so that transitions to states $x_1, \ldots, x_N$ occurred at times $t_1, \ldots, t_N$, then the likelihood is of the form

$$L(\theta) = \prod_{i=1}^{N} q_{x_{i-1},x_i}(t_i, \theta) \exp\left( \int_{t_{i-1}}^{t_i} q_{x_{i-1},x_{i-1}}(u, \theta) du \right)$$

where $q_{rs}(t; \theta)$ is the transition intensity between states $r$ and $s$ at time $t$ given parameters $\theta$ and $q_{rr}(t; \theta)$ is the $r$th diagonal entry of the transition intensity matrix at time $t$. Hence provided the transition intensities are of an appropriate functional form so that their partial integrals with respect to $t$ are easy to compute, the overall likelihood is also straightforward.

Time homogeneous semi-Markov processes can be similarly represented. Now, it is the sojourn times within each state that are important, so suppose we again have transitions to states $x_1, \ldots, x_N$ at times $t_1, \ldots, t_N$ and that the process was initiated in state $x_0$ at time $t_0$. This implies there were sojourns of length $u_0 = t_1 - t_0, \ldots, u_{N-1} = t_N - t_{N-1}$ in states $x_0, \ldots, x_{N-1}$. Thus the likelihood can be written as:

$$L(\theta) = \prod_{i=0}^{n-1} q_{x_i, x_{i+1}}(u_i, \theta) \exp\left(-Q_{x_i}(u_i, \theta)\right)$$

where now $q_{rs}(t, \theta)$ is the transition intensity from state $r$ to $s$ given time $t$ since entry into state $r$ and $Q_r(t, \theta) = \int_0^t \sum_{j \neq r} q_{rj}(u, \theta) du$, the overall integrated hazard of intensities out of state $r$, given time $t$ since entry into $r$. As with the time inhomogeneous Markov case, provided these hazards are of an easily integrable form, this complete likelihood is easy to calculate.

The difficulty in applying the EM algorithm to these situations is in performing the E-step, the conditional expectation of the complete-data likelihood, given the observed data $\mathbf{y}$ and the parameters $\theta$. Such an expectation in the case of inhomogeneous Markov or homogeneous semi-Markov data will involve both a summation over all admissible sequences of states and integration over the space of possible transition times given each particular sequence of states. This computation is at least as difficult as simply evaluating the observed likelihood directly as it also involves evaluating intractable integrals. Hence the EM algorithm itself is not useful when the likelihood is intractable.

### 5.4.3 Monte Carlo Expectation-Maximisation algorithm

When the E-step of the EM algorithm is difficult to compute a potential solution is the Monte-Carlo Expectation-Maximisation algorithm [133]. In this variation of the EM algorithm, the E-step of the algorithm is approximated by Monte Carlo methods. This involves drawing $M$ samples $Q_1^*(\theta|\theta_r), \ldots, Q_M^*(\theta|\theta_r)$ from the distribution of $l(\theta, x)$ given the observed data $y$, and taking

$$Q^*(\theta|\theta_r) = \frac{1}{M} \sum_{i=1}^{M} Q_i^*(\theta|\theta_r)$$

as the quantity to be maximised in the M-step. To apply the MCEM algorithm, we therefore only need to be able to sample from the distribution of $\mathbf{z}|\mathbf{y}$.

However, replacing the E-step with a Monte Carlo approximation does not come without some loss of efficiency. Firstly, since the sample $Q_1^*(\theta|\theta_r), \ldots, Q_M^*(\theta|\theta_r)$ is random, unlike

the EM algorithm, the path $\theta_0, \theta_1, \ldots$, conditional on $\theta_0$ and $\mathbf{y}$, is not deterministic. Moreover, there is no guarantee that at each iteration, there will be an improvement in the observed likelihood.

Intuitively, we expect that it is much easier to get an improvement in the observed likelihood when we are a long way from the optimal $\theta$ and harder in the locality of the optimum. As $M$ tends to infinity, by the weak law of large numbers, the step performed in the MCEM algorithm tends in probability to the equivalent step in the EM algorithm, which will have the ascent property. Thus near to the optimum, a large $M$ is needed so that the Monte-Carlo error does not dominate $|\theta_r - \theta_{r-1}|$. Conversely, it is inefficient to have a large $M$ at the beginning, when $|\theta_r - \theta_{r-1}|$ is large and an improvement in the likelihood has a high probability for small $M$.

Hence, in order to ensure convergence to the optimum, it is necessary to increase $M$ as $r$ increases. Methods for how to increase $M$ with $r$ are an active area of research [13, 15]. In the application to multi-state models we will adopt an *ad hoc* approach to increasing $M$. We start with $M = M_0$ and keep this value for the first 10 iterations. Subsequently, if on two occasions the optimised full likelihood on the $r$th step is not an improvement on the $(r-1)$th step, then the sample size $M$ is doubled. Final convergence is determined by considering the difference in parameter estimates between two steps. If this is less than some limit, for instance $1 \times 10^{-6}$, then the algorithm terminates.

### 5.4.4 MCEM for multi-state models

The observed data for a multi-state model is the set of states occupied at the sampling times. These provide information on the transition times and transition types. In order to 'complete' the dataset based on the observed data, it is necessary to sample from the conditional distributions of the transition times, given the data.

#### Conditional distributions

For the case in which the times of transition are only interval censored, i.e. the type of transition is known but the time is only known to lie within an interval, $[t_1, t_2]$, it is only the transition times, or equivalently the sojourn times in each state, that are needed. From panel observation, such interval censored data occur if the model is unidirectional. Suppose, we have an $R$ state unidirectional process. For an individual who is right censored

in state $J$ at time $s_J$, the observed data, $D$, imply that the transition times between state $j$ and $j+1$, $T_j$, is such that $T_j \in [t_{j1}, t_{j2}]$ for $j = 1, \ldots, J-1$, where $t_{j1} \leq t_{j2} \quad \forall j$ and $t_{1k} \leq t_{2k} \leq \ldots \leq t_{J-1,k} \quad \text{for} \quad k = 1, 2$.

We therefore have that the conditional distribution of transition times $u_1, ..., u_{J-1}$

$$f(u_1, \ldots, u_{J-1}|D) \propto f_1(u_1; 0) f_2(u_2; u_1) \ldots f_{J-1}(u_{J-1}; u_{J-2}) F_J(u_J; u_{J-1})$$

for values of $u_1, \ldots, u_{J-1}$ satisfying $u_j \in T_j$ and $u_1 < u_2 < \ldots < u_{J-1}$. Here $f_r(u; t)$ refers to the probability density that arises for entering state $r$ at time $t$ and exiting at time $u$. This notation allows generality for time inhomogeneous Markov and time homogeneous semi-Markov models.

For more general progressive models, panel observation will lead to situations where a number of different sequences of states is possible given the observed data. In this situation the missing data can be partitioned into $\delta$, defining the sequence of states that occurred, and $\mathbf{s}|\delta$ the transition times conditional on the sequence of states. Suppose $\delta$ implies a set of states $d_1, \ldots, d_r$ are visited in that order. Conditional on $\delta$ the transition time between state $d_j$ and $d_{j+1}$ will lie in an interval $[t_{d_j,1}, t_{d_j,2}]$. We then have

$$f(\delta, u_{d_1}, \ldots, u_{d_{r-1}}|D) \propto f_{d_1,d_2}(u_{d_1}; 0) f_{d_2,d_3}(u_{d_2}; u_{d_1}) \ldots f_{d_{r-1},d_r}(u_{d_{r-1}}; u_{d_{r-2}}) F_{d_r}(u_{d_r}; u_{d_{r-1}})$$

when $u_{d_j} \in [t_{d_j,1}, t_{d_j,2}]$ and $u_{d_1} < \ldots < u_{d_j}$.

Note that the probability of a particular sequence of states $\delta$ given observed data is not easily calculable. Moreover, the likelihood contribution of a particular $\delta$ may depend on the particular values of $u_{d_j}$.

When the subject reaches the absorbing state the form of the conditional distribution does not change markedly. The only change is that there is no final term involving a cdf $F$. Moreover, when the time of entry into the absorbing state, $R$, is known exactly, there is no difference except that the time of entry into the absorbing state is not part of the missing data and a $f_{d-1,R}$ term appears in the product.

If the model is not progressive then the above method is less feasible. The potential number of intermediate sojourns between observed states is unbounded, and therefore so is the potential amount of missing data. This makes determining the conditional distributions of transition times and state sequences from which to sample more difficult and time consuming.

### 5.4.5 Monte Carlo sampling methods

The conditional distributions which arise, both for unidirectional and more general progressive models, are non-standard and, unless particular choices (e.g. piecewise exponential) of functions $f$ are chosen, will not allow sampling by simple methods such as inversion.

#### Gibbs sampler

In the special case of data from a unidirectional model, where all the transition times are interval censored, and a time homogeneous semi-Markov model is assumed, a Gibbs sampler approach [54] can be used. In order to do this, it is necessary to reparametrise in terms of the sojourn times $\tilde{u}_j$ in each state. Conditional on sojourn times $\mathcal{S}_{(k)} = \{\tilde{u}_1, \dots, \tilde{u}_{k-1}, \tilde{u}_{k+1}, \dots, \tilde{u}_{J-1}\}$, namely all except $\tilde{u}_k$

$$f(\tilde{u}_k|\mathcal{S}_{(k)}, D) \propto \quad \tilde{f}_k(\tilde{u}_k)\mathbb{1}^*(\mathcal{S}_{(k)}, D), \quad t_{u1} < \tilde{u}_k < t_{u2}$$

where $t_{u1}, t_{u2}$ depend on $\mathcal{S}_{(k)}$, $\mathbb{1}^*(\mathcal{S}, D)$ is an indicator function taking value 1 if a set of sojourn times $\mathcal{S}$ is consistent with data $D$ and zero otherwise, and $\tilde{f}_k(t)$ refers to the pdf of the sojourn distribution in state $k$. Hence the conditional sojourn distribution will simply be a truncated version of the chosen sojourn distribution. Therefore, if a distribution e.g. Gamma or Weibull, with easily calculable inverse cdf, has been assumed, sampling from these conditional distributions will be straightforward.

A standard Gibbs sampler algorithm can then be used, taking some feasible initial set of sojourn times.

#### Metropolis-Hastings based sampling

It does not seem possible to use a Gibbs sampler for time inhomogeneous Markov models because the transition time alters the sojourn distributions in subsequent states. Nor can the Gibbs sampler be used for more general progressive models where the sequence of states is not known. In these situations we propose the use of a Metropolis-Hastings (MH) algorithm [58].

Suppose we wish to sample from a distribution with density $f$, which is known up to a constant. The MH algorithm allows a Markov chain, $\{X^{(r)} : r = 1, 2, \dots\}$, with stationary

density $f$ to be constructed. The $(r+1)$th value iteration of the algorithm depends on the $r$th.

**1.** Sample a candidate value $Y$ from a proposal distribution $\pi(Y|X^{(r)})$. This distribution may depend on the current value $X^{(r)}$.

**2.** Accept $X^{(r+1)} = Y$ with probability

$$\min\left(1, \frac{f(Y)\pi(X^{(r)}|Y)}{f(X^{(r)})\pi(Y|X^{(r)})}\right).$$

Although the stationary distribution of the chain will be $f$, regardless of the choice of proposal $\pi$, the rate of convergence may vary greatly.

For the models fitted in the proceeding section, an independence sampler, that assumes uniform distributions over the space of transition times conditional on the sequence of states $\delta$, seems to perform well, provided reasonable values for $\mathbb{P}(\delta)$ are chosen. Various methods of approximating the probabilities of each sequence are possible. However, a simple, and usually adequate approach is to simply consider the midpoint of the space of transition times for each $\delta$ and assume the density is proportional to this value throughout. This method seems to perform well for the examples in this chapter.

### 5.4.6 Standard Error estimates

Louis [91] showed that the observed Fisher information matrix for a model fitted with the EM algorithm is given by:

$$I_O(\theta) = \mathbb{E}(I_F(\theta)) - Cov(U_F(\theta))$$

where $I_O$ is the observed Fisher information, $I_F$ is the Fisher information of the full data and $U_F$ is the score vector of the full data. Using a MCEM algorithm, we cannot compute the expectation of the information or the covariance of the score analytically. Instead, like Deltour *et al* [41], we use the sample mean and sample covariance using generated data at the maximum likelihood estimates, $\hat{\theta}$. Hence we take

$$\frac{1}{M}\sum_{i=1}^{M}\frac{\partial^2 l}{\partial\theta^2}(\hat{\theta}, x_i) - \frac{1}{M}\sum_{i=1}^{M}\left[\frac{\partial l}{\partial\theta}(\hat{\theta}, x_i)\right]\left[\frac{\partial l}{\partial\theta}(\hat{\theta}, x_i)\right]^T$$

to be our estimate of the observed information matrix, where $x_i$ is the full data generated for dataset $i$. The second term is just $\mathbb{E}(U(\theta)U(\theta)^T)$ because we can assume $\mathbb{E}(U(\hat{\theta})) = 0$. Standard errors can then be derived in the usual way.

### 5.4.7 Limitations of the method

As already mentioned, the MCEM algorithm method can only be applied to the case of progressive models because for bi-directional models the space of 'missing' data is unbounded.

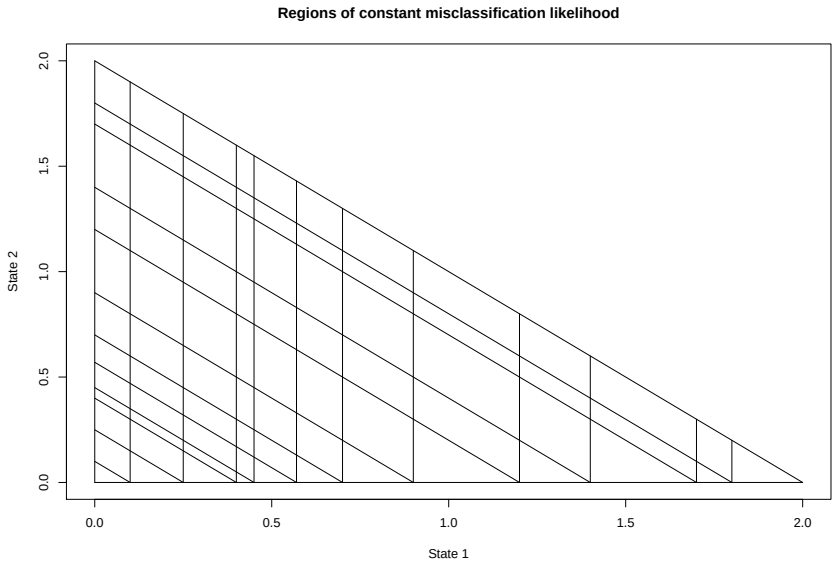
Theoretically, there is no reason why the method cannot be extended to the case of data with misclassified observed states. The form of the MCEM algorithm can in principle remain the same. But the sojourn distributions conditional on the observed states become intricate and sampling from them is more challenging. As before, provided the initial state is known, the observed states determine a region of minimum and maximum sojourn times. The misclassification model widens the possible range of sojourn times, but makes the range of values within the conditional distribution much greater. Within this region there will be subregions with boundaries corresponding to all the observation times of the process. Figure 5.4 depicts the range of boundaries for an individual observed 11 times and censored at time 2 under the 4-state CAV model (figure 2.1) with misclassification to adjacent states. Within each of these subregions the density is smooth, but there are large discontinuities at the boundaries. Typically, the subregions corresponding to the fewest misclassified states have a higher probability density. However, these subregions are not necessarily adjacent, so the density may be multi-modal.

It is not a trivial task to devise a proposal distribution for the MH algorithm that allows adequate mixing. An independence sampler and a random walk proposal were tried, but both resulted in an over-representation of the more likely sojourn times and an under-representation of rare sojourn times. For the MCEM algorithm method to be effective, the burn-in times need to be short because a different chain has to be run for each subject at each iteration. Thus even if a reasonably effective proposal distribution could be found, it might still prove to require too many iterations to be practical.

## 5.5 Application to CAV data

In this section the methods for fitting multi-state models via an MCEM algorithm are applied to the CAV data without misclassification as described in section 2.1.1. Firstly, a semi-Markov model with Weibull transition intensities is fitted. Subsequently a restricted model where only the onset rate has a Weibull hazard and which is therefore time inhomogeneous Markov, is fitted. This second model can also be fitted using the approach of

Figure 5.4: Regions of constant misclassification likelihood contributions for an individual observed 11 times and censored at time 2.




section 5.3.1 where the ODEs are solved numerically.

The Weibull distribution is commonly used in survival analysis and has also been used by some authors for the intensities in time inhomogeneous Markov models [99, 125]. The distribution has two parameters, the rate parameter $\lambda$ and the shape parameter $\alpha$. The density is given by

$$f(x;\alpha,\lambda) = \lambda^{\alpha} x^{\alpha-1} \exp\left(-(\lambda x)^{\alpha}\right), x > 0.$$

When $\alpha = 1$, the distribution is Exponential with rate $\lambda$, if $\alpha < 1$ there is a decreasing hazard and if $\alpha > 1$ there is an increasing hazard.

### 5.5.1 Semi-Markov model

The first model assumes that the transition intensities within each state depend on the time since entry into that state:

$$q_{rs}(t) = \alpha_{rs}\lambda_{rs}(\lambda_{rs}t)^{\alpha_{rs}-1}$$

where $\alpha_{rs}$ and $\lambda_{rs}$ denote the shape and rate parameters respectively and $t$ is time since entry to state $r$. Note that if $\alpha_{rs} = 1$, the intensity is constantly $\lambda_{rs}$. Sojourn time

Table 5.7: Comparison of parameter estimates and 95% confidence intervals for a time homogeneous Markov model and a semi-Markov model with Weibull intensities for the CAV data without misclassification

| | Markov model | | semi-Markov model | |
|---|---|---|---|---|
| Parameter | Estimate | 95% CI | Estimate | 95% CI |
| $\alpha_{12}$ | 1 | | 1.203 | (1.086,1.332) |
| $\lambda_{12}$ | 0.094 | (0.082,0.107) | 0.102 | (0.091,0.115) |
| $\lambda_{14}$ | 0.023 | (0.017,0.030) | 0.031 | (0.024,0.041) |
| $\lambda_{23}$ | 0.200 | (0.162,0.246) | 0.224 | (0.177,0.283) |
| $\lambda_{24}$ | 0.040 | (0.022,0.073) | 0.014 | (0.001,0.198) |
| $\alpha_{34}$ | 1 | | 0.734 | (0.547,0.986) |
| $\lambda_{34}$ | 0.146 | (0.116,0.184) | 0.167 | (0.119,0.235) |
| $-2\times$ LL | 3553.0 | | 3537.5 | |

distributions in this model are not necessarily Weibull. In general the sojourn distribution for state $r$ will be given by

$$\min_s (T_{rs})$$

for $s = 1, \ldots, R$ where $T_{rs}$ are independent Weibull$(\lambda_{rs}, \alpha_{rs})$.

Initially a model allowing complete flexibility in the transition intensities was fitted. However, the algorithm failed to converge to a point with a positive definite Hessian indicating that the likelihood is quite flat. The parameter values reached however, gave an indication that the time effect of state 2 was insignificant as the estimated shape parameter was close to 1.

To allow the parameters to be estimated various constraints were made. Firstly, the sojourn time in state 2 was fixed as Exponential, or equivalently $\alpha_{23} = \alpha_{24} = 1$. Secondly, the competing hazards in state 1 were constrained to have the same shape parameter, $\alpha_{12} = \alpha_{14}$. This implies that the relative hazard of progressing to state 2 compared to the hazard of going directly to death remains constant in time.

Table 5.7 compares the estimated parameters for the semi-Markov model with those of the time homogeneous Markov model which is a special case of the semi-Markov model. The semi-Markov model estimates an increasing hazard in state 1, and a decreasing hazard in state 3. Both estimates are significantly different from 1 (the parameter value for a

Table 5.8: Comparison of estimates and 95% confidence intervals for time inhomogeneous Markov model on the CAV data using numerical solutions to ODE or an MCEM algorithm.

| | Numerical ODE | | MCEM algorithm | |
|---|---|---|---|---|
| Parameter | Estimate | 95% CI | Estimate | 95% CI |
| $\alpha_{12}$ | 1.204 | (1.088,1.332) | 1.204 | (1.089,1.332) |
| $\lambda_{12}$ | 0.103 | (0.091,0.115) | 0.103 | (0.091,0.116) |
| $\lambda_{14}$ | 0.031 | (0.024,0.041) | 0.031 | (0.024,0.041) |
| $\lambda_{23}$ | 0.202 | (0.164,0.249) | 0.203 | (0.165,0.248) |
| $\lambda_{24}$ | 0.038 | (0.020,0.070) | 0.037 | (0.020,0.070) |
| $\lambda_{34}$ | 0.145 | (0.115,0.182) | 0.145 | (0.115,0.182) |
| $-2\times$ LL | 3541.1 | | 3541.1 | |

Markov model). A likelihood ratio test of the Markov versus the semi-Markov model gives $\Lambda = 15.5$ on 2 degrees of freedom. There is general agreement between the estimates of this model and the semi-Markov model with piecewise-constant intensities of section 5.2.2.

### 5.5.2 Time inhomogeneous Markov model

Given that $\hat{\alpha}_{34}$ was only marginally significant (p=0.04) and that its inclusion as an unknown parameter caused significant difficulty in estimating $\lambda_{24}$, a more appropriate model may be the time inhomogeneous Markov model where only state 1 transition intensities are time dependent, or equivalently $\alpha_{34}$ is fixed at 1. Now, we have a choice of methods available; the MCEM algorithm or numerically solving the forward equations using the LSODA package. Each method converges to a similar set of parameter estimates. The confidence intervals are in close agreement for all parameters. $-2 \times \text{LL} = 3541.1$ for this model on 6 parameters. This presents a significant improvement compared to the time homogeneous model presented in section 2.1.2 ($-2 \times \text{LL} = 3552.92$, 5 parameters) and the model with $Q(t) = Q_0 \exp(-\mu t)$ presented in section 2.5.3 ($-2 \times \text{LL} = 3549.64$, 6 parameters). However, the piecewise constant intensities Markov model presented in section 5.2.1, although based on a greater number of parameters, represents a better fit ($-2 \times \text{LL} = 3531.60$ from 8 parameters).

## 5.6 Conclusions about the CAV data

The implications of this chapter to our understanding of the CAV data are that we have demonstrated there is significant time dependency. This is particularly clear for onset rates where the transition intensity is increasing with time. Since the mortality rate increases with onset, this has implications for estimates of mean sojourn time and prevalence. The extent and nature of any time dependencies for other transitions is harder to determine. A time inhomogeneous Markov model with piecewise constant intensities suggested a decrease in $2 \rightarrow 3$ transition rates after 3 years. However, a semi-Markov model found a significantly decreasing hazard for $3 \rightarrow 4$ transitions, measured on time since entry into the state. In terms of the likelihood, the time inhomogeneous Markov model is preferred. However, it is not clear whether the semi-Markov effects would become insignificant if time dependency with respect to time since transplant was taken into account.

Figure 5.5 gives the estimated survival curves from the Kaplan-Meier estimate, the time homogeneous Markov model, and the Markov and semi-Markov piecewise-constant intensities models.

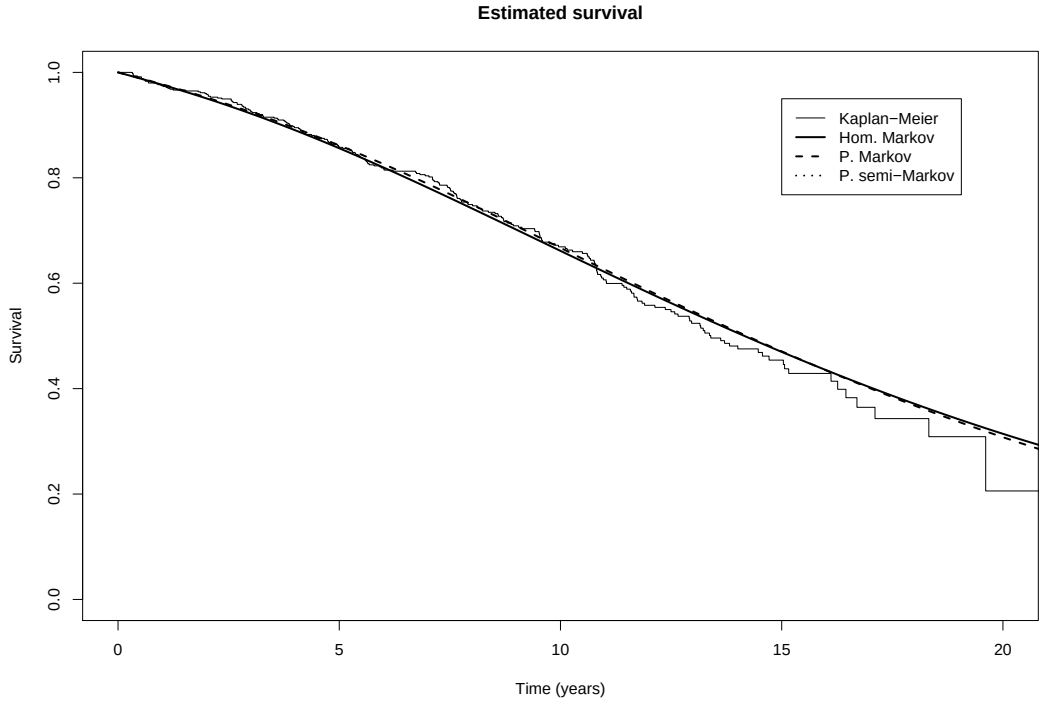
All estimates are in broad agreement up to around 10 years. Beyond this point the time homogeneous Markov model begins to diverge from the Kaplan-Meier estimate, although, as established in chapter 2, by not a significant amount. Despite the significant improvement in the fit between the homogeneous Markov model and the piecewise time inhomogenous Markov model, there is virtually no difference between their fitted survival curves.

Table 5.9 gives the estimated mean post-transplant lifetime for each model based upon the CAV data without misclassification. There is some variability in the point estimates between models, these differences are larger due to extrapolating the fitted models beyond the follow-up time in the data. The width of the 95% confidence intervals also varies, with the Markov models having the least uncertainty and the Weibull semi-Markov model having the most.

## 5.7 Conclusion

This chapter has presented a range of methods for fitting time inhomogeneous Markov or semi-Markov models. The piecewise constant hazards approach provides a simple way of fitting time inhomogeneous Markov models. For the CAV data, the piecewise time

Figure 5.5: Estimated survival curves for different models on CAV data without misclassification




inhomogeneous Markov model had the best fit in terms of $-2\times$ LL. This suggests that the piecewise-constant hazard approach can provide more flexible models for time dependency. However, piecewise models are limited by the necessity to make a choice about the location and number of change points for the intensities.

Here we have extended the concept of piecewise-constant intensities to allow the fitting of progressive semi-Markov models. While the likelihood can be computed without using quadrature, a complicated and intricate method of identifying regions of constant hazard is needed. This makes the method impractical for a model with more than one change point in hazard, or for models with more than 4 states. The advantage of this method over the alternative MCEM approach is that, at least for data where the number of observations per patient is moderate (e.g. less than 20), hidden semi-Markov models with misclassification of observed states can also be accommodated.

The use of the Euler method to numerically solve the Kolmogorov forward equations for

Table 5.9: Estimates of mean post-transplant lifetime for the competing CAV models

| Model | Estimate | 95% CI | -2 × LL | Parameters |
|---|---|---|---|---|
| Homogeneous Markov | 16.530 | (15.105,17.970) | 3553.0 | 5 |
| Piecewise Markov | 15.829 | (14.335, 17.256) | 3531.6 | 8 |
| Weibull Markov | 15.658 | (14.371,17.073) | 3541.1 | 6 |
| Piecewise semi-Markov | 16.090 | (14.362,17.684) | 3538.0 | 7 |
| Weibull semi-Markov | 16.632 | (12.345,19.007) | 3537.5 | 7 |

inhomogeneous Markov models used by Anisimov *et al* [11], was explored. Whilst for standard time homogeneous Markov models, this approach works well, more generally the estimated solution can diverge significantly from the true solution. It does not therefore seem a viable approach in more complicated cases. However, more sophisticated methods for solving differential equations exist and are available in standard software packages such as R. These are effective on a much greater range of models.

The bulk of the chapter was devoted to an MCEM algorithm for fitting progressive time inhomogeneous Markov or semi-Markov models. For time inhomogeneous Markov models, it will generally be preferable to use a direct numerical solution of the Kolmogorov forward equations, as the fitting procedure for the MCEM algorithm is slower, particularly if the likelihood surface is quite flat. Moreover, misclassification HMMs with time dependent intensities can also be fitted by the direct numerical solution approach. However, the ability to fit semi-Markov models with standard parametric sojourn distributions like Weibull or Gamma is a unique feature of the MCEM method. Also, the MCEM algorithm approach for semi-Markov models is generally simpler to apply than the piecewise constant hazards method. However, the inability to fit models which can cope with misclassification of states is an obvious limitation.

Further constraints to the adoption of more complicated models to panel observed data relate to problems of parameter identifiability and estimability. Strict identifiability problems (where the likelihood function is identical for two distinct parameter values) are rare. However, greater uncertainty about parameter estimates arise if constant transition intensities are not assumed. This is particularly the case if a semi-Markov model is chosen, because sojourn as opposed to transition times determine the likelihood, and these are harder to determine from discretely observed states. The mortality rates from states are

difficult to estimate. This is because the state occupied just before death is never observed. These problems become greater the more flexible the model being fitted. These issues are discussed further in chapter 6.

Despite the advances the methods presented in this chapter provide, there remain considerable gaps in the available methodology. The phase-type sojourn distributions approach in chapter 6 fills many of these gaps by allowing a relatively straightforward method of fitting (hidden) semi-Markov models for both progressive and bi-directional models.

# Chapter 6

# Semi-Markov models with phase-type sojourn distributions

This chapter details methodology for fitting semi-Markov and hidden semi-Markov models to interval censored or panel observed data, both for progressive and bi-directional models. This is achieved through the use of Coxian phase-type sojourn distributions. These allow the semi-Markov models to be expressed as a type of hidden Markov model, allowing relatively straightforward likelihood analysis.

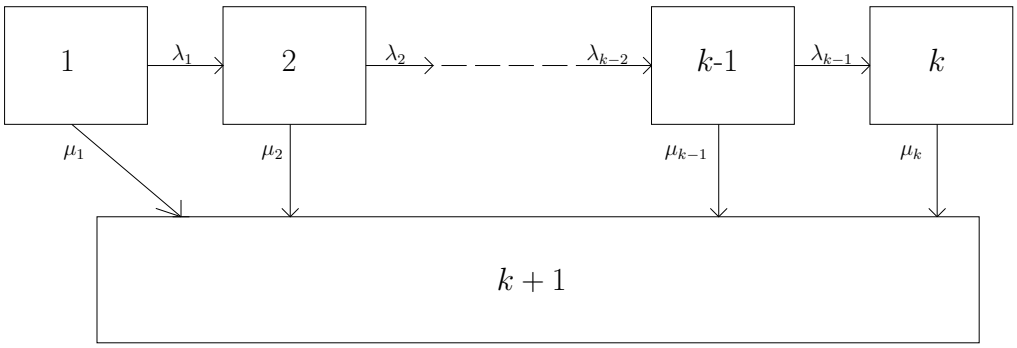
## 6.1 Phase-type distributions

This section gives a general introduction to phase-type distributions in contexts not directly related to multi-state modelling.

### Definition

A phase-type distribution describes the time to absorption of a finite-state Markov chain. A general continuous phase-type distribution is the distribution of the time to absorption of a $k+1$ state homogeneous Markov process, in which states $1, \ldots, k$ are transient and state $k+1$ is an absorbing state. There is an initial vector determining the probability of starting in each of the $k+1$ states. Any continuous distribution with non-negative support can be arbitrarily closely approximated by some phase-type distribution. The most basic phase-type distribution is simply the exponential distribution which describes a 2 state

Figure 6.1: General Coxian Phase-type distribution with $k$ phases.




process where state 1 is transient and state 2 is absorbing.

Although each intermediate transition may have constant hazards, time from initiation to absorption will, in general, not have constant hazards, except in the simplest 2 state case.

### 6.1.1 Coxian phase-type distribution

A Coxian phase-type distribution [32] is a special class of phase-type distributions. A $k$ phase Coxian phase-type distribution describes the time to absorption in a $k+1$ state homogeneous Markov process where progression from transient state $r$ is only possible to the absorbing state $k+1$ or to the adjacent state $r+1$ (figure 6.1). At time zero, the process is in state 1. The parameters of this distribution are $(\lambda_1, \ldots, \lambda_{k-1})$, the transition intensities between transient states, and $(\mu_1, \ldots, \mu_k)$, the transition intensities to the absorbing state.

Phase-type distributions and Coxian phase-type distributions in particular, have a wide range of applications in applied probability and statistics. Coxian phase-type distributions have been used to provide ways of fitting smooth curves to fully observed or right censored data, for instance on length of stay of hospital patients [46, 93], as an alternative both to restrictive parametric models and to non-parametric models.

In survival analysis, consideration of the shape of the hazard function, as discussed in section 2.2.2, is related to phase-type distributions. In particular, Aalen [2] has emphasised the importance of phase-type distributions in survival analysis.

A dichotomy exists between those analyses which attempt to give a physical meaning to the latent states (phases) of the distribution, and those analyses which simply use the

phases as a way of better describing the system as a whole. In this chapter, we generally take the second view.

## 6.2 Application of phase-type distributions to semi-Markov models

As discussed in section 1.2.3, semi-Markov models are difficult to fit to panel observed data primarily because closed form solutions for the transition probabilities between occupied states cannot be found. However, an exception to this exists when the semi-Markov model has sojourn time distributions which are phase-type distributions.

### 6.2.1 Review of uses of phase-type distribution in a multi-state setting

The use of phase-type distributions in the context of complicated stochastic processes is common in applied probability, having first been proposed by Cox in 1955 [31].

The central idea is to express a semi-Markov process by making it a function of an unobservable Markov process which has an expanded state space. This allows quantities such as the equilibrium distribution or the transition probabilities to be determined in a *matrix-analytic* form. The ability of phase-type distributions to provide an arbitrary close approximation to any non-negative distribution means that general semi-Markov processes, for instance with Weibull or Gamma sojourn distributions, can be approximated by semi-Markov processes with phase-type sojourn distributions. The transition probabilities for semi-Markov processes can be approximated in this way [88].

The use of phase-type sojourn distributions for semi-Markov models fitted to panel data is less common. Matis *et al* [96] fitted a 3-state bi-directional semi-Markov model to current status data on marked shrimp. The two transient states had Erlang distributions (Gamma distributions with integer shape). This was achieved by expanding the state space, such that in order to pass through observable state 1, a subject has to progress through $k$ latent states in sequence. The model was fitted using non-linear least squares estimation rather than maximum likelihood. As the data consisted of a single observation for each subject, only the transition probabilities were needed to fit the model.

Åhlström *et al* [7] used phase-type distributions to provide a general methodology for relapse clinical trials for recurrent diseases. When a patient was diagnosed as cured,

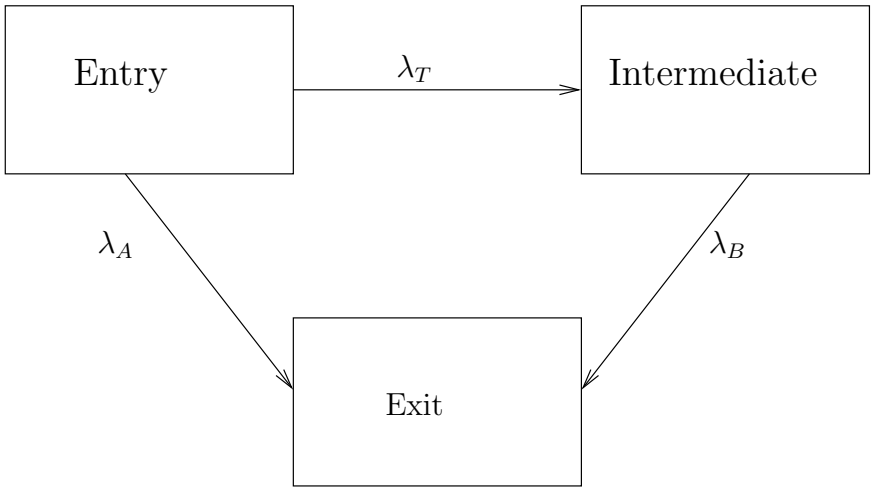
he/she was assigned a schedule of medical examination times, $t_1, \ldots, t_N$. The time to relapse $X$ was the quantity of interest, but the observation process was the realisation of a random variable $Y$, the time to symptoms. Either a patient was diagnosed as having relapsed (without the emergence of symptoms) at the $k$th examination time, in which case $X$ is interval censored between $t_{k-1}$ and $t_k$, or the patient could develop symptoms at a time $y$ between examination times $t_{k-1}$ and $t_k$, in which case $X$ is interval censored between $t_{k-1}$ and $y$. In order for data in the latter case to be used it is necessary to model the joint distribution of $(X, Y)$. The aim was to get an unbiased empirical estimate of the distribution of $X$. It was assumed that overall the process could be described by a Markov process whose state space is split into two groups of transient states and an absorbing state. The patient was healthy whilst in the first group of transient states, was in the pre-clinical state of the disease (relapse without symptoms) in the second group of states, and the absorbing state corresponded to the clinical state (relapse with symptoms). Thus the time of transition between the set of pre-clinical and set of clinical states is known up to interval censoring.

Crespi *et al* [36] modelled a two state recurrent process for episodes of genital herpes, for which the data were panel observed. The observed process $O(t)$, which took values 0 or 1, was assumed to be linked to an unobserved latent process $X(t)$, which took values $0, 1, 2, \ldots$, where $X(t)$ was a time-homogeneous birth-death process, with arrivals occurring with intensity $\lambda > 0$ and deaths occurring with intensity $\mu > 0$. When $X(t) = 0$, the observed process $O(t) = 0$, while when $X(t) > 0$, $O(t) = 1$. Hence, $O(t)$ described a semi-Markov process, since the sojourn time distribution for observable state 1 was non-exponential, but also a hidden Markov process, since $O(t)$ was a function of a Markov process. In this situation, standard methods for hidden Markov models can be applied in order to calculate both the transition probabilities, and the overall likelihood. It is this approach we develop in this chapter.

### 6.2.2 Simple illustrative example

As a basic illustrative example, suppose we have survival data in which subjects are either in state 1 (alive) or state 2 (dead). A possible parametric approach to modelling the time to death is to use Coxian phase-type distributions. For instance, suppose we choose to use a two-phase Coxian phase-type distribution (figure 6.2). In this setting we let the observable two-state (survival) process be explained by a three-state latent process. When

Figure 6.2: Two-phase Coxian Phase-type distribution




the subject is alive, he/she may be in one of two latent states, entry or intermediate.

The two phase Coxian phase-type distribution is characterised by three parameters: $\lambda_A, \lambda_B$ and $\lambda_T$. If $\lambda_A < \lambda_B$ then the hazard function for the distribution is increasing. If $\lambda_A > \lambda_B$ the hazard is decreasing. $\lambda_T$ determines the rate at which the hazard changes. The initial hazard is always given by $\lambda_A$. However, the limiting hazard, defined as the limit of the hazard function as $t \to \infty$, is bounded by $\lambda_B$, but only equals $\lambda_B$ when $\lambda_A + \lambda_T > \lambda_B$. This is due to general results about the quasi-stationary distribution of a Markov process as discussed in section 2.2.2.

The two-phase Coxian phase-type distribution therefore has a somewhat analogous parametrisation to the piecewise constant semi-Markov model with two zones. $\lambda_T$ can be thought of as corresponding broadly to the change point in the piecewise constant model. But the phase-type distribution has a continuous hazard function. Compared to the Weibull intensities considered in section 5.5.1, the two-phase Coxian distribution is more versatile, allowing a hazard which can start at any given level and reduce to any given level. The Weibull hazard is constrained to either begin at $\infty$ and decrease to zero if the shape parameter is less than 1, or begin at zero and increase to $\infty$ if it is greater than 1. Of course, an additional parameter is required for the two-phase distribution as it requires 3 rather than 2 parameters.

The likelihood for data under this phase-type model can be expressed as a hidden Markov model. Specifically, we say that the latent process, $X(t)$, is a homogeneous Markov process

with transition intensity matrix

$$Q = \begin{bmatrix} -\lambda_A - \lambda_T & \lambda_T & \lambda_A \\ 0 & -\lambda_B & \lambda_B \\ 0 & 0 & 0 \end{bmatrix}.$$

The observable process $O(t)$ is defined through the fixed misclassification probabilities

$$P(O(t) = s|X(t) = r) = e_{rs}$$

where $e_{rs}$ is the $(r, s)$ entry of a matrix

$$e = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Suppose instead that the two state model represents some other unidirectional process, e.g. the development of a chronic disease such that state 1 is disease free and state 2 is ill, but that the disease status cannot be observed with complete accuracy. Suppose also there is a probability $\alpha$ of being wrongly classified as ill, and $\beta$ the probability of being wrong classified as disease free. The same hidden Markov model framework as above can be used, but the matrix of misclassification probabilities becomes

$$e = \begin{bmatrix} 1-\alpha & \alpha \\ 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

and these misclassification probabilities must be estimated from the data.

### 6.2.3 General procedure for phase-type semi-Markov models

In a more general setting the same principle of describing a semi-Markov process as a hidden-Markov process applies.

Consider a semi-Markov process $X(t)$ with state space $\mathcal{S} = \{1, \ldots, R\}$, where $R$ is an absorbing state. To maintain clarity, we will assume there is one absorbing state. However, only small modifications are required for the case of no absorbing state or multiple absorbing states. Let $\mathcal{S}^*$ be the state space for a latent Markov process, $X^*(t)$. Let the sojourn distributions of each state of $X(t)$ be $k$-phase Coxian phase-type distributions, as defined in section 6.1.1 and depicted in figure 6.1, with parameters $\lambda_1^{(r)}, \ldots, \lambda_{k-1}^{(r)}$ and $\mu_1^{(r)}, \ldots, \mu_k^{(r)}$, for the $r$th state in $X(t)$.

**Latent process**

To describe $X(t)$ we let $\mathcal{S}^*$ consist of states $r_1, \ldots, r_k$ for each $r = 1, \ldots, R-1$, and state $R$, such that

$$\mathcal{S}^* = \{1_1, 1_2, \ldots, 1_k\} \cup \{2_1, 2_2, \ldots, 2_k\} \cup \ldots \cup \{(R-1)_1, \ldots, (R-1)_k\} \cup R$$

meaning it has dimension $k(R-1)+1$.

Note that it is not necessary for each observable state to have the same number of latent states.

There is the added complication of there being multiple destinations that the process can go to from an existing state. Additional parameters are needed to determine, given a subject leaves a state, what the probability of going to a particular state is. Let $\rho_{sj}^{(r)}$, denote the probability of going from observable state $r$ to observable state $s$, given that a subject leaves state $r$ from the $j$th phase of state $r$ (i.e. from latent state $r_j$), and the $\rho_{sj}^{(r)}$ satisfy

$$\sum_{s \neq r} \rho_{sj}^{(r)} = 1$$

for $r = 1, \ldots, R$.

The transition intensities for $X^*(t)$ are then as follows: the transition intensity between state $r_j$ and $s_1$ for $j = 1, \ldots, k$ and $r = 1, \ldots, R-1$, $s = 1, \ldots, R-1$, $r \neq s$ is given by

$$\rho_{sj}^{(r)} \mu_j^{(r)}.$$

Similarly the intensity between $r_j$ and $R$ is given by

$$\rho_{Rj}^{(r)} \mu_j^{(r)}.$$

In addition, for $r = 1, \ldots, R-1$, transitions from $r_j$ to $r_{j+1}$ for $j = 1, \ldots, k-1$ have a transition intensity given by $\lambda_j^{(r)}$. All other transition intensities are zero. This implies that in the model, a patient enters a new state $s$ in phase $s_1$. They then must pass through consecutive phases until the state is exited. Exiting the state can occur from any phase.

**Relation to observable process**

$X(t)$ relates to $X^*(t)$ in the following way: if $X^*(t) \in \{r_1, \ldots, r_k\}$ then $X(t) = r$ for $r = 1, \ldots, R$. Hence $X(t)$ has the same structure as a hidden Markov model with a

$(k(R-1)+1) \times R$ misclassification probability matrix, $e^*$,

$$e^* = \left[\begin{array}{cccccc} 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 & 0 \\ \hline 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \hline \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ \hline 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ \hline 0 & 0 & 0 & \dots & 0 & 1 \end{array}\right]. \tag{6.1}$$

**Extension for misclassification models**

A straightforward extension allows hidden semi-Markov models to be fitted. The extension is as follows: suppose the hidden semi-Markov process, has observed states $O(t)$ related to the states of the underlying semi-Markov process $X(t)$, by misclassification probability matrix $e$, such that $e_{rs} = P(O(t) = s|X(t) = r)$. $O(t)$ can be related to $X^*(t)$ as a hidden Markov model with an $(k(R-1)+1) \times R$ misclassification probability matrix $e^*$ where

$$e^* = \left[\begin{array}{cccccc}
e_{11} & e_{12} & e_{13} & \cdots & e_{1(R-1)} & e_{1R} \\
e_{11} & e_{12} & e_{13} & \cdots & e_{1(R-1)} & e_{1R} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\
e_{11} & e_{12} & e_{13} & \cdots & e_{1(R-1)} & e_{1R} \\
\hline
e_{21} & e_{22} & e_{23} & \cdots & e_{2(R-1)} & e_{2R} \\
e_{21} & e_{22} & e_{23} & \cdots & e_{2(R-1)} & e_{2R} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\
e_{21} & e_{22} & e_{23} & \cdots & e_{2(R-1)} & e_{2R} \\
\hline
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\
\hline
e_{(R-1)1} & e_{(R-1)2} & e_{(R-1)3} & \cdots & e_{(R-1)(R-1)} & e_{(R-1)R} \\
e_{(R-1)1} & e_{(R-1)2} & e_{(R-1)3} & \cdots & e_{(R-1)(R-1)} & e_{(R-1)R} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\
e_{(R-1)1} & e_{(R-1)2} & e_{(R-1)3} & \cdots & e_{(R-1)(R-1)} & e_{(R-1)R} \\
\hline
e_{R1} & e_{R2} & e_{R3} & \cdots & e_{R(R-1)} & e_{RR}
\end{array}\right]. \tag{6.2}$$

The likelihood for an individual is derived in the same way as a HMM. Suppose we have a subject observed in states $O_1, \ldots, O_N$ at times $t_1, \ldots, t_N$. To calculate $L = \mathbb{P}(O_1, \ldots, O_N)$, we can sum over all the possible sequences of latent states $x_1^*, \ldots, x_N^*$ occupied. The Markov property of the underlying states and the conditional independence of $o_1, \ldots, o_N$ given $x_1^*, \ldots, x_N^*$ allows us to write

$$L = \sum_{X_1^*} \mathbb{P}(O_1|X_1^*)\mathbb{P}(X_1^*) \sum_{X_2^*} \mathbb{P}(O_2|X_2^*)\mathbb{P}(X_2^*|X_1^*) \ldots \sum_{X_N^*} \mathbb{P}(O_N|X_N^*)\mathbb{P}(X_N^*|X_{N-1}^*).$$

Since $\mathbb{P}(O_i = s|X_i^* = r) = e_{rs}^*$ and $\mathbb{P}(X_i^* = s|X_{i-1}^* = r) = p_{rs}(t_i - t_{i-1})$, we can define matrices $M_1 \ldots M_N$ where $M_i$ is a $k(R-1)+1 \times k(R-1)+1$ matrix with (r,s) entry

$$e_{s,O_i}^* p_{rs}(t_i - t_{i-1})$$

then the likelihood can be written as a matrix product

$$L = \pi_{\mathbf{0}} M_1 M_2 \ldots M_N \mathbf{1}$$

where $\pi_{\mathbf{0}}$ is the vector of initial state probabilities for the latent process $X^*(t)$ and $\mathbf{1}$ is a vector of ones of length $k(R-1)+1$.

The above formulation can be made even more general. For instance there is no requirement for there to be an absorbing state. Also, the number of phases in the Coxian phase-type distributions does not need to be equal across states.

### 6.2.4 Practical issues

A general semi-Markov model with phase-type sojourn times could have an arbitrarily large number of unknown parameters. Moreover, the time taken to compute the likelihood will be a function of the number of latent states. The need to limit both the number of parameters and the number of latent states motivates the use of Coxian phase-type distributions with a small number of phases. For many purposes a two-phase Coxian phase-type distribution will be adequate to provide a sufficiently flexible hazard function, either because it could be reasonably assumed that the hazard functions are monotonic or because there are insufficient data to allow more than two-phases.

Similarly, if one allows the transition probabilities to other observable states, $\rho_{sj}^{(r)}$, to vary between phases, the number of unknown parameters will also be very large. The simplest way of dealing with this is to constrain

$$\rho_{s1}^{(r)} = \rho_{s2}^{(r)} = \ldots = \rho_{sk}^{(r)}$$

so that the probability of making an $r \rightarrow s$ transition, given a subject leaves state $r$, is constant.

### 6.2.5 Further possible extensions

#### Unknown initiation times

Whilst in the transplantation data considered in this thesis, the transplantation time for a subject defines the initiation of the process, in other contexts the initiation time of the process may not be known. This is not problematic for Markov models where only the current state and current time are required, but it does have an effect on the likelihood calculations for semi-Markov models since the time of entry into the current state is needed. In the context of recurrent processes, most authors assume the process is in equilibrium [36, 109]. For models with absorbing states this is not appropriate because at equilibrium all individuals are in the absorbing state. Satten and Sternberg [116] in the context of non-parametric modelling of unidirectional processes assumed the unknown initiation times into a state were independent of subsequent transition times. They treat the time from initial observation until the first transition as separate nuisance functions to be estimated. This leads to some information loss because in reality the time between the first observation and the first transition is related to the sojourn time in the initially

observed state. However, this formulation means that methods applicable to unidirectional models when the times of initiation are known can be easily adapted. Kang and Lagakos [72], in a parametric setting, only dealt with the case of known initiation times, but suggested methods similar to Satten and Sternberg's might be applicable.

For phase-type semi-Markov models, the time since the initiation of the process is not needed if the underlying latent Markov state at the first observation time is known. One approach is therefore to assume that the probability of occupancy of latent state $r_j$ at the first observation time is given by $p_{r_j}$, the $r_j$th entry of $\mathbf{p}$, where $\mathbf{p}$ is a $k(R-1)$ vector summing to 1 and $k$ is the number of phases. The entries of $\mathbf{p}$ would need to be estimated from the data, leading to $k(R-1)-1$ additional parameters. If it is reasonable to assume the times from initiation to first observation are identically distributed, then this method is most suitable if the subjects are assumed to be homogeneous. If there are covariates affecting the progression of the semi-Markov process, we would instead expect there to be a correlation between the covariates and the vector of initial state occupancy probabilities. In theory, covariate dependencies on the initial state occupancy vector could be incorporated, though this would greatly increase the number of unknown parameters to be estimated.

**Inhomogeneous semi-Markov processes**

The semi-Markov model presented above is dependent on the time since entry into a state, but not on the overall calendar time. In some situations this may be restrictive. For instance in illness-death models the patient age is likely to have an effect on mortality regardless of disease status. This dependence means that a time inhomogeneous Markov model, in which age affects mortality, may be favoured over a homogeneous semi-Markov model, in which the mortality at onset of illness is fixed, regardless of the subject's age at that time. Time since onset may nevertheless have a significant influence on mortality. This motivates the use of a time inhomogeneous semi-Markov model, in which a subject's transition intensities depend both on the current time and the time since entry into the current state.

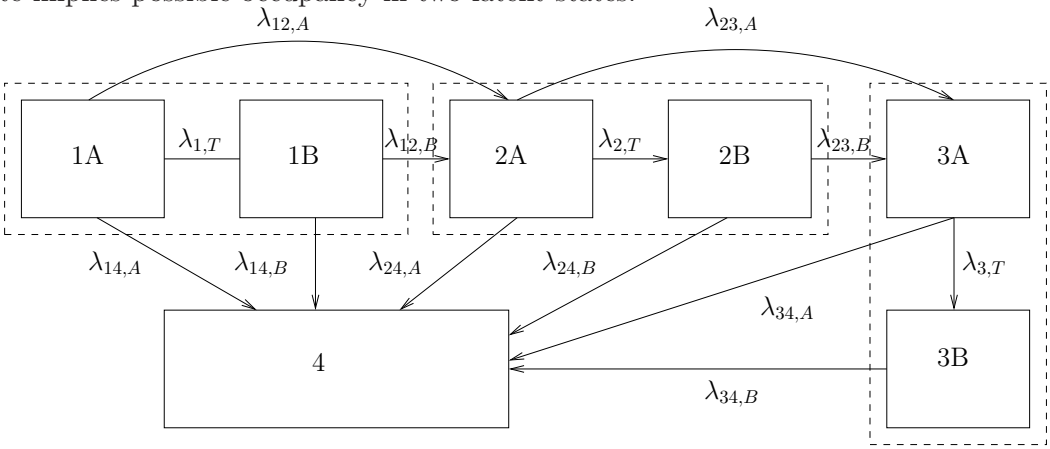
In the context of illness-death models time-inhomogeneous models have also been called 'excess mortality' models - the time spent ill governing the amount of additional hazard of death compared to if the subject was yet to develop the disease. Commenges *et al* [27] applied this type of model, using smooth penalised likelihood estimates, to data concerning

elderly patients at risk of developing dementia. In their application the excess mortality model fitted the data better than a homogeneous semi-Markov model, but not significantly better than a time inhomogeneous Markov model.

The framework of phase-type sojourn distributions can, in principle, be readily extended to allow for time dependence with respect to the initiation time of the process. Instead of a time homogeneous semi-Markov process being made through a time homogeneous HMM, we let it be a time inhomogeneous HMM and produce an inhomogeneous semi-Markov process. The methods of chapter 5 can be employed to do this, either through piecewise constant intensities or smooth intensities where transition probabilities can be calculated by numerically solving the differential equations. However, it may not be practically possible to fit such models to panel observed data in many cases because identifiability and estimability problems, already inherent for semi-Markov models (see section 6.4.1), will be even more pronounced.

## 6.3 Application to the CAV dataset

Figure 6.3: A phase-type semi-Markov model for the CAV data. Each observable transient state implies possible occupancy in two latent states.




The phase-type method can be easily applied to the CAV data. Using 2-phase Coxian phase-type distributions for each sojourn time gives a model as depicted in 6.3. This has 7 latent states. Whilst in latent states $1A$ or $1B$, the subject is in true state 1 (CAV free), whilst in latent states $2A$ or $2B$ the subject is in true state 2 (mild CAV) and so on. In addition, as with the analogous misclassification hidden Markov model, the observed state

may be one of the states adjacent to the true state. As discussed in section 6.2.4, for some parsimony, for states 1 and 2, we constrain the probability of death given exit from the state, to be constant. Thus we apply the constraints

$$\frac{\lambda_{14,A}}{\lambda_{12,A}+\lambda_{14,A}} = \frac{\lambda_{14,B}}{\lambda_{12,B}+\lambda_{14,B}}$$

and

$$\frac{\lambda_{24,A}}{\lambda_{23,A}+\lambda_{24,A}} = \frac{\lambda_{24,B}}{\lambda_{23,B}+\lambda_{24,B}}.$$

This is achieved by reparametrising so that $\kappa_1$ and $\kappa_2$ define

$$\lambda_{12,B} = \kappa_1\lambda_{12,A}, \lambda_{14,B} = \kappa_1\lambda_{14,A}$$

and

$$\lambda_{23,B} = \kappa_2\lambda_{23,A}, \lambda_{24,B} = \kappa_2\lambda_{24,A}.$$

Hence $\kappa_1$ and $\kappa_2$ determine the degree of time dependency in states 1 and 2 respectively. A value of $\kappa_r < 1$ implies that the transition intensities from state $r$ are decreasing with time since entry into state $r$, if $\kappa_r > 1$ then the transition intensities are increasing.

A similar restriction is unnecessary for state 3 transitions as it is only possible to enter state 4 from there. Despite these restrictions, there are problems with estimating all these parameters. As with the semi-Markov models in chapter 5, it appears state 2 does not have significant time dependency, having a point estimate for $\kappa_2$ very close to 1. We therefore again assume an exponential distribution for the sojourn in state 2. This reduces the number of latent states in the model to 6.

We assume the covariates donor age and IHD primary diagnosis, only affect CAV onset rates. Specifically we let

$$\lambda_{12,j} = \lambda_{12,j}^{(0)} \exp\left(\beta_{12}^{(\mathrm{IHD})} \times \mathrm{IHD} + \beta_{12}^{(\mathrm{dage})} \times \mathrm{dage}\right)$$

for $j = A, B$.

The maximum likelihood estimate for this model gives $-2 \times \mathrm{LL} = 3915.4$ from 15 parameters. This represents a significant improvement compared to the time homogeneous hidden Markov model ($-2 \times \mathrm{LL} = 3933.6$ from 11 parameters). However, on the basis of AIC, the time inhomogeneous piecewise-constant HMM of section 5.2.1 is preferred ($-2 \times \mathrm{LL} = 3913.6$ from 14 parameters).

Parameter estimates are presented in table 6.1. There are problems obtaining the standard errors for the phase-type model because the maximum likelihood estimate occurs at the boundary of the parameter space. Specifically, the estimated transition intensity from state 2 to state 4 is 0. $\kappa_1$ and $\kappa_3$ define the degree of time dependency in states 1 and 3 respectively. Values below 1 signify a decreasing hazard whilst values above 1 imply an increasing hazard. $\kappa_1 = 4.801$ suggesting an increasing hazard, whilst $\kappa_3 = 0.086$ suggesting a strongly decreasing hazard. The confidence intervals are conditional on $\lambda_{24} = 0$ fixed. Asymptotic results for obtaining confidence intervals only apply conditional on the estimate being in the interior of the parameter space. The stated confidence intervals will tend to be narrower than they should be. This is particularly clear for $\kappa_3$. The given confidence interval is (0.022, 0.342), which suggests strong evidence of time dependency, being very far from 1 with Wald $p < 0.001$ for this parameter equalling 1. However, a significant part of the hazard soon after entry into the state arises from deaths directly from state 2 being transferred via state 3. By fitting a similar model, but with no time dependencies in either state 2 or state 3, so that it is a time inhomogeneous Markov model with time dependence in the first state only, we get $-2 \times LL = 3921.82$. On this basis, a better approximation of the significance of the time dependency in state 3 can be obtained. The likelihood ratio test gives $p = 0.04$, so there is evidence of time dependency but it is weak.

The covariate effects of IHD and donor age on disease onset rates remain largely unaffected compared with the time homogeneous hidden Markov model. There is a slight increase in the magnitude of effects, but it is not possible to make direct comparisons because the parameter now has a different influence on the transition intensities.

The point estimates imply that instead of facing any risk of death in state 2, a subject instead proceeds to state 3 and is then exposed to a high hazard of death on entry into state 3. To a lesser extent the same thing was seen with the piecewise constant intensities model and the Weibull model in which the $2 \rightarrow 4$ transition intensity is significantly lower for these models than in the Markov models. Primarily, the boundary estimate is a problem specific to the CAV dataset. In particular, we note that the shortest interval between a state 2 observation and death is 94 days, compared with 24 days and 5 days for deaths after state 1 and state 3 observations. There is little empirical evidence of direct $2 \rightarrow 4$ transitions occurring. The phase-type model has greater support for all patients who enter state 2 passing through state 3 before death.

Note that a non-parametric bootstrap, sampling the subjects with replacement, would

Table 6.1: Parameter estimates for the phase-type sojourn distribution hidden semi-Markov model on the CAV data

| Parameter | Estimate | 95% CI |
|---|---|---|
| $\lambda_{12,A}$ | 0.015 | (0.012,0.018) |
| $\lambda_{14,A}$ | 0.012 | (0.007,0.021) |
| $\lambda_{1,T}$ | 0.090 | (0.038,0.217) |
| $\kappa_1$ | 4.801 | (1.282,17.985) |
| $\lambda_{23}$ | 0.241 | (0.194,0.299) |
| $\lambda_{24}$ | 0 | |
| $\lambda_{34,A}$ | 1.618 | (0.419,6.252) |
| $\lambda_{3,T}$ | 4.095 | (0.937,17.897) |
| $\kappa_3$ | 0.086 | (0.022,0.342) |
| $e_{12}$ | 0.025 | (0.015,0.042) |
| $e_{21}$ | 0.172 | (0.112,0.255) |
| $e_{23}$ | 0.069 | (0.041,0.069) |
| $e_{32}$ | 0.092 | (0.046,0.178) |
| $\beta_{12}^{(\text{IHD})}$ | 0.598 | (0.260,0.935) |
| $\beta_{12}^{(\text{dage})}$ | 0.033 | (0.017,0.050) |
| $-2\times$ LL | 3915.4 | |

not be helpful in determining confidence intervals without assuming $\lambda_{24} = 0$. In all the bootstrap sample datasets, the minimum time between a state 2 observation and death would be greater than or equal to the 94 days observed in the original dataset.

## 6.4 Identifiability and Estimability for semi-Markov models

While highly flexible models are useful to account for departures from the straightforward Markov model, in practice, they may be too complex to estimate from a given dataset or even unidentifiable. In this section issues of parameter identifiability and estimability will be explored in the context of semi-Markov models applied to panel observed data.

### 6.4.1 Identifiability

Parameter identifiability is a relatively weak condition on the likelihood function and parametrisation. For likelihood function $l(.)$ of parameter $\theta \in \Theta$ and data $x \in \mathcal{X}$, in order for identifiability to be violated it is necessary that $\exists \theta_1, \theta_2$ with $\theta_1 \neq \theta_2$ such that $\forall x \in \mathcal{X}$,

$$l(\theta_1; x) = l(\theta_2; x).$$

#### Trivial identifiability issues

If we apply the Coxian Phase-type distribution framework for the semi-Markov model, as in section 6.2, then there are some obvious identifiability issues. Suppose a state has a sojourn time distribution given by a $k$-phase Coxian phase-type distribution and the transition intensities to the exit state are given by $\mu_1, \ldots, \mu_k$ in phases $1, \ldots, k$ respectively, and call $\lambda_1, \ldots, \lambda_{k-1}$ the transition intensities between consecutive phases. Consider the case where $\mu = \mu_1 = \ldots = \mu_k$. In this situation the hazard function for the phase-type distribution is $h(t) = \mu$ since the hazard of absorption is $\mu$ regardless of the phase occupied. Therefore the sojourn time distribution is just $Exp(\mu)$ and does not depend on $\lambda_1, \ldots, \lambda_{k-1}$.

Hence the likelihood function of any dataset is independent of the values of $\lambda$'s. When the differences between the $\mu$'s are small but non-zero the $\lambda$'s are hard to estimate, but at the same time they have little impact. If we are interested in the mean or median sojourn time in a particular state, when the differences between the $\mu$'s are small, the standard

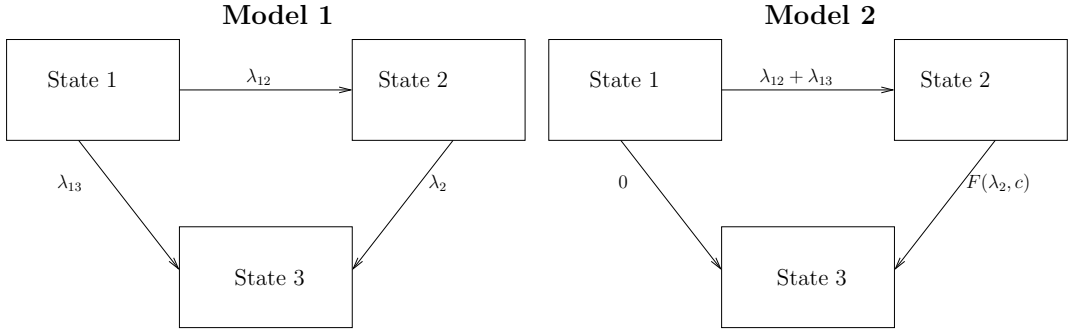



Figure 6.4: Two possible models for three state disease model: Model 1 is time homogeneous Markov. Model 2 is semi-Markov in state 2.

error on the $\lambda$'s will be large, but this will not translate to much extra uncertainty about the mean or median sojourn time.

**Non-trivial identifiability issues**

However, more significant identifiability issues exist. Suppose panel data are observed from a three state illness-death model. Specifically, we assume there is some minimum time interval between consecutive observations, $\epsilon > 0$. Let model 1 be a time homogeneous Markov model where the intensities are constant with values $\lambda_{12}, \lambda_{13}$ and $\lambda_2$. Let model 2 allow state 2 to have an arbitrary sojourn distribution and therefore be semi-Markov. For model 2, we make the $1 \rightarrow 2$ transition intensity be equal to $\lambda_{12} + \lambda_{13}$ and set the intensity between 1 and 3 to zero. In addition we make the sojourn distribution in state 2 equal to $F(\lambda_2, c)$ which describes a distribution that has a point mass of $c = \frac{\lambda_{13}}{(\lambda_{12}+\lambda_{13})}$ at $t = 0$, and is then proportional to an Exponential distribution with parameter $\lambda_2$ for $t > 0$. The layout of these models is given in figure 6.4.

Since the minimum time interval between observations is $\epsilon$, in model 1, it will not be possible to determine whether entry into state 3 occurred from state 1 or from state 2. With such a sampling scheme, the likelihood functions for the two models will be identical conditional on any data $x$. Both models have the same sojourn distribution in state 1. In model 1, given that a subject leaves state 1, there is a

$$\frac{\lambda_{13}}{(\lambda_{12} + \lambda_{13})}$$

probability they enter state 3 directly. In model 2, any patient who leaves state 1, enters

state 2, but faces an immediate

$$\frac{\lambda_{13}}{(\lambda_{12}+\lambda_{13})}$$

probability of entering state 3 from state 2 at time 0.

Model 2 is somewhat pathological as it includes a sojourn distribution for state 2 with a discontinuity at zero. However, the flexibility of Coxian phase type distributions are such that it is possible to get arbitrarily close to such a distribution. For instance, take a 2-phase Coxian phase-type distribution. Let $\lambda_A = K\lambda_{13}, \lambda_T = K\lambda_{12}, \lambda_B = \lambda_2$. The time to the absorbing state has a cumulative density function given by

$$F(t) = \left(\frac{\lambda_{13}}{\lambda_1} + \frac{\lambda_{12}}{\lambda_1(\lambda_2 - K\lambda_1)}\right)(1 - \exp(-K\lambda_1 t)) - \frac{K\lambda_{12}}{(\lambda_2 - K\lambda_1)}(1 - \exp(-\lambda_2 t))$$

where $\lambda_1 = \lambda_{12} + \lambda_{13}$. Now

$$\lim_{K\to\infty} F(t) = \frac{\lambda_{13}}{(\lambda_{12}+\lambda_{13})} + \frac{\lambda_{12}}{(\lambda_{12}+\lambda_{13})}(1 - \exp(-\lambda_2 t))$$

this corresponds precisely with $F(\lambda_2, c)$ as defined above.

More generally, we can also express model 1 as being a semi-Markov model where state 1 involves competing exponential hazards

$$\lambda_{12A} = \lambda_{12} + \tau, \lambda_{13A} = \lambda_{13} - \tau$$

for $0 \leq \tau \leq \min(\lambda_{12}, \lambda_{13})$ and let the state 2 sojourn distribution be given by

$$F(\lambda_2, \frac{\tau}{\lambda_{12}+\lambda_{13}}).$$

This class of models includes all semi-Markov models in which the probability of death at time $0^+$ in state 2 plus the probability of death from state 1, equals the probability of death from state 1 in the original model. Thus there is a class of models which cannot be distinguished through panel observation. To our knowledge this class of models has not be previously considered, we term them *observably-Markov*.

### 6.4.2 Estimability

It is clear from section 6.4.1 that there will be problems identifying some of the parameters when the process is close to being Markov. It is less clear to what extent these problems, particularly in estimating the probability of dying from a particular state, persist when the true process is semi-Markov.

Similar methods to section 4.2.1 in chapter 4, involving computation of the expected likelihood function, can be used to investigate the estimability of the parameters for panel observed data from a phase-type semi-Markov or hidden semi-Markov model. In particular we are interested in the shape of the profile likelihood with respect to the probability of death from an intermediate state. We use a simple example for illustration. Suppose we have panel data from a 3-state illness-death process in which all patients are initiated in state 1 at time 0 and recovery from illness is not permitted within the model. Moreover, let subjects be observed at yearly intervals up to a maximum of 10 years.

Two sampling cases are considered. In the first case, all transitions are interval censored. In the second case, entry times into state 3 are known exactly. For the first case, there are 66 patterns of observed states, which have probabilities, $p(\theta_0)$, defined by the true underlying model parameters $\theta_0$. The expected likelihood can then be found by

$$\mathbb{E}_{\theta_0} \log L(\theta) = \sum_{i=1}^{66} p_i(\theta_0) \log \left(L_i(\theta)\right).$$

In the second case, calculation of the expected likelihood is more complicated as it requires integration over possible times of deaths. Details of the calculations are given in Appendix D.

Three scenarios of underlying models are considered.

1. Underlying process time inhomogeneous Markov. State 1 has a two-phase Coxian phase-type distribution. Sojourn time in state 2 is Exponential.

2. Underlying process semi-Markov. State 1 has a two-phase Coxian phase-type distribution. State 2 has a two-phase Coxian phase-type distribution with a weakly decreasing hazard.

3. Underlying process semi-Markov. State 1 has a two-phase Coxian phase-type distribution. State 2 has a two-phase Coxian phase-type distribution with a strongly decreasing hazard.

Parameter values are shown in table 6.2. These were chosen to allow around 60% of subjects to reach state 3 within the follow-up time and to roughly mimic the state progression seen in the CAV dataset.

In particular, the probability of dying directly from state 1 given exit from state 1 is $\rho = \frac{9}{40}$ in both cases. This is roughly what was estimated in the original time homogeneous

Table 6.2: Parameter values for 3-state phase-type model

| Parameter | Case 1 | Case 2 | Case 3 |
|---|---|---|---|
| $\lambda_{12A}$ | 0.2 | 0.2 | 0.2 |
| $\lambda_{12B}$ | 0.15 | 0.15 | 0.15 |
| $\lambda_{1T}$ | 0.3 | 0.3 | 0.3 |
| $\rho$ | 0.225 | 0.225 | 0.225 |
| $\lambda_{23A}$ | 0.15 | 0.2 | 0.2 |
| $\lambda_{23B}$ | NA | 0.1 | 0.05 |
| $\lambda_{2T}$ | NA | 0.15 | 0.5 |

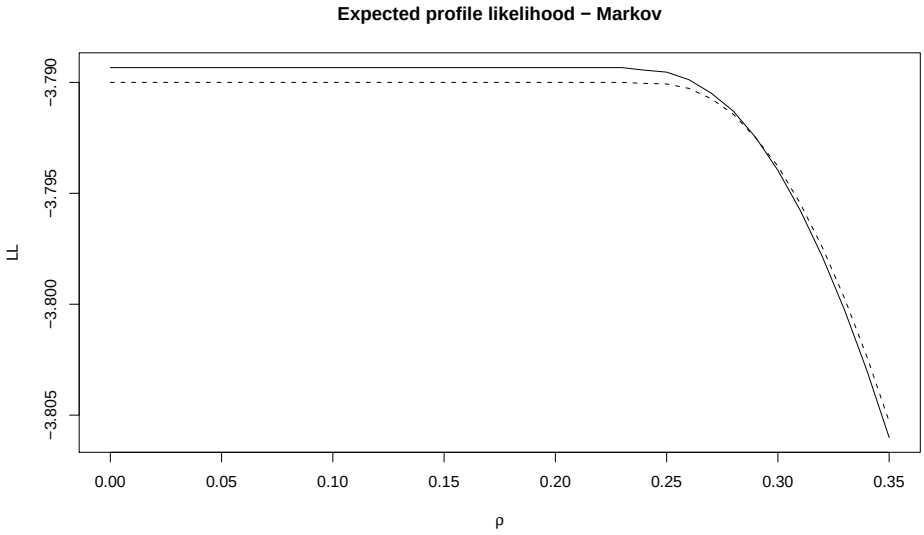
Markov model for the CAV data for the $2 \rightarrow 4$ transition intensity. In all cases a phase-type semi-Markov model is fitted to the data.

Figure 6.5 gives the expected profile likelihood of an individual when the true model is Markov, taking the parameter of interest to be $\rho$, the probability of death from state 1. Here, the lack of identifiability discussed in section 6.4.1 is evident. The profile likelihood is flat between 0 and the true value of $\rho$. This is because the same model can be expressed by a range of different *observably-Markov* models. This pattern is the same regardless of whether death times are interval censored or known exactly. As expected, the gradient of the profile likelihood for $\rho > \frac{9}{40}$ is greater when exact death times are known, reflecting the additional information available in the data.

When the underlying process is semi-Markov, the expected profile likelihood has a distinct maximum at the true value of $\rho$ (figure 6.6). Hence for large enough sample size, we can expect convergence of the maximum likelihood estimate to the true parameter values. However, the expected profile likelihood is flat near to $\rho = 0$. This flat area corresponds to the interval for which the optimal estimate conditional on a fixed $\rho$ gives equivalent *observably-Markov* models. The size of the region of flat likelihood depends on how close the semi-Markov model is to being Markov and the precise sampling scheme. In both cases, knowing the exact death times improves the situation considerably. There is more power to detect the non-exponential state 2 sojourn times and hence the optimum is more pronounced and the interval of flat profile likelihood shorter.

These results suggest that for datasets with a large number of subjects and exact death times, provided the true model is semi-Markov, maximum likelihood estimates should

Figure 6.5: Expected profile likelihood when underlying process is Markov. Bold line = exact death times, dashed line = panel observed

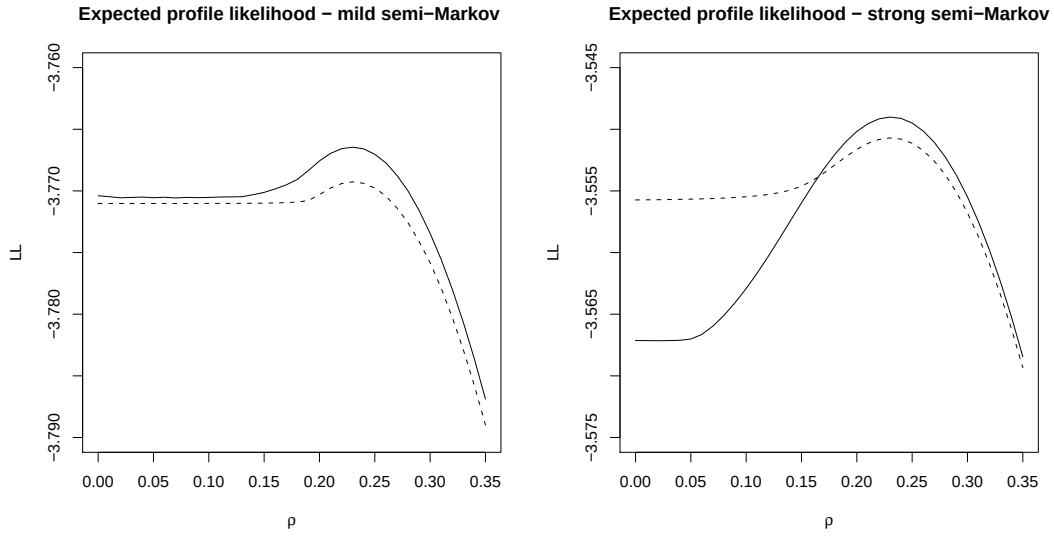



be consistent. However, some care may be needed to ensure numerical algorithms for optimisation have obtained the global optimum.

### 6.4.3 CAV dataset

In the CAV dataset, an estimate of $\rho = 0$ for state 2 was seen since the estimate of $\lambda_{24}$ in table 6.1 is 0. This cannot directly be due to the identifiability problems discussed in this section because the estimate of $\lambda_{3T}$ was only moderately large - so the estimated parameters give a process that is genuinely semi-Markov. However, except when the semi-Markov effect is strong and exact death times are known, the difference in the expected profile likelihood between $\rho = 0$ and the true value of $\rho$ was shown in the previous section to be small. This suggests that, for moderate sample sizes, only a small number of events would be required to allow $\rho = 0$ to be preferred. For the CAV data, there are two main factors which seem to have caused the boundary estimate. Firstly, there were no $2 \rightarrow 4$ transitions observed in short intervals. Secondly, in the piecewise-constant time inhomogeneous HMM, a decreasing transition intensity between 2 and 3 was observed. The semi-Markov model fitted does not allow any time dependency in state 2, but it does allow it in state 3. The mildly significant decreasing hazard in state 3 is possibly

Figure 6.6: Expected profile likelihood when state 2 is semi-Markov. Bold lines = exact death times, dashed lines = panel observed




a statistical artifact due to a decreasing hazard in state 3 giving a similar result to a decreasing transition intensity in state 2, particularly if the $2 \rightarrow 4$ transition intensity is set to zero. In general, boundary estimates can occur quite often when fitting these models.

## 6.5 Conclusion

### Application as a goodness-of-fit diagnostic

Phase-type sojourn distributions allow semi-Markov and hidden semi-Markov models to be fitted to panel observed multi-state data with relative ease. The likelihood for such models can be expressed as a particular type of HMM. Standard methods for fitting HMMs (section 1.2.4) can therefore be used. Fitting these alternative models provides a specific test of a time homogeneous Markov against semi-Markov assumption in the form of a likelihood ratio test, without the need to specify times of changing hazards as in piecewise-constant models. These tests will be more powerful than the general goodness-of-fit test presented in chapter 3 at detecting lack-of-fit related to semi-Markov intensities.

For the CAV data there was a likelihood ratio statistic of 18.2 on 4 degrees of freedom between the time homogeneous Markov model and the hidden semi-Markov model. This yields a p-value of 0.001. However, no time dependency was detected for state 2 and moreover the time dependency in state 3 was only marginally significant (p=0.04). It is most likely that this is an artifact of the significant decreasing $2 \rightarrow 3$ transition intensity found in the piecewise-constant HMM fitted in chapter 5.

### General application

Phase-type semi-Markov models may also have a wider applicability as models to more accurately describe the process of interest. The primary advantage of the phase-type approach is its general flexibility, which is comparable to what can be achieved using piecewise-constant intensities but without the need to choose the locations of change-points. Similarly unlike the piecewise-constant intensities approach, the resulting intensities in the phase-type model are smooth and continuous. The other distinct advantage of the approach is the ease with which the models can be fitted and the range of models that can be fitted at least in theory. Unlike the methods of chapter 5, where only progressive models could be fitted and misclassification of observed states presented considerable difficulties, the phase-type approach allows models with reverse transitions to be fitted and state misclassification requires only a small change to the likelihood. As the likelihood can be expressed as the likelihood for a HMM, only a small modification would be required to fit such models in existing software packages for HMMs such as **msm**. It is already possible to fit them in the most general case where there are no restrictions on $\rho_{sj}^{(r)}$ parameters.

However, a general problem with panel observed data from semi-Markov models is the lack of information in the data. In particular, the transition intensities at times soon after entry into a state, may be difficult to estimate because of the lack of consecutive observations in short time intervals. An objective of fitting a multi-state model to data with a survival context may be to estimate the hazard of death from occupancy within each state. As discussed in section 2.2.2, typically in a disease model where the states represent levels of disease, it is usually expected that higher disease states will give higher instantaneous risk of death. In the context of time homogeneous Markov models, the hazard of death is constant within each state and so jumping to a higher state results in a fixed increment to the hazard of death. For a time inhomogeneous Markov model,

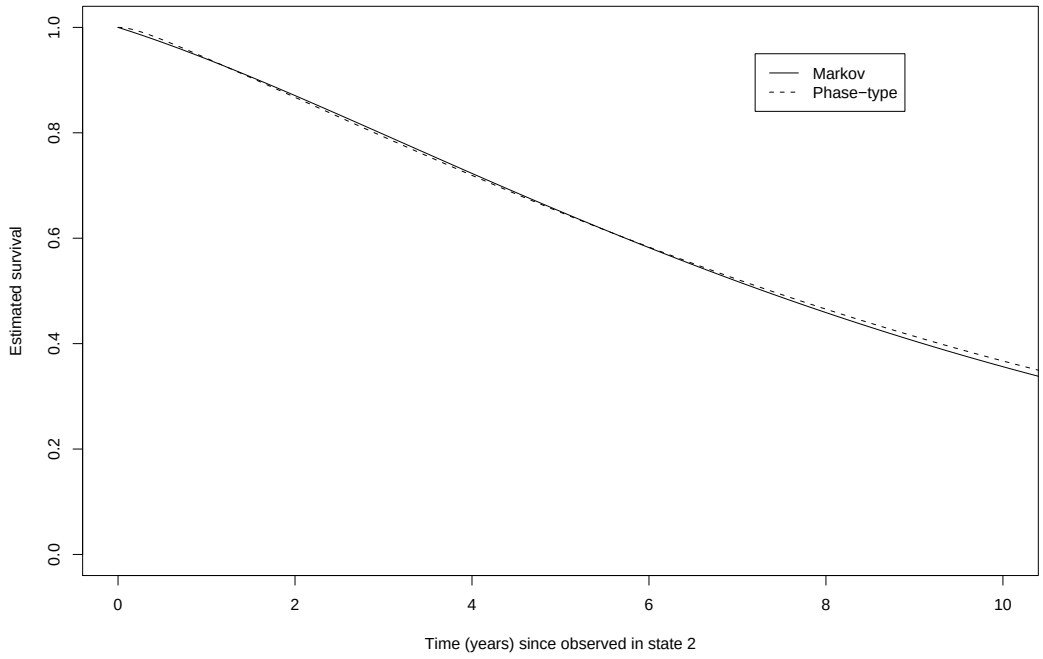
the situation is similar, except the increment in the hazard will depend on the time at which the jump occurred. For semi-Markov models however, we can get counter-intuitive changes in hazard. For instance, a semi-Markov model with Weibull intensities results in an initial hazard on entry into the state of 0 if the shape parameter $\alpha < 1$, regardless of the rate parameter. Similarly, Kang and Lagakos [72] required a *guarantee time* in each state, meaning the transition intensity is zero for a fixed time in a state.

These problems can arise in phase-type models. For the CAV data, the maximum likelihood estimate found that the transition intensity to death from state 2 (mild CAV) was 0. This presents two problems, both that the m.l.e. is difficult to interpret and that, because the m.l.e. is a boundary solution, standard asymptotic results cannot be applied to get estimates of the standard errors. Section 6.4.2 showed that these problems are likely to occur more generally in panel data with an absorbing state. The effect of state occupancy on mortality is often of interest, yet phase-type models are unable to give reliable estimates of this. However, the model can still give reasonable estimated survival given entry into a particular state at time 0. Figure 6.7 compares the estimated survival for the time homogeneous Markov and the phase-type semi-Markov models for the CAV data. As can be seen, the lack of risk of death whilst in state 2 according to the phase-type model only results in a short period where the survival is significantly higher than estimated in the Markov model.

More interpretable estimates for survival type multi-state models from the phase-type method could be obtained by further restricting the model. For instance, we could restrict the semi-Markov aspects of the model to only govern progression between transient states. For the CAV application this would mean the rate of progression from mild to severe CAV might depend on the time since CAV onset, but the transition intensities to death are fixed as constant whilst a subject is in a particular state. This in itself is somewhat restrictive. For the CAV data, the time dependency in state 3 was significant according to a likelihood ratio test. A further extension therefore is to allow the transition intensities to depend on time since the initiation of the process, but not on time since entry into a particular state. The resultant model is formally an inhomogeneous semi-Markov model but, due to the restrictions, it has parameters that can more easily be estimated than the general semi-Markov model. In particular, zero transition intensities to death are much less likely to be estimated.

Such a model does not fit better than the time inhomogeneous HMM fitted in chapter 5 for the CAV data because there is no evidence that state 2 is semi-Markov, whereas the

Figure 6.7: Estimated survival given observation in state 2 at time 0 for time homogeneous HMM and phase-type hidden semi-Markov model




model in section 5.2.1 did suggest some time dependency with respect to transplantation time, for state 2. This framework may however be useful for data which exhibit significant semi-Markov effects in intermediate states.

Overall, the flexibility and ease of implementation of the phase-type method makes it useful, particularly as a diagnostic to test Markov assumptions. The method also has potential for providing better models to describe the process of interest, particularly if there is some expectation that a semi-Markov model may be more appropriate for the specific application. Problems with parameter estimability may arise, meaning restrictions on the model may be needed. However these estimation issues are inherent in semi-Markov models on panel data and not confined to phase-type models.

# Chapter 7

# Final overview and discussion

This thesis investigated and developed diagnostics for testing model fit in multi-state models from panel observed data. The primary original contribution has been to draw together all existing approaches to model assessment into one work. The principal methodological developments are a general goodness-of-fit test which allows formal assessment of fit for models with misclassification as well as exact death times, and the development of phase-type sojourn time semi-Markov and hidden semi-Markov models, which allow the Markov assumption to be relaxed and are also relatively straightforward to fit. In addition, the use of adaptive ODE solvers to the Kolmogorov forward equations provides a reliable and effective way of fitting inhomogeneous Markov models and HMMs. The thesis also advanced the knowledge of the illustrative datasets. In particular, the onset rate for CAV was shown to increase with time since transplant.

## 7.1 Conclusions

The literature review in chapter 1 showed that existing methods for assessing goodness-of-fit are not well developed. Chapter 2 mainly focused on less formal diagnostics of model fit. The comparison of the overall empirical survival function with the fitted survival function from the multi-state model was explored. In the presence of covariates these comparisons become more difficult. Comparison with estimated survival functions from a Cox proportional hazards model are not recommended as they will not necessarily coincide, even if the Markov model is correctly specified. The related method of comparing observed and expected prevalence counts requires the observed counts to be estimated by some form

of interpolation. Apparent lack of fit may be due to a poor choice of times at which the observed and expected counts are to be calculated and the form of interpolation in the context of the observation scheme. A graphical generalisation of prevalence counts is developed which removes the need to choose particular times to calculate the counts. A further generalisation for misclassification HMMs is also given.

The tracking model of Satten [117] is extended to the case of any progressive model and the presence of exact death times in the data and is used to provide a likelihood ratio test of simple patient heterogeneity. However, it is also shown that a process with tracking closely resembles that of a time inhomogeneous Markov process with decreasing transition intensities. Chapter 2 also demonstrates that the graphical plots proposed by Bureau *et al* [14] are useful in assessing the fit of misclassification HMMs. In contrast, the prediction of future observations method of Satten and Longini [115] is shown to be less effective, particularly when the main departures are related to lack of conditional independence of the observed states. A formal test of this conditional independence assumption is proposed, based on a likelihood ratio test.

Chapter 3 extended previous work by Aguirre-Hernández and Farewell [6] to provide a Pearson-type goodness-of-fit test for Markov models. Firstly the form of the null distribution of the AH/F test statistic is explored. It is shown that a much better asymptotic approximation to the null distribution than $\chi^2$ can be found which means bootstrapping is not required for an accurate p-value in cases where the complexity of the model and the sample size make bootstrapping for the null distribution very time consuming. The AH/F test is also extended to the case of misclassification HMMs. AH/F's test is shown to be inappropriate when applied to models on data with exact death times. However, a modified test statistic, which involves modelling the sampling distribution in order to impute times at which subjects would have next been observed had they not died, is developed.

Chapter 4 investigated some of the effects of model misspecification of multi-state models. The results of the chapter give an indication of the pattern and degree of bias of estimates in some simple examples. The complexity of the models considered make it hard to obtain general conclusions. However, the methods of the chapter provide a general way of assessing the effect of model misspecification in particular cases.

Chapters 5 and 6 considered methods for fitting more complicated time dependent models. In chapter 5 the concept of piecewise-constant intensities, a well-established method of applying time inhomogeneous Markov models, is applied to semi-Markov models. Whilst

such a method is shown to be feasible in simple cases, computing and maximising the likelihood is significantly more difficult than in the Markov case. Chapter 5 also considers time dependent models with smooth intensities. Firstly, the use of numerical solutions to the Kolmogorov forward equations is explored. This is shown to be effective at allowing time inhomogeneous Markov models to be fitted, as long as an adaptive method to solving the differential equations is used rather than the Euler method. The use of a Monte Carlo Expectation-Maximisation algorithm is also considered. For time-inhomogeneous Markov models this is an inferior method compared to direct numerical solution of the Kolmogorov forward equations. However the MCEM algorithm approach also allows progressive semi-Markov models to be fitted. Here the MCEM algorithm is preferable to the alternative of direct numerical integration. The methods for semi-Markov models in chapter 5, particularly in the presence of misclassification, are limited in scope and quite slow to implement. In contrast, chapter 6 develops an approach to fitting semi-Markov and hidden semi-Markov models with phase-type sojourn distributions. The likelihood for these models can be specified as a type of hidden Markov model, hence calculating the likelihood is relatively straightforward. However, there can be parameter estimation problems when fitting these models to panel observed data.

Two motivating datasets were used through the thesis. The misclassification HMM applied to the BOS dataset of post lung transplantation patients was shown to fit poorly by various different criteria in chapter 2. The main problem was departures from the assumption of conditional independence of the observed states conditional on the true states. The models applied to the CAV dataset of post heart transplantation patients were in the main shown to be more appropriate. The methods of chapter 2 identified few problems with the fit of a time homogeneous model, except some small departures from independent misclassification in the Bureau *et al* plots. However, the Pearson-type goodness-of-fit test of chapter 3 showed that the fit of a time homogeneous model was quite poor. In particular, there were higher than expected $1 \rightarrow 2$ and $1 \rightarrow 3$ transitions in short intervals, perhaps indicating some form of time inhomogeneity. This was verified in chapter 5 where a piecewise-constant time inhomogeneous Markov model represented a significant improvement in likelihood. Other time inhomogeneous Markov and semi-Markov models also proved to have more support than the time homogeneous Markov model. However, any time dependency in state 2 is with respect to time since transplant rather than time from entry into the state.

## 7.2 Areas of further work

### 7.2.1 Empirical estimates

In chapter 2, comparisons of the fitted parametric Markov model with non-parametric empirical estimates, either of overall survival using Kaplan-Meier estimates, or of estimated prevalence in each state using prevalence counts, were used. Comparison with empirical survival works well when patients can be assumed to be homogeneous. However, in the presence of covariates, the correct semi-parametric estimate is difficult to obtain. One basic solution is to apply a Cox proportional hazards model to the overall survival curve. However, if the covariates have different effects on different transition intensities, then the overall hazard function of the fitted Markov model will not have the proportional hazards property. The Markov model is not embedded within the Cox proportional hazards model space, thus it is quite possible to have disagreement between the survival curves, even when the Markov model is correct. The current methodology for assessing fit using prevalence counts, uses very crude empirical estimators for the observed prevalence which, unless subjects are observed frequently, can be considerably biased.

A partial solution to both these problems would be methodology for non-parametric and semi-parametric estimation of panel observed multi-state data. It seems unlikely this will be possible except under a Markov assumption. In that context the work of Frydman [49, 50, 51] and Gaüzère [53] on illness-death Markov models with interval censored transition times using self-consistent estimators may be extendible to the case of progressive panel observed models. These non or semi-parametric estimates could then provide both consistent empirical survival estimates in the presence of covariates, and better prevalence estimates. As a goodness-of-fit diagnostic however, making the Markov assumption implies comparing a parametric Markov model (e.g. homogeneous), with a non-parametric or semi-parametric inhomogeneous Markov model. The performance of these estimators when the true model is non-Markov is less clear.

### 7.2.2 General goodness-of-fit test

Chapter 3 developed methods for formally assessing general goodness-of-fit in Markov and hidden Markov models. When all transition times, including death, are interval censored the Aguirre-Hernández/Farewell test, or the modified version for hidden Markov models, can be applied. Section 3.3 gave a way of getting a better asymptotic approximation to

the null distribution than $\chi^2$ which means it is not necessary to use bootstrapping to determine the null distribution unless cell counts in the contingency table are small.

When there are exact death times, the modified test can be applied. However, the form of the null distribution is different in this case and depends on the sampling distribution $f(t)$ and the proportion of observations that are exact deaths. An upper bound for the 95% point of the null distribution is given by the 95% point of $\chi^2_{C-|\theta|}$, where $C$ is the number of independent cells in the constructed contingency table. Similarly, the mean of the null distribution is known to lie between $(C - |\theta|)$ and $C$. Hence if the value of the statistic $T$ surpasses $\chi^2_{C-|\theta|}(0.95)$, it can be assumed that the test indicates that the fit is poor. Similarly, if $T < (C - |\theta|)$, it can be assumed that the test accepts the null hypothesis. However if $(C - |\theta|) < T < \chi^2_{C-|\theta|}(0.95)$, it is necessary to bootstrap to determine a p-value for $T$. For Markov or particularly hidden Markov models, with large datasets and a large number of parameters, fitting the model once may take a non-trivial time, e.g. 10 minutes. This causes the necessary bootstrapping to take an unacceptable amount of time, e.g. over a week for 1000 samples.

A possible solution is to consider other types of goodness-of-fit test which can ensure a known asymptotic null distribution. The information matrix test proposed by White [134] is one such test. This is a general test of model specification that exploits the familiar identity that when the model is correctly specified,

$$\mathbb{E}I(\theta) = \mathbb{E}(U(\theta)U(\theta)^T)$$

where $U$ is the score function and $I$ the Fisher information, but that, as shown in Appendix C and used in chapter 4, this identity is not true otherwise. The general form of the test therefore considers whether

$$D = \frac{1}{n}\sum_{i}^{n}\left(\frac{\partial l_i}{\partial \theta}\right)\left(\frac{\partial l_i}{\partial \theta}\right)^T - \frac{1}{n}\frac{\partial^2 l}{\partial \theta^2} \tag{7.1}$$

is significantly different from a matrix of zeroes, where $l_i(\theta)$ is the log-likelihood contribution for the $i$th individual in a sample of $n$ and $l(\theta)$ is the overall log-likelihood. A consistent estimate of the covariance of $D$ can be found, requiring only the first two derivatives of the likelihood [82]. However, the number of distinct entries in the Fisher information matrix increases quadratically with the number of unknown parameters. Similarly, some entries of the sample product of scores and Fisher information matrices can be highly dependent. In practice therefore, it is usually necessary to base the statistic on a linear combination of a subset of the entries from $D$.

In biostatistics, information matrix tests have been proposed as goodness-of-fit tests for Cox proportional hazard models [87], more general proportional hazards models [37], logistic regression models based on case-control data [137] and in binomial or beta-binomial models [16]. It remains to be seen whether this approach can be applied to multi-state models and whether the power of such a test is comparable to a Pearson-type test.

### 7.2.3 Models for time dependent misclassification

In chapter 2, the misclassification HMM model for the BOS dataset was shown to be a very poor fit, particularly because the assumption of independent observed states conditional on the true underlying states does not hold. Instead the observed state was more likely to follow the previous observed state. In chapter 4, such model misspecification was shown to cause significant bias in estimates of the mean sojourn time in specific states. In some cases, it may be beneficial to consider alternative models, which allow state misclassification without the assumption that the misclassification is independent conditional on true states. In chapter 2, a very basic model where the current observed state depends both on the current true state and the previous observed state was fitted to test the assumption of independent misclassification. This model is unrealistic, particularly if the observation times are not regularly spaced. A better model would need to take into account that the level of dependence on the previous observation should be lower if the observation was longer ago.

Methods related to the phase-type approach could be used to provide such a model. For each true state $r$, $K$ latent states could be introduced, $r_1, \ldots, r_K$. Occupancy in a particular one of these $K$ states would not affect mortality rates or rate of entry into another true state $s$, but instead affects the misclassification probabilities for the observed states. For instance being in $r_1$ might imply a higher probability of being misclassified to the lower state $r-1$, whilst being in $r_k$ might imply a low probability of being misclassified to a lower state but a higher probability of being misclassified to state $r+1$. A further simplification is to assume movement between the $K$ latent states is independent of movement between true states.

There may be situations where it is not reasonable to assume independent misclassification, but otherwise the usual HMM assumptions hold. Here a time dependent misclassification model may provide both less biased estimates with more realistic standard errors, than a HMM would provide. For the BOS dataset however, the degree of time dependency is very

severe and there are also problems with the assumption of exponential underlying states. A better analysis of BOS may require a model which deals with the raw $FEV_1$ counts rather than a discretised state. A less rigid correlation structure between the observed $FEV_1$ and the true underlying disease would be needed.

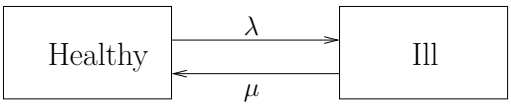
### 7.2.4 Phase-type semi-Markov models

As discussed in chapter 6, the full applied potential of phase-type semi-Markov models is not clear. Certainly in some cases there may be insufficient information in the data to allow all intensities to be semi-Markov. The development of models with some additional constraints would therefore be useful. A time inhomogeneous semi-Markov model, where transitions between transient states are semi-Markov, but mortality intensities are time-inhomogeneous Markov, is one possibility.

Phase-type models may be less desirable because the sojourn distributions do not have a standard parametric form. Similarly, parametric formulations like Gamma or Weibull sojourn distributions, require fewer parameters, though lack some flexibility as a result. Any non-negative distribution can be approximated arbitrarily closely by a phase-type distribution (of sufficient order). Phase-type distributions can therefore be used to approximate the transition probabilities in Weibull or Gamma semi-Markov models [88]. Potentially the methods of chapter 6 could be extended to allow Gamma or Weibull semi-Markov models to be fitted to close accuracy.

The general concept of using hidden Markov structures to create non-Markov observed processes could also be used to create more complicated non-Markov processes. For instance a recurrent disease process, in which a subject jumps between periods of health and periods of illness (figure 7.1), might be such that the hazard of relapse depends on the number of times a patient has already relapsed. An underlying Markov structure, in

Figure 7.1: Two-state recurrent disease model




which the latent process goes through a series of states

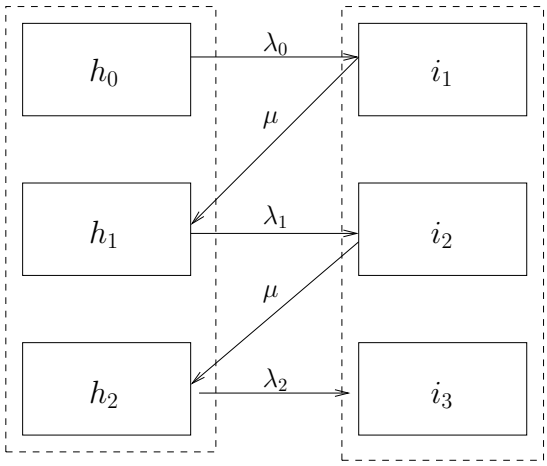
$$h_0, i_1, h_1, \ldots, i_{n-1}, h_{n-1}, i_n, \ldots,$$

where $h_j$ denotes the state representing healthy, having had $j$ periods of illness, and $i_j$ denotes the state representing the $j$th period of illness (figure 7.2). The subject is observed as healthy if in states $\{h_0, h_1, \dots\}$ and ill if in states $\{i_1, i_2, \dots\}$. A non-Markov observed process can be achieved if the transition intensities between states $h_j$ and $i_{j+1}$, $\lambda_j$, vary with $j$. For identifiability, it would either be necessary to constrain

$$\lambda_m = \lambda_{m+1} = \lambda_{m+2} = \dots$$

beyond some $m > 1$, or alternatively assume some form like $\lambda_{m+1} = q\lambda_m$. For practical computation purposes, it would be necessary to truncate the number of healthy and illness states at some point.

Figure 7.2: Latent recurrent disease model to allow non-Markov observable process




### 7.2.5 Development of software

Finally, many of the methods developed in this thesis have required specific programming to be implemented. The necessity of such programming would be a discouragement to other researchers. There is therefore a need for software to allow some of the methods to be routinely applied. The **msm** package [67] in R has made fitting a range of Markov and hidden Markov models relatively straightforward. This would be the natural software for inclusion of some of the methods of this thesis. The graphical generalisation of prevalence counts of chapter 2, the general goodness-of-fit test of chapter 3 and the phase-type semi-Markov methods of chapter 6 offer the most appropriate mix of usefulness and relative ease of implementation.

# Appendix A

# Impact of non-identical counts on $\chi^2$ statistic

**Lemma.** Let $\mathbf{X}_1, ..., \mathbf{X}_N$ be random variables with $\mathbf{X}_j \sim$ Multinomial$(1, \mathbf{p}_j)$ where $\mathbf{p}_1, ..., \mathbf{p}_N$ are known vectors of length $R$, s.t. $\sum_{r=1}^{R} p_{rj} = 1$ for all $j$ and let

$$T = \sum_{r}^{R} \frac{(\sum_j^N X_{rj} - \sum_j^N p_{rj})^2}{\sum_j^N p_{rj}}.$$

Then the limiting distribution of $T$ is not in general $\chi^2_{R-1}$.

*Proof.* Let $\mathbf{Y} = \sum_{j=1}^{N} \mathbf{X}_j$. Then $\mathbb{E}(\mathbf{Y}) = \sum_j^N \mathbf{p}_j$. The covariance matrix of $\mathbf{Y}$ is the $R \times R$ matrix:

$$\Sigma = \begin{bmatrix} \sum_j p_{1j}(1-p_{1j}) & -\sum_j p_{1j}p_{2j} & \dots & -\sum_j p_{1j}p_{Rj} \\ -\sum_j p_{1j}p_{2j} & \sum_j p_{2j}(1-p_{2j}) & \dots & -\sum_j p_{2j}p_{Rj} \\ \vdots & \vdots & \ddots & \vdots \\ -\sum_j p_{1j}p_{Rj} & -\sum_j p_{2j}p_{Rj} & \dots & \sum_j p_{Rj}(1-p_{Rj}) \end{bmatrix}.$$

Since $\sum Y_j = N$, $\mathbf{Y}$ is entirely defined by $\mathbf{Y}^* = (Y_1, \dots, Y_{R-1})$. This has covariance matrix $\Sigma^*$, defined as the upper left $(R-1) \times (R-1)$ of $\Sigma$, and expectation $\mathbf{p}^* = (\sum_j^N p_{1j}, \dots, \sum_j^N p_{R-1,j})$.

$T$ can be expressed as

$$T = (\mathbf{Y}^* - \mathbf{p}^*)^T V (\mathbf{Y}^* - \mathbf{p}^*)$$

where

$$V = \begin{bmatrix} \frac{1}{\sum p_{1j}} + \frac{1}{\sum p_{Rj}} & \frac{1}{\sum p_{Rj}} & \cdots & \frac{1}{\sum p_{Rj}} \\ \frac{1}{\sum p_{Rj}} & \frac{1}{\sum p_{2j}} + \frac{1}{\sum p_{Rj}} & \cdots & \frac{1}{\sum p_{Rj}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sum p_{Rj}} & \frac{1}{\sum p_{Rj}} & \cdots & \frac{1}{\sum p_{(R-1),j}} + \frac{1}{\sum p_{Rj}} \end{bmatrix}.$$

Let $\mathbf{Z} = (V^{\frac{1}{2}})(\mathbf{Y}^* - \mathbf{p}^*)$, then by the Lindeberg-Feller central limit theorem

$$\mathbf{Z} \xrightarrow{d} \mathrm{N}_{R-1}(0, V^{\frac{1}{2}}\Sigma^* V^{\frac{1}{2}}).$$

Since $V$ and $\Sigma^*$ are both symmetric, $V^{\frac{1}{2}}\Sigma^* V^{\frac{1}{2}} = V\Sigma^*$.

$V\Sigma^*$ is the $(R-1) \times (R-1)$ matrix

$$\begin{bmatrix} 1 - \frac{\sum p_{1j}^2}{\sum p_{1j}} + \frac{\sum p_{1j}p_{Rj}}{\sum p_{Rj}} & \frac{\sum p_{1j}p_{Rj}}{\sum p_{Rj}} - \frac{\sum p_{1j}p_{2j}}{\sum p_{2j}} & \cdots & \frac{\sum p_{1j}p_{Rj}}{\sum p_{Rj}} - \frac{\sum p_{1j}p_{R-1,j}}{\sum p_{R-1,j}} \\ \frac{\sum p_{1j}p_{Rj}}{\sum p_{Rj}} - \frac{\sum p_{1j}p_{2j}}{\sum p_{2j}} & 1 - \frac{\sum p_{2j}^2}{\sum p_{2j}} + \frac{\sum p_{2j}p_{Rj}}{\sum p_{Rj}} & \cdots & \frac{\sum p_{2j}p_{Rj}}{\sum p_{Rj}} - \frac{\sum p_{2j}p_{R-1,j}}{\sum p_{R-1,j}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sum p_{1j}p_{Rj}}{\sum p_{Rj}} - \frac{\sum p_{1j}p_{R-1,j}}{\sum p_{R-1,j}} & \frac{\sum p_{2j}p_{Rj}}{\sum p_{Rj}} - \frac{\sum p_{2j}p_{R-1,j}}{\sum p_{R-1,j}} & \cdots & 1 - \frac{\sum p_{R-1,j}^2}{\sum p_{R-1,j}} + \frac{\sum p_{R-1,j}p_{Rj}}{\sum p_{Rj}} \end{bmatrix}.$$

This is equal to the identity matrix only when $p_{r1} = \ldots = p_{rN}$ for $r = 1, \ldots, R$. Hence, since $T = \mathbf{Z^T Z}$, the limiting distribution of $T$ is the scalar product of an $R-1$ dimensional multivariate normal distribution, and not $\chi^2_{R-1}$. □

# Appendix B

# Derivation of the asymptotic null distribution of AH/F

**Theorem**: The asymptotic null distribution of AH/F, conditional on the true parameter values, the sampling times and the total group counts, can be expressed as a scalar product of a multivariate normal distribution with zero mean vector and some known covariance matrix.

*Proof.* The proof of this theorem comes in a series of steps. Firstly we need to establish a more general form of the statistic.

Suppose we have panel observed data assumed to come from a Markov model. Suppose each observation is arbitrarily categorised into category $c = 1, \dots, C$. Since each observation from a panel observed Markov model can be considered multinomial, we have that each observation $i = 1, \dots, n_c$ within category $c$ is non-identical multinomial, such that

$$\mathbf{x_{c,i}} \sim \text{Multinomial}(1, (p_1(z_{c,i}, \theta), \dots, p_R(z_{c,i}, \theta)))$$

where $z_{c,i}$ is the covariate vector corresponding to that observation (which we allow to include both the last observed state and the time between observations) and $\theta$ the parameter vector of length $M$. We can therefore write the AH/F statistic as

$$T(\mathbf{x}, \hat{\theta}) = \sum_{c=1}^{C} \sum_{r}^{R} \frac{(o_{rc} - e_{rc}(\hat{\theta}))^2}{e_{rc}(\hat{\theta})} \tag{B.1}$$

where $o_{rc} = \sum_i \mathbb{1}\{\mathbf{x_{c,i}} = \delta_r\}$ and $e_{rc}(\hat{\theta}) = \sum_i p_r(z_{c,i}, \hat{\theta})$, where $\delta_r$ is a vector of length $R$ with $r$th entry 1 and all other entries zero.

The second step of the proof is to establish the correlation between the observed counts vector, $\mathbf{O} = \{o_{rc} : r = 1, \dots, R, c = 1, \dots, C\}$ and the maximum likelihood estimate based on the complete data, $\hat{\theta}$. For a standard chi-squared test on multinomial data, $\mathbf{O}$ is a sufficient statistic for $\theta$ and so $\hat{\theta}$ is a deterministic function of $\mathbf{O}$. However, this isn't the case when $\mathbf{O}$ are not multinomial.

**Proposition 1**: Asymptotically

$$Cov(U_m(\theta), o_{rc}) = \sum_{i \in I_c} \frac{\partial p_r(z_{ci}, \theta)}{\partial \theta_m} \tag{B.2}$$

for $m = 1, \dots M$, $c = 1, \dots, C$ and $r = 1, \dots, R$, where $I_c$ is the set of observations in the $c$th category and $U(\theta) = \frac{\partial l(\theta)}{\partial \theta}$ is the score function.

**Proof of Proposition 1**: First we note standard asymptotic results,

$$\hat{\theta} \xrightarrow{d} \mathrm{N}(\theta, (\mathbb{E}I(\theta))^{-1}). \tag{B.3}$$

Also $\hat{\theta}$ satisfies the score equation $U(\hat{\theta}) = 0$. Taylor expansion of the score function about $\theta$ gives

$$0 = U(\hat{\theta}) = U(\theta) - (\hat{\theta} - \theta)^T I(\theta) + o_p(1)$$

where $I(\theta)$ is the observed Fisher information and $N$ the total number of observations and $o_p(1)$ denotes a remainder term which converges in probability to zero. Since $I(\theta) \xrightarrow{p} \mathbb{E}I(\theta)$, we may replace $I(\theta)$ by $\mathbb{E}I(\theta)$ to give

$$\hat{\theta} - \theta \xrightarrow{d} U(\theta)(\mathbb{E}I(\theta))^{-1},$$

where

$$\mathbb{E}I(\theta) = -\mathbb{E}\left(\frac{\partial^2 l(\theta)}{\partial \theta^T \partial \theta}\right).$$

For each category $c$, we have that

$$\sqrt{n_c}(\mathbf{o_c} - \mathbf{e}_c(\theta)) \xrightarrow{d} \mathrm{N}(0, \Sigma_c)$$

as $n_c \to \infty$, where

$$(\Sigma_c)_{rs} = \begin{cases} \sum_{i \in I_c} p_r(z_{c,i}, \theta)(1 - p_r(z_{c,i}, \theta)) & r = s \\ \sum_{i \in I_c} p_r(z_{c,i}, \theta) p_s(z_{c,i}, \theta) & r \neq s \end{cases}$$

for $c = 1, \ldots, C$.

To proceed we need to consider the asymptotic limit of $Cov(U(\theta), \mathbf{O})$. Since both quantities are functions of the full data $\mathbf{x}$, we can condition on $\mathbf{x}$.

$$Cov(U(\theta), \mathbf{O}) = Cov\left[\mathbb{E}(U(\theta)|\mathbf{x}), \mathbb{E}(\mathbf{O}|\mathbf{x})\right] + \mathbb{E}\left[Cov(U(\theta), \mathbf{O}|\mathbf{x})\right] \tag{B.4}$$

The second term of the RHS of equation B.4 is zero because given $\mathbf{x}$, $U(\theta)$ and $\mathbf{O}$ are fully determined. Moreover

$$Cov(U(\theta), \mathbf{O}) = Cov(U(\theta)|\mathbf{x}, \mathbf{O}|\mathbf{x}),$$

where $Cov(U(\theta)|\mathbf{x}, \mathbf{O}|\mathbf{x})$ is the covariance between the random variables obtained from $U(\theta)$ and $\mathbf{O}$ after conditioning on a particular value of the full data $\mathbf{x}$ and the implied expectation is over possible values of $\mathbf{x}$. The $m$th component of $U(\theta)$ can be written as

$$U_m(\theta) = \sum_c \sum_{i \in I_c} \left( \frac{\partial p(z_{c,i}, \theta)}{\partial \theta_m} \frac{1}{p(z_{c,i}, \theta)} \right)^T \mathbf{x_{c,i}},$$

while the $(r, c)$ entry of $\mathbf{O}$ is just $\sum_{i \in I_c} x_{(c,i)(r)}$ where $x_{(c,i)(r)}$ denotes the $r$th entry of the vector $\mathbf{x_{c,i}}$. Each of the individual observations $\mathbf{x_{c,i}}$ are independent. Hence

$$Cov(U_m(\theta), o_{rc}) = \sum_{i \in I_c} Cov\left[ \left( \frac{\partial p(z_{c,i}, \theta)}{\partial \theta_m} \frac{1}{p(z_{c,i}, \theta)} \right)^T \mathbf{x_{c,i}}, \delta_r^T \mathbf{x_{c,i}} \right].$$

Further

$$\begin{aligned} Cov(U_m(\theta), o_{rc}) = & \sum_{i \in I_c} \mathbb{E}\left[ \left( \left( \frac{\partial p(z_{c,i}, \theta)}{\partial \theta_m} \frac{1}{p(z_{c,i}, \theta)} \right)^T \mathbf{x_{c,i}} \right) \left( \delta_r^T \mathbf{x_{c,i}} \right) \right] \\ & - \mathbb{E}\left[ \left( \frac{\partial p(z_{c,i}, \theta)}{\partial \theta_m} \frac{1}{p(z_{c,i}, \theta)} \right)^T \mathbf{x_{c,i}} \right] \mathbb{E}\left[ \delta_r^T \mathbf{x_{c,i}} \right]. \end{aligned} \tag{B.5}$$

The $i$th observation only contributes a non-zero value for both $U_m$ and $o_{rc}$ if the observation is in cell $r$ with probability $p_r(z_{c,i}, \theta)$. Moreover, the expected contribution to $U_m$ from observation $i$ is 0. Hence equation B.5 reduces to

$$Cov(U_m(\theta), o_{rc}) = \sum_{i \in I_c} \frac{\partial p_r(z_{c,i}, \theta)}{\partial \theta_m}.$$

Thus we have proved Proposition 1. Denote $\Psi = Cov(U(\theta), \mathbf{O})$.

**Proposition 2**: Asymptotically

$$T(\mathbf{x}, \hat{\theta}) = \upsilon^*(\hat{\theta})^{\mathbf{T}} \upsilon^*(\hat{\theta})$$

where $\upsilon^*(\hat{\theta})$ is a vector of dimension $RC$

$$\upsilon^*(\hat{\theta}) = \upsilon(\theta) + B(\hat{\theta} - \theta),$$

where $B$ is an $RC \times M$ matrix with $(R(c-1)+r, m)$ entry $\frac{\partial e_{rc}(\theta)}{\partial \theta_m} \frac{1}{e_{rc}(\theta)^{\frac{1}{2}}}$ and

$$\upsilon_{rc}(\theta) = \frac{(o_{rc} - e_{rc}(\theta))}{e_{rc}(\theta)^{\frac{1}{2}}}.$$

**Proof of Proposition 2**: From the definition of $T(\mathbf{x}, \hat{\theta})$ in equation B.1, it follows that $T(\mathbf{x}, \hat{\theta}) = \upsilon(\hat{\theta})^{\mathbf{T}} \upsilon(\hat{\theta})$. Since $\hat{\theta} \xrightarrow{p} \theta$ we may Taylor expand $\upsilon(\hat{\theta})$ about $\theta$. This gives

$$\upsilon(\hat{\theta}) = \upsilon(\theta) + B(\hat{\theta} - \theta) + o_p(1), \tag{B.6}$$

with $B$ as defined above. Hence,

$$T(\mathbf{x}, \hat{\theta}) = \upsilon^*(\hat{\theta})^{\mathbf{T}} \upsilon^*(\hat{\theta}) + o_p(1)$$

as required.

**Proposition 3**: Asymptotically the combined $M + RC$ dimension vector $(\hat{\theta} - \theta, \upsilon(\theta))$ has mean vector zero and covariance matrix $A \Psi A^T$ where

$$A = \begin{bmatrix} \mathbb{E}I(\theta)^{-1} & 0 \\ 0 & P \end{bmatrix}$$

where $P$ is a $RC \times RC$ diagonal matrix with elements $(e_{rc}(\theta))^{-\frac{1}{2}}$.

**Proof of Proposition 3**: Given that $\Psi = Cov(U(\theta), \mathbf{O})$, we can note from the proof of proposition 1 that

$$\hat{\theta} - \theta \xrightarrow{d} U(\theta)(\mathbb{E}I(\theta))^{-1}.$$

Moreover, $\upsilon(\theta)$ is a linear function of $\mathbf{O}$. Hence, asymptotically $\xi = (\hat{\theta} - \theta, \upsilon(\theta))^T$ is a linear function of $\lambda = (U(\theta), \mathbf{O})^T$. This linear function has derivative given by the matrix $A$. Thus, by the delta method, asymptotically $\xi$ is multivariate normal with mean vector zero and covariance matrix $A \Psi A^T$. This concludes the proof of proposition 3.

Finally, let $W$ be a $(M + RC) \times RC$ matrix with

$$W = \begin{bmatrix} B & I \end{bmatrix},$$

where $I$ is a $RC \times RC$ identity matrix. Let $\kappa = W\xi$. Then from proposition 2,

$$T(\mathbf{x}, \hat{\theta}) \xrightarrow{d} \upsilon^*(\hat{\theta})^{\mathbf{T}} \upsilon^*(\hat{\theta})$$

as $N$ tends to infinity. Moreover by proposition 3, $\kappa$ is a linear function of $\lambda = (U(\theta), \mathbf{O})^T$. By the delta method, $\kappa$ is asymptotically MVN and so $T(\mathbf{x}, \hat{\theta})$ can be expressed as a scalar product of a MVN with mean vector 0 and, by proposition 1, covariance matrix $WA\Psi A^T W^T$. □

# Appendix C

# Derivation of the asymptotic distribution of a misspecified maximum likelihood estimator

Suppose that data are assumed to be from some probability model with parameters $\beta \in \mathcal{B}$, giving misspecified likelihood function $\tilde{l}(\beta; x)$, but that in fact they are from some other probability model with parameters $\alpha \in \mathcal{A}$.

Then there exists a value $\beta_\alpha$ that satisfies

$$\hat{\beta} \xrightarrow{p} \beta_\alpha \tag{C.1}$$

where $\hat{\beta}$ is the maximum likelihood estimate of $\beta$ under the misspecified model.

$\hat{\beta}$ solves the score equation

$$0 = \tilde{U}(\hat{\beta}) = \sum_i \frac{\partial \tilde{l}_i(\hat{\beta}; x)}{\partial \beta}.$$

Due to equation C.1, we can Taylor expand $\tilde{U}(\hat{\beta})$ about $\beta_\alpha$

$$0 = \tilde{U}(\hat{\beta}) = \tilde{U}(\beta_\alpha) + (\hat{\beta} - \beta_\alpha)^T \frac{\partial \tilde{U}}{\partial \beta}(\beta_\alpha) + o_p(1).$$

Hence, in the asymptotic limit

$$(\hat{\beta} - \beta_\alpha)^T = \tilde{U}(\beta_\alpha)\tilde{I}(\beta_\alpha)^{-1}$$

where

$$\tilde{I}(\beta_\alpha) = -\sum_i \frac{\partial^2 \tilde{l}_i(\hat{\beta}; x)}{\partial \beta^T \partial \beta}.$$

We can, for the purposes of an asymptotic expansion, replace $\tilde{I}(\beta_\alpha)$ with $\mathbb{E}_\alpha(\tilde{I}(\beta_\alpha))$. The asymptotic distribution of $\hat{\beta}$ is then multivariate normal with expectation $\beta_\alpha$ and covariance matrix

$$\Sigma_\alpha = \mathbb{E}_\alpha(\tilde{I}(\beta_\alpha))^{-1} V_\alpha \mathbb{E}_\alpha(\tilde{I}(\beta_\alpha))^{-1} \tag{C.2}$$

where

$$V_\alpha = \mathbb{E}_\alpha(\tilde{U}(\beta_\alpha)\tilde{U}^T(\beta_\alpha)).$$

Note that for a correctly specified model,

$$\mathbb{E}(I(\beta)) = \mathbb{E}(U(\beta)U^T(\beta)),$$

so that (C.2) reduces to $\mathbb{E}(I(\beta))^{-1}$.

# Appendix D

# Expected likelihood for a mixture of panel observed data and expected deaths

Suppose $R_i$ denotes that a patient died in the interval $(i, i+1)$. $T$ is the time of death, Then we can define

$$\tau_i(t, \theta_0) = \mathbb{P}(T = t|\theta_0, \mathcal{D}, T \in (i, i+1))$$

where $\theta_0$ is the true parameter vector and $\mathcal{D}$ is the set of observed states. Given $\theta_0$ and the observed states up to time $i$, we can obtain $\xi_i$, the state occupancy probabilities vector, by using equation 3.7 from chapter 3. We can write

$$\mathbb{P}(T = t|\theta_0, \mathcal{D}) = \sum_{j=1}^{R-1} \sum_{r \neq R} q_{rR} p_{jr}(t; \theta_0) \xi_{ij}$$

where $\xi_{ij}$ is the $j$th entry in $\xi_i$. Then

$$\tau_i(t, \theta_0) = \sum_{j=1}^{R-1} \sum_{r \neq R} \frac{q_{rR} p_{jr}(t; \theta_0) \xi_{ij}}{\sum_{j=1}^{R-1} \xi_{ij}(1 - p_{jR}(1; \theta_0))}$$

for $t \in (0, 1)$ is the conditional density for $T = i + t$ given response $\mathcal{D}$. To obtain the expected likelihood we need

$$\sum \rho_i(\theta_0) \int_0^1 \log\left(L_i(\theta|t)\right) \tau(t|\theta_0) dt$$

where $L_i(\theta|t)$ is the likelihood contribution for response $R_i$ with a death time $i + t$ given parameter $\theta$. This integral can be quickly computed numerically using numerical quadrature.

# Bibliography


[1] Aalen O.O, Johansen S. An empirical transition matrix for nonhomogeneous Markov chains beased on censored observations. *Scandinavian Journal of Statistics* 1978; **5**: 141-150.

[2] Aalen O. Phase type distributions in survival analysis. *Scandinavian Journal of Statistics* 1995; **22**: 447-463.

[3] Aalen O.O, Farewell V.T, De Angelis D, Day N.E, Gill O.N. A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. *Statistics in Medicine* 1997; **16**: 2191-2210.

[4] Aalen O.O, Gjessing H. K. Understanding the shape of the hazard rate: A process point of view. *Statistical Science* 2001; **16**: 1-22.

[5] Agresti A. *Categorical Data Analysis* 2nd edn. Wiley, 2002.

[6] Aguirre-Hernández R, Farewell V.T. A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models. *Statistics in Medicine* 2002; **21** :1899-1911.

[7] Åhlström L, Olsson M. A parametric estimation procedure for relapse time distributions. *Lifetime Data Analysis* 1999; **5** :113-132.

[8] Andersen P.K, Sommer Hansen L, Keiding N. Assessing the influence of reversible disease indicators on survival. *Statistics in Medicine* 1991; **10**: 1061-1067.

[9] Anderson G.L, Fleming T.R. Model misspecification in proportional hazards regression. *Biometrika* 1995; **82**: 527-541.

[10] Anderson T.W, Goodman L.A. Statistical inference about Markov chains. *Annals of Mathematical Statistics* 1957; **28**: 89-110.

[11] Anisimov V, Maas H.J, Danhof M, Della Pasqua O. Analysis of responses in migraine modelling using hidden Markov models. *Statistics in Medicine* 2007; **26**: 4163-4178.

[12] Baum L.E, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* 1970; **41**: 164-171.

[13] Booth J.G, Hobert J.P. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B* 1999; **61**: 265-285.

[14] Bureau A, Shiboski S, Hughes J.P. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine* 2003; **22**: 441-462.

[15] Caffo B.S, Jank W, Jones G.L. Ascent-based Monte Carlo expectation-maximization. *Journal of the Royal Statistical Society: Series B* 2005; **67**: 235-251.

[16] Capanu M, Presnell B. Misspecification tests for binomial and beta-binomial models. *Statistics in Medicine* 2007; *In press* DOI: 10.1002/sim.3049

[17] Chan K.S, Muñoz-Hernández. A generalized linear model for repeated ordered categorical response data. *Statistica Sinica* 2003; **13**: 207-226.

[18] Chen H-H, Duffy S.W, Tabar L. A mover-stayer mixture of Markov chain models for the assessment of dedifferentiation and tumour progression in breast cancer. *Journal of Applied Statistics* 1997; **24**: 265-278.

[19] Chen H-H, Yen M-F, Shiu M-N, Tung T-H, Wu H-M. Stochastic model for non-standard case-cohort design. *Statistics in Medicine* 2004; **23**: 633-647.

[20] Chen P-L, Bernard E.J, Sen P.K. A Markov chain model used in analyzing disease history applied to a stroke study. *Journal of Applied Statistics* 1999; **4**: 413-422.

[21] Chen P-L, Sen P.K. A piecewise transition model for analyzing multistate life history data. *Journal of Statistical Planning and Inference* 1999; **78**: 385-400.

[22] Chen P-L, Sen P.K. Markov chain model selection by misclassified model probabilities. *Communications in Statistics: Theory and Methods* 2007; **36**: 143-153.

[23] Chen P-L, Tien H-C. Semi-Markov models for multistate data analysis with periodic observations. *Communications in Statistics: Theory and Methods* 2004; **33**: 475-486.

[24] Chen Y, Xie J, Liu J.S. Stopping-time resampling for sequential Monte Carlo methods. *Journal of the Royal Statistical Society: Series B* 2005; **67**: 199-217.

[25] Commenges D. Multi-state models in epidemiology. *Lifetime Data Analysis* 1999; **5**: 315-327.

[26] Commenges D. Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research* 2002; **11**: 167-182.

[27] Commenges D, Joly P, Gégout-Petit A, Liquet B. Choice between semi-parametric estimators of Markov and non-Markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics* 2007; **34**: 33-52.

[28] Cook R.J. A mixed model for two-state Markov processes under panel observation. *Biometrics* 1999; **55**: 915-920.

[29] Cook R.J, Kalbfleisch J.D. A generalized mover-stayer model for panel data. *Biostatistics* 2002; **3**: 407-420.

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

[31] Cox D.R. The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proceedings of the Cambridge Philosophical Society* 1955; **51**: 33-41.

[32] Cox D.R. A use of complex probabilities in the theory of stochastic processes. *Proceedings of the Cambridge Philosophical Society* 1955; **51**: 313-319.

[33] Cox D.R. Tests of separate families of hypotheses. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability* 1961: 105-123. University of California Press, Berkeley, CA.

[34] Cox D.R, Miller H.D. *The theory of stochastic processes*. Chapman and Hall, London, 1965.

[35] Cox D.R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B* 1972; **34**: 187-220.

[36] Crespi C.M, Cumberland W.G, Blower S. A queuing model for chronic recurrent conditions under panel observation. *Biometrics* 2005; **61**: 193-198.

[37] Crouchley R, Pickles A. A specification test for univariate and multivariate proportional hazards models. *Biometrics* 1993; **49**: 1067-1076.

[38] Datta S, Sundaram R. Nonparametric estimation of stage occupation probabilities in a multistage model with current status data. *Biometrics* 2006; **62**: 829-837.

[39] Davies R.B. Algorithm AS 155: The distribution of a linear combination of $\chi^2$ random variables. *Journal of the Royal Statistical Society: Series C.* 1980; **29**: 323-333.

[40] De Gruttola V, Lagakos S.W. Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* 1989; **45**: 1-11.

[41] Deltour I, Richardson S, Le Hesran J-V. Stochastic Algorithms for Markov models estimation with intermittent missing data. *Biometrics* 1999; **55**: 565-573.

[42] Dempster A.P, Laird N.M, Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B.* 1977; **39**: 1-38.

[43] Duffy S.W, Chen H-H, Tabar L, Day N. Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Statistics in Medicine* 1995; **14**: 1531-1543.

[44] Efron B, Tibshirani R.J. *An Introduction to the Bootstrap* Chapman and Hall: New York, 1993.

[45] Faddy M.J. A note on the general time-dependent stochastic compartmental model. *Biometrics* 1976; **32**: 443-448.

[46] Faddy M.J, McClean S.I. Analysing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry* 1999; **15**: 311-317.

[47] Fisher R.A. The conditions under which $\chi^2$ measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society* 1924; **87**: 442-450.

[48] Foulkes A.S, De Gruttola V. Characterizing the progression of viral mutations over time. *Journal of the American Statistical Association* 2003; **98** 859-867.

[49] Frydman H. A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *Journal of the Royal Statistical Society: Series B.* 1992; **54**: 853-866.

[50] Frydman H. Nonparametric estimation of a Markov 'illness-death' process from interval-censored observations, with application to diabetes survival data. *Biometrika* 1995; **82**: 773-789.

[51] Frydman H, Szarek M. Nonparametric estimation in a Markov "illness-death" process from interval censored observations with missing intermediate transition status. *Department of Biostatistics, University of Copenhagen, Research Report*, 2007; 12.

[52] Gaüzère F, Commenges D, Barberger-Gateau P, Letenneur L, Dartigues JF. Maladie et dépendance: Description des évolutions par des modèles multi-états. *Population* 1999; **54**: 202-222

[53] Gaüzère F. *Approche non-paramétrique pour un modèle 3 états avec censures par intervalles - application à la dépendance.* PhD thesis, Université Victor Segalen Bordeaux 2, 2000.

[54] Gelfand A.E, Smith A.F.M. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**: 479-488.

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

[56] Genz A, Meyer M, Lumley T, Maechler M. *adapt: multidimensional numerical integration. R package version 1.0-3*

[57] Grüger J, Kay R, Schumacher M. The validity of inferences based on incomplete observations in disease state models. *Biometrics* 1991; **47**: 595-605.

[58] Hastings W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**: 97-109.

[59] Healy B, De Gruttola V. Hidden Markov models for settings with interval censored transition times and uncertain time origin: application to HIV genetic analyses. *Biostatistics* 2007; **8**: 438-452.

[60] Hollander M, Proschan F. Testing to determine the underlying distribution using randomly censored data. *Biometrics* 1979; **35**: 393-401.

[61] Hougaard P. Multi-state models: a review. *Lifetime Data Anaylsis* 1999; **5**: 239-264.

[62] Hsieh H-J, Chen T. H-H, Chang S-H. Assessing chronic disease progression using non-homogeneous exponential regression Markov models: an illustration using a selective breast cancer screening in Taiwan. *Statistics in Medicine* 2002; **21**: 3369-3382.

[63] Hwang W-T, Brookmeyer R. Design of panel studies for disease progression with multiple stages. *Lifetime Data Analysis* 2003; **9**: 261-274.

[64] Jackson C.H. *Statistical models for the latent progression of chronic disease using serial biomarkers*. PhD thesis, University of Cambridge 2002.

[65] Jackson C.H, Sharples L.D. Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine* 2002; **21**: 113-128

[66] Jackson C.H, Sharples L.D, Thompson S.G, Duffy S.W, Couto E. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D* 2003; **52**: 193-209.

[67] Jackson C.H. *msm: Multi-state Markov and hidden Markov models in continuous time. R package version 0.7.3* 2007.

[68] Joly P, Commenges D. A penalized likelihood approach for a progressive three-state model with censored and truncated data: application to AIDS. *Biometrics* 1999; **55**: 887-890.

[69] Joly P, Commenges D, Helmer C, Letenneur L. A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* 2002; 433-443.

[70] Kalbfleisch J.D, Lawless J.F. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* 1985; **80**: 863-871.

[71] Kang M, Lagakos S.W. Evaluating the role of human papillomavirus vaccine in cervical cancer prevention *Statistical Methods in Medical Research* 2004; **13**: 139-155.

[72] Kang M, Lagakos S.W. Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics* 2007; **8**: 252-264.

[73] Kay R. A Markov model for analysing cancer markers and diseases states in survival studies. *Biometrics* 1986; **42**: 855-865.

[74] Keats J.B, Nahar P.C, Korbel K.M. A study of the effects of mis-specification of the Weibull shape parameter on confidence bounds based on the Weibull-to-exponential transformation. *Quality and Reliability Engineering International* 2000; **16**: 27-31.

[75] Keilson, J. *Markov chain models: rarity and exponentiality.* New York, Springer, 1979.

[76] Kendall M.G, Stuart A. *The advanced theory of statistics. Vol.2* London, 1961.

[77] Klein J.P, Klotz J.H, Grever M.R. A biological marker model for predicting disease transitions. *Biometrics* 1984; **40**: 927-936.

[78] Klotz J.H, Sharples L.D. Estimation for a Markov heart transplant model. *Journal of the Royal Statistical Society: Series D* 1994; **43**: 431-438.

[79] Kosorok M.R, Chao W.H. The analysis of longitudinal ordinal response data in continuous time. *Journal of the American Statistical Association* 1996; **91**: 807-817

[80] Kosorok M.R, Chao W.H. Further details on the analysis of longitudinal ordinal response data in continuous time. *Technical Report, UW Madison* 1995.

[81] Lancaster T, Nickell S. The analysis of re-employment probabilities for the unemployed. *Journal of the Royal Statistical Society: Series A* 1980; **143**: 141-165.

[82] Lancaster T. The covariance matrix of the information matrix test. *Econometrica* 1984; **52**: 1051-1053.

[83] Lawless J.F. *Statistical Models and Methods for Lifetime Data.* Wiley, 1982.

[84] Lawless J.F, Yan, P. Some statistical methods for followup studies of disease with intermittent monitoring, In *Multiple comparisons, selection, and applications in Biometry*, F.M. Hoppe (ed), 427-446. New York, Marcel Dekker, 1993.

[85] Lee E.W, Kim M.Y. The analysis of correlated panel data using a continuous-time Markov model. *Biometrics* 1998; **54**: 1638-1644.

[86] Li Y-H, Klein J.P, Moeschberger M.L. Effects of model misspecification in estimating covariate effects in survival analysis for small sample sizes. *Computational Statistics and Data Analysis* 1996; **22**: 177-192.

[87] Lin D.Y, Wei L.J. Goodness-of-fit tests for the general Cox regression model. *Statistica Sinica* 1991; **1**: 1-17.

[88] Limnios N, Oprisan G. *Semi-Markov processes and reliability.* Birkhauser, 2001.

[89] Little R.J.A, Rubin D.B. *Statistical analysis with missing data.* J. Wiley and Sons, New York, 1987.

[90] Longini I.M, Clark W.S, Byers R.H, Ward J.W, Darrow W.W, Lemp G.F, Hethcote H.W. Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine* 1989; **8**: 831-843.

[91] Louis T.A. Finding the observed information matrix when using the EM algorithm. *Journal of Royal Statistical Society: Series B* 1982; **33**: 226-233.

[92] Lystig T.C, Hughes J.P. Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics* 2002; **11**: 678-689.

[93] Marshall A.H, Shaw B, McClean S.I. Estimating the costs for a group of geriatric patients using the Coxian phase-type distribution. *Statistics in Medicine* 2007; **26**: 2716-2729.

[94] Marshall G, Jones R.H. Multi-state models and diabetic retionpathy. *Statistics in Medicine* 1995; **14**: 1975-1983.

[95] Mathieu E, Loup P, Dellamonica P, Daurès J. P. Markov modelling of immunological and virological states in HIV-1 infected patients. *Biometrical Journal* 2005; **47**: 834-846.

[96] Matis J.H, Grant W.E, Miller T.H. A semi-Markov process model for migration of marine shrimp. *Ecological Modelling* 1992; **60**: 167-184.

[97] Müller H-G, Wang J-L. Hazard rate estimation under random censoring with varying kernels and bandwiths. *Biometrics* 1994; **50**: 61-76.

[98] Nelder, J.A, Mead R. A simplex method for function minimization. *The Computer Journal* 1965; **7** 308-313.

[99] Omar R.Z, Stallard N, Whitehead J. A parametric multistate model for the analysis of carcinogenicity experiments. *Lifetime Data Analysis* 1995; **1**: 327-346.

[100] Ocaña-Riola R. Two methods to estimate homogeneous Markov processes. *Journal of Modern Applied Statistical Methods* 2002; **1**: 131-138.

[101] Ocaña-Riola R. Non-homogeneous Markov processes for biomedical data analysis. *Biometrical Journal* 2005; **47**: 369-376.

[102] Pérez-Ocón R, Ruiz-Castro J.E, Gámiz-Pérez M.L. A multivariate model to measure the effect of treatments in survival to breast cancer. *Biometrical Journal* 1998; **40**: 703-715.

[103] Pérez-Ocón R, Ruiz-Castro J.E, Gámiz-Pérez M.L. Markov models with lognormal transition rates in the analysis of survival times. *TEST* 2000; **9**: 353-370.

[104] Pérez-Ocón R, Ruiz-Castro J.E, Gámiz-Pérez M.L. A piecewise Markov process for analysing survival from breast cancer in different risk groups. *Statistics in Medicine* 2001; **20**: 109-122.

[105] Pérez-Ocón R, Ruiz-Castro J.E, Gámiz-Pérez M.L. Non-homogeneous Markov models in the analysis of survival after breast cancer. *Journal of the Royal Statistical Society: Series C* 2001; **50**: 111-124.

[106] Pérez-Ocón R, Ruiz-Castro J.E. A multiple-absorbent Markov process in survival studies: application to breast cancer. *Biometrical Journal* 2003; **45**: 783-797.

[107] Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A* 1972; **135**: 185-207.

[108] Petzold, L.R. Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM Journal on Scientific and Statistical Computing* 1983; **4**: 136-148.

[109] Rosychuk R.J, Thompson M.E. A semi-Markov model for binary longitudinal responses subject to misclassification. *Canadian Journal of Statistics* 2001; **29**: 395-404.

[110] Rosychuk R.J, Thompson M.E. Bias correction of two-state latent Markov process parameter estimates under misclassification. *Statistics in Medicine* 2003; **22**: 2035-2055.

[111] Rosychuk R.J, Thompson M.E. Parameter identifiability issues in a latent Markov model for misclassified binary responses. *Journal of Iranian Statistical Society* 2004; **3**: 38-57.

[112] Rubin D.B. *Multiple Imputation for nonresponse in surveys*. J Wiley and Sons, New York, 1987.

[113] Ruiz-Castro J.E, Pérez-Ocón R. A semi-Markov model in biomedical studies. *Communications in Statistics: Theory and Methods* 2004; **33**: 437-455.

[114] Saint-Pierre P, Combescure C, Daurès J.P, Godard P. The analysis of asthma control under a Markov assumption with use of covariates. *Statistics in Medicine* 2003; **22**: 3755-3770.

[115] Satten G.A., Longini I.M. Markov chains with measurement error: estimating the 'true' course of a marker of the progression of Human Immunodeficiency Virus disease. *Journal of the Royal Statistical Society: Series C* 1996; **45**: 265-309.

[116] Satten G.A, Sternberg M.R. Fitting semi-Markov models to interval-censored data with unknown initiation times *Biometrics* 1999; **55**: 507-513.

[117] Satten G.A. Estimating the Extent of Tracking in Interval-Censored Chain-of-Events Data. *Biometrics* 1999; **55**: 1228-1231.

[118] Self S.G, Liang K-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 1987; **82**: 605-610.

[119] Setzer R.W. *odesolve: Solvers for ordinary differential equations. R package version 0.5-16* 2006.

[120] Shanno D.F. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation* 1970; **24**: 647-656.

[121] Shampine L.F, Gear C.W. A user's view of solving stiff ordinary differential equations. *SIAM Review* 1979; **21**: 1-17.

[122] Sharples L.D. Use of the Gibbs sampler to estimate transition rates between grades of coronary disease following cardiac transplantation. *Statistics in Medicine* 1993; **12**: 1155-1169.

[123] Sharples L.D, Taylor G.J, Faddy M. A piecewise-homogeneous Markov chain process of lung transplantation. *Journal of Epidemiology and Biostatistics* 2001; **6**: 349-355.

[124] Sharples L.D, Jackson C.H, Parameshwar J, Wallwork J, Large S.R. Diagnostic accuracy of coronary angiography and risk factors for post-heart-transplant cardiac allograft vasculopathy. *Transplantation* 2003; **76**: 679-682.

[125] Siannis F, Farewell V.T, Head J. A multi-state model for joint modelling of terminal and non-terminal events with application to Whitehall II. *Statistics in Medicine* 2007; **26**: 426-442.

[126] Solomon P.J. Effect of misspecification of regression models in the analysis of survival data. *Biometrika* 1984; **71**: 291-298.

[127] Stavola B.L de. Sampling designs for short panel data. *Econometrica* 1986; **54**: 415-424.

[128] Stavola B.L de. Testing departures from time homogeneity in multistate Markov processes. *Journal of the Royal Statistical Society: Series C* 1988; **37**: 242-250.

[129] Sternberg M.R, Satten G.A. Discrete-time nonparametric estimation for semi-Markov models of chain-of-events data subject to interval censoring and truncation. *Biometrics* 1999; **55**: 514-522.

[130] Struthers C.A, Kalbfleisch J.D. Misspecified proportional hazard models. *Biometrika* 1986; **73**: 363-369.

[131] Therneau T.M, Grambsch P.M. *Modeling survival data: extending the Cox model.* Springer. 2000.

[132] Turnbull B.W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B* 1976; **38**: 290-295.

[133] Wei G.C.G, Tanner M.A. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 1990; **85**: 699-704.

[134] White H. Maximum likelihood estimation of misspecified models. *Econometrica* 1982; **50**: 1:25.

[135] Wilks S.S. The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* 1938; **9**: 60-62.

[136] Yen A. M-F, Chen T. H-H. Mixture multi-state Markov regression model. *Journal of Applied Statistics* 2007; **34**: 11-21.

[137] Zhang B. An information matrix test for logistic regression models based on case-control data. *Biometrika* 2001; **88**: 921-932.