Statistical Inference for Heterogeneous Event History Data

by

Edmund Tze-Man Ng

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

Waterloo, Ontario, Canada, 1997

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

There has recently been considerable interest in the development of statistical methodology for the analysis of event history data. Most of the existing methods are directed to single-event time data or to transitional data based on Markov and semi-Markov assumptions. In many longitudinal studies, however, extensive subject-to-subject variability is present. Although the literature of statistical methods for the analysis of heterogeneous failure time data is vast, there remains a need to further investigate a number of issues pertaining to frailty models for failure time and more general event history data. The goal of this thesis is to develop and investigate statistical methods for modeling heterogeneous event history data. Specifically, we will focus on three areas: (*i*) tests of homogeneity; (*ii*) estimation with multiplicative random effects for intensity models; and (*iii*) marginal models based on cumulative mean functions for point processes. A strategy used throughout this thesis is to adopt piecewise constant baseline functions as a compromise between standard parametric and semi-parametric models.

Score tests are often used to test for homogeneity. We provide empirical evidence that score tests tend to have poor performance in the context of point processes with small to moderately large samples. Adjustments for the bias of the score statistics, induced by the substitution of parameter estimates, are derived for Poisson processes with parametric and semi-parametric specifications. Simulation studies suggest that the adjustment to the score test leads to much better performance in small samples. The tests based on piecewise constant intensities proves to be particularly attractive in terms of the type I error rate.

Methods of parameter estimation for mixed point processes are investigated by simulation based on Gauss-Hermite integration and the EM algorithm for log-normal and non-parametric random effects distributions respectively. Mixed Poisson and mixed renewal processes are considered. We find that the parameters of the intensity function can be estimated with negligible bias and with quite efficient variance estimates by these

methods, regardless of the true underlying mixing distribution. However, the estimate for the dispersion parameter tends to be positively biased for the Gauss-Hermite integration when the true mixing distribution is highly discrete. In contrast, variance estimates for the estimates of the masses and mass points are inflated based on the EM algorithm if the true dispersion parameter is large. These methods of estimation are also investigated in the context of a mixed two-state processes. Models which accommodate multiple time scales are also examined here.

Finally, when interest lies in relating the number of events of a point process to covariates, an alternative approach based on mean functions and estimating functions may be employed. We develop and investigate such a model in the context of bivariate point processes. The model formulation only requires correct specification of the mean functions and thus full probabilistic specification of the processes is avoided. An optimal criterion is proposed for the estimation function of the mean function parameters. Estimating functions arising from mixed bivariate Poisson processes are introduced as a working model when the covariance structure is unknown. Simulation studies indicate that the mixed Poisson estimating function performs satisfactorily. Data from a recently completed asthma clinical trial are used to illustrate this approach.

Although only univariate and bivariate processes are studied here, the methods developed here provide insight and lay the groundwork for methodology directed at the analysis of higher dimensional processes.

Acknowledgements

I wish to express my most sincere gratitude to my supervisor, Professor Richard Cook, for his valuable guidance, perpetual encouragement, and financial support during the course of this research. I would like to thank Professors John Hirdes, John Klein, Jerry Lawless, David Matthews, and Jeanette O'Hara Hines for their helpful comments and suggestions.

I would also like to thank Bayer Canada Inc. for providing the data in the CHEST study, Dr. Jayanti Mukherjee for discussions of the CHEST study, and Dr. Peter Herbison for providing the bronchial asthma data.

This work is supported partially by the Ontario Graduate Scholarship and the University of Waterloo Graduate Scholarship.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

In longitudinal studies, subjects are often observed over a period of time during which certain events of interest are recorded. For example, in studies of small bowel motility, times between the migrating motor complex are an important measure for gastrointestinal function in humans (Aalen and Husebye, 1991). As another example, patients with chronic bronchitis experience acute exacerbations of symptoms which typically alternate with periods of good respiratory health (Fietta et al., 1992), and the durations of these symptomatic and symptom-free periods are of interest. Other clinical examples arise from fields such as gastroenterology (Rokkas et al., 1995), infectious disease (Nagelkerke et al., 1990), and psychiatry (Frank et al., 1990). Such chronic disease processes are often very naturally modeled by means of event history analysis.

A frequently occurring complication in chronic diseases, however, is that individuals exhibit extensive subject-to-subject variability in their disease processes. When this variation is inadequately characterized by available covariates, this phenomenon is termed *heterogeneity*. Latent (unobserved) genetic and environmental factors, or covariates subject to

measurement errors, are common causes of heterogeneity. An artifact of this heterogeneity is that a clustering effect is induced in the data. That is, responses such as event times, arising from subjects with comparable latent factors tend to be more similar than responses arising from subjects with very different latent covariates. There has recently been considerable interest in the development of methods for the analysis of heterogeneous event history data with the two most common approaches being based on random effect (frailty) models (Clayton and Cuzick, 1985; Aalen, 1988; Klein, 1992; Nielsen et al., 1992) and marginal models (Lin, 1994; Lawless and Nadeau, 1995). Focus is mainly given to multivariate survival data (Wei et al., 1989), mixed Poisson processes (Lawless, 1987) and mixed renewal processes (Aalen and Husebye, 1991). There is, however, a lack of literature on more general stochastic processes.

In this thesis, we aim to develop statistical models and methods of inference for heterogeneous multivariate processes. In what follows, we use the term multivariate processes to refer to either multi-dimensional point processes, or multi-state processes. The general objectives of this thesis are to

(*i*) develop hypothesis tests of homogeneity;

(*ii*) consider issues in model formulation and statistical inference for random effect models in event history data; and

(*iii*) propose model formulation and develop robust inference procedure for mean and covariance functions based on the cumulative number of events.

Since the frailty model plays an important role in the analysis of heterogeneous time to event data, a review of frailty models is given in chapter 2. In chapters 3 to 5, we will study the above three topics in details. Topics for further research are discussed in chapter 6.

In sections 1.2 and 1.3 of this chapter, the above three topics of research are briefly

introduced. Section 1.4 presents three examples in which the proposed methods will be applied.

1.2 Model Formulation, Estimation and Inference

1.2.1 Mixed Univariate Point Processes

A point process is a sequence of event times governed by an intensity function. Let $N(s,t)$ be the number of events of a subject occurring in an interval $(s,t]$ for $s < t$. Furthermore let $N(t)$ denote $N(0,t)$, where t is an appropriate time scale, such as the time since the diagnosis of the disease, age, time since randomization, or calendar time. Let $\mathcal{H}(t) = \{N(s)|0 < s < t\}$ denote the history of the process up to time t. Given $\mathcal{H}(t)$, the intensity function, $\lambda(t|\mathcal{H}(t))$, is defined through the transition probabilities in a small time interval $[t, t+\delta t)$, as $\delta t \to 0^+$,

$$\begin{aligned} \Pr(N(t,t+\delta t) = 1|\mathcal{H}(t)) &= \lambda(t|\mathcal{H}(t))\delta t + o(\delta t), \\ \Pr(N(t,t+\delta t) > 1|\mathcal{H}(t)) &= o(\delta t). \end{aligned}$$

We assume that this is an orderly process, i.e., at most one event can happen at any instant (Cox and Isham, 1980). Two widely applied classes of point processes are the Poisson process and the renewal process. Statistical properties of the Poisson and renewal processes can be found in some standard textbooks such as Cox (1962), Karlin and Taylor (1975) and Cox and Isham (1980).

The transition probabilities of a Poisson process do not depend on $\mathcal{H}(t)$. Its intensity is a deterministic function of time t, and we said that it possesses a time trend. It is also well known that $N(t)$ has a Poisson distribution with mean $\Lambda(t) = \int_0^t \lambda(s)ds$ and that the numbers of events in non-overlapping time intervals are independent.

On the other hand, the transition probabilities of a renewal process depend only on the time since the occurrence of the last event, i.e., its intensity is a function of backward recurrence time $t - t_{N(t-)}$ where t_k is the event time for the kth event. This implies that the inter-event times are independent and identically distributed (iid) with hazard function $h(s) = \lambda(s + t_{N(t-)}|\mathcal{H}(t))$ for $s > 0$. Thus, a renewal process has a stable cyclical behavior.

Furthermore, if both a time trend and cycles are expected, Cox (1972a) introduced the modulated renewal process which incorporates a time trend into a renewal process by considering the intensity as a product of a function of time t and a function of backward recurrence time $t - t_{N(t-)}$. Methods of estimation for parametric and semi-parametric models under homogeneity assumption were investigated by Lawless and Thiagarajah (1996) and Oakes and Cui (1994) respectively.

In order to study the effects of covariates over a sample of m subjects, we consider the popular multiplicative intensity model (Andersen et al., 1993). Now the history for the ith subject becomes

$$\mathcal{H}_i(t) = \{N_i(s), X_i(s), Y_i(s)|0 \leq s < t\}, \tag{1.2.1}$$

where $\{X_i(t)\}$ is a $p \times 1$ covariate process, and $Y_i(t)$ is a censoring process for subject i. We make the following essential assumptions on the covariate and censoring processes (Andersen et al., 1993 Chapter VII):

1. the covariate processes are predictable with respect to the underlying filtration generated from the $\mathcal{H}_i(t)$'s and locally bounded;

2. the censoring processes are predictable and do not depend on the parameters in the intensity function.

Typically, $Y_i(t) = I(t \in [\tau_{i1}, \tau_{i2}])$, where $I(\cdot)$ is the indicator function, and τ_{i1} is the time of entry of the study and τ_{i2} is the right-censoring time for the ith subject. We also assume subsequently that τ_{i1} and τ_{i2} are stopping times with respect to $\mathcal{H}_i(t)$, see Aalen

and Husebye (1991) and Andersen et al. (1993, Chapter III) for discussion. In the case of random effect models, we further assume that τ_{i1} and τ_{i2} are independent of the random effects given $\mathcal{H}_i(t)$ (Andersen et al., 1993 Chapter IX).

Note that all inferences are based conditionally on the covariate and censoring processes. The intensity function for subject i is formulated as

$$\lambda_i(t|\mathcal{H}_i(t)) = \exp(\boldsymbol{x}_i'(t)\boldsymbol{\beta})\lambda_{i0}(t|\mathcal{H}_i(t)), \tag{1.2.2}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients and $\lambda_{i0}(t|\mathcal{H}_i(t))$ is an unknown subject-specific baseline intensity which may be function of past events.

The baseline intensities may not be the same for all subjects due to heterogeneity. Although this specification seems to be more flexible than the multiplicative frailty models, there is a serious drawback. If the baseline intensities are unspecified, the time-independent covariate effects are completely confounded with the baseline intensities and thus are not estimable. Even when parametric baseline intensities are assumed, we still run into the trouble of having a large number of nuisance parameters which is proportional to the sample size.

Since the observed covariates act multiplicatively, we may assume that the subject-specific baseline intensities are also proportional to each other and write

$$\lambda_{i0}(t|\mathcal{H}_i(t)) = v_i\lambda_0(t|\mathcal{H}_i(t)), \tag{1.2.3}$$

where the v_i's are the unknown proportionality constants and $\lambda_0(\cdot|\cdot)$ are the common baseline transition intensity. Additional model structure is still needed to overcome the problem of over-parameterization. As in many random effect models, we postulate that the V_i's are sampled independently from a distribution. We further assume that the three components of the model, the covariates, the random effects and the baseline intensity, are

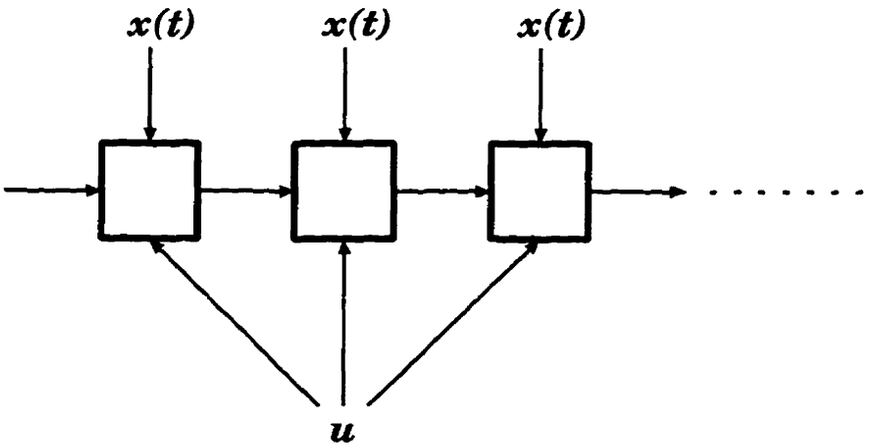

Figure 1.1: A random effect model for point processes, where the box represents the ocurrence of an event.

mutually independent. As a log-link function is adopted for the covariates, it is sometimes more convenient to apply the log-transform for V_i and denote $U_i = \log(V_i)$. For the sake of identifiability, we restrict the mean of U_i to be 0. Let $G(u)$ be the distribution function of U_i and σ be its variance. Very often, $G(u)$ is also called mixing distribution. Given u_i (or v_i), (1.2.2) is the conditional intensity with respect to the random effect. To distinguish the conditional and unconditional intensities, we adopt the universal notation and write the conditional intensity (1.2.2) as

$$\lambda_i(t|u_i; \mathcal{H}_i(t)) = \exp(\boldsymbol{x}_i'(t)\boldsymbol{\beta} + u_i)\lambda_0(t|\mathcal{H}_i(t)) \tag{1.2.4}$$

Note that now $\mathcal{H}_i(t)$ contains only the observed information of the process in $(0, t)$.

Model (1.2.4) is depicted in Figure 1.1. As covariates can be functions of past events, autoregressive structures are allowed. For instance, the intensity may depend on the previous sojourn times.

Maximum likelihood estimation for random effect models (1.2.4) for univariate point process requires the knowledge of the mixing distribution which is usually unknown. Mis-

specification of the mixing distribution may lead to inconsistency and loss of efficiency in parameter estimation. Fortunately, Pickles and Crouchley (1995) suggested that this is not a serious problem. They carried out a comparison study for different mixing distributions used in bivariate failure time data based on simulations and demonstrated that there is some robustness to a misspecified mixing distribution for the estimation and testing of regression coefficients. Models using the popular log-gamma or normal frailties perform well even in the presence of a substantial non-susceptible group (i.e., its hazard rate is zero). However, the non-parametric mixing distribution, which was thought to be robust to misspecification, did not perform well for heavily censored data. They suggested that computational convenience and the choice of the baseline intensity seem to be more important criteria than the choice of the mixing distribution.

Model (1.2.4) can be estimated by a number of approaches which will be discussed in chapter 2. Some approaches are particularly useful for certain types of mixing distribution, for example, the EM algorithm is usually applied for a non-parametric mixing distribution. A comparison of some of these estimation methods via simulations is considered in chapter 4.

1.2.2 Mixed Bivariate Processes

The concept of heterogeneity may be extended to multivariate processes in a similar fashion. This amounts to specify a multivariate mixing distribution such that each mixing component acts multiplicatively on exactly one intensity function. The choice of mixing distribution becomes heavily reliant on the tractability of likelihood function. Aalen (1987) suggested approaches in estimation for Markov processes in which the baseline transition intensities are time-invariant. A model of this sort was also considered by Cook and Ng (1997) for a two-state Markov chain in which a logistic-bivariate normal model was used to describe heterogeneity in data from an infection field study (Chunge, 1989). In addition, a

bivariate frailty model for in-patient mental health care was proposed by Xue and Brookmeyer (1996) using semi-Markov processes with bivariate log-normal frailties. However, not much work has been done for more general processes so far.

Since many applications give rise to bivariate point processes or two-state processes, we will focus on modeling such processes.

Suppose m subjects are followed over time generating m independent bivariate point processes characterizing two types of events. We assume that the processes are orderly. Let $N_{ij}(t)$ be the number of the jth type events occurring in $(0,t]$ and $N_{i\cdot}(t) = N_{i1}(t) + N_{i2}(t)$. Let $\lambda_{ij}(t|\mathcal{H}_i(t))$ denote the jth intensity function for subject i at time t given the history of the process up to time t, $\mathcal{H}_i(t) = \{N_{i1}(s), N_{i2}(s), X_i(s), Y_i(s)|0 \leq s < t\}$, for $j = 1,2$ and $i = 1,\ldots,m$, where the covariate and censoring processes are defined as in section 1.2.1. That is, as $\delta t \to 0^+$,

$$\begin{aligned}
\Pr(N_{ij}(t,t+\delta t) = 1|\mathcal{H}_i(t)) &= Y_i(t)\lambda_{ij}(t|\mathcal{H}_i(t))\delta t + o(\delta t), \\
\Pr(N_{ij}(t,t+\delta t) > 1|\mathcal{H}_i(t)) &= Y_i(t)o(\delta t),
\end{aligned}$$

where $N_{ij}(s,t) = N_{ij}(t) - N_{ij}(s)$ is the number of the jth type events occurred in $(s,t]$ for $s < t$. If simultaneous occurrence of both types of event is allowed, we define the instantaneous rate as $\Pr(N_{i1}(t,t+\delta t) = 1, N_{i2}(t,t+\delta t) = 1|\mathcal{H}_i(t)) = Y_i(t)\lambda_{i12}(t|\mathcal{H}_i(t))\delta t + o(\delta t)$.

When a two-state process is entertained, we define $N_{ij}(t)$ to be the number of $j \to 3-j$ transitions in $(0,t]$ and $Y_{ij}(t) = 1$ if subject i is in state j at time t^- and 0 otherwise. The history of the process up to time t is the collection

$$\mathcal{H}_i(t) = \{N_{i1}(s), N_{i2}(s), Y_{i1}(s), Y_{i2}(s), X_i(s), Y_i(s)|0 < s < t\}.$$

Let $\lambda_{ij}(t|\mathcal{H}_i(t))$ be the $j \to 3-j$ transition intensity given $\mathcal{H}_i(t)$. Simultaneous transition

in both directions is impossible and thus $\lambda_{i12}(t|\mathcal{H}_i(t)) = 0$. In addition, the transition probabilities over a small time interval $[t, t+\delta t)$ are defined as, for $\delta t \to 0^+$,

$$\begin{aligned}
\Pr(N_{i1}(t,t+\delta t)=1, N_{i2}(t,t+\delta t)=0|\mathcal{H}_i(t)) &= Y_i(t)Y_{i1}(t)(\lambda_{i1}(t|\mathcal{H}_i(t))\delta t + o(\delta t)),\\
\Pr(N_{i1}(t,t+\delta t)=0, N_{i2}(t,t+\delta t)=1|\mathcal{H}_i(t)) &= Y_i(t)Y_{i2}(t)(\lambda_{i2}(t|\mathcal{H}_i(t))\delta t + o(\delta t)),\\
\Pr(N_{i1}(t,t+\delta t)=0, N_{i2}(t,t+\delta t)=0|\mathcal{H}_i(t)) &= 1 - Y_i(t)\sum_{j=1}^{2} Y_{ij}(t)[\lambda_{ij}(t|\mathcal{H}_i(t))\delta t\\
&\qquad + o(\delta t)],\\
\Pr(N_{ij}(t,t+\delta t)>1|\mathcal{H}_i(t)) &= Y_i(t)Y_{ij}(t)o(\delta t), \qquad j=1,2.
\end{aligned}$$

The two-state process differs from the bivariate point process in that we observe not only $\{N_{i1}(t), N_{i2}(t)\}$ over time, but also $\{Y_{i1}(t), Y_{i2}(t)\}$ such that $Y_{i1}(t) + Y_{i2}(t) = 1$ for $t \geq 0$.

We are now going to formulate the intensity functions for the bivariate point process or the transition intensity functions for the two-state process. Assuming that simultaneous occurrence of both types of event is not allowed, a mixed multiplicative intensity model similar to (1.2.4) may be formulated as follows:

$$\lambda_{ij}(t|\boldsymbol{u}_i; \mathcal{H}_i(t)) = \exp(\boldsymbol{x}_{ij}'(t)\boldsymbol{\beta}_j + u_{ij})\lambda_{j0}(t|\mathcal{H}_i(t)), \tag{1.2.5}$$

where $\boldsymbol{x}_{ij}'(t)$ is a $p_j \times 1$ vector of covariates, $\boldsymbol{\beta}_j$ is a $p_j \times 1$ vector of fixed-effect regression coefficients, $\boldsymbol{u}_i = (u_{i1}, u_{i2})'$ is a bivariate vector of random effects with distribution function $G(\cdot;\boldsymbol{\sigma})$ indexed by $\boldsymbol{\sigma}$, and $\lambda_{j0}(t|\mathcal{H}_i(t))$ is an unknown baseline intensity for the jth type of event or the $j \to 3-j$ transition and is common to all subjects.

As in the univariate point process, we assume the $\boldsymbol{U}_i$'s are independent and identically distributed, and the covariates, the random effects and the baseline intensities are mutually independent. Figure 1.2 shows this model diagrammatically for a two-state process.

Markov processes and semi-Markov processes are widely applied in which the baseline

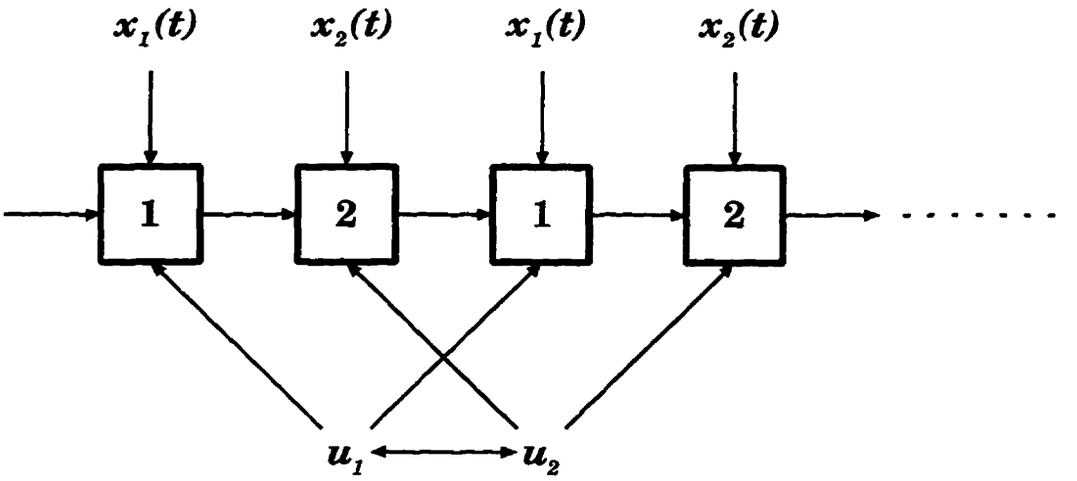

Figure 1.2: A random effect model for two-state processes, where the number inside the box represents the current state.

intensities are specific functions of $\mathcal{H}_i(t)$. We first introduce the formulations for a semi-Markov process and a process combining the Markov and semi-Markov properties, and then describe some issues pertaining to model inference.

Alternating Renewal Process

An ordinary two-state semi-Markov process is also known as an alternating renewal process. This process has the following three properties (Karlin and Taylor, 1975, p207 and Cox and Isham, 1980, p55):

1. the distribution of the duration in either state depends only on the time since the entry to that state;

2. durations in each state are independently and identically distributed; and

3. durations in different states are independent.

Loosely speaking, these properties imply that the process under consideration is relatively stable with fairly regular cycles of relapsing and remitting disease activity. Such a structure

is quite plausible for applications to many chronic diseases.

Given the renewal characteristic, the baseline intensities are functions of time since the most recent entry into the current state, in other words,

$$\lambda_{j0}(t|\mathcal{H}_i(t)) = \lambda_{j0}(t - t_{N_{i\cdot}(t^-)}), \tag{1.2.6}$$

where t_k is the kth transition time.

If the covariates are also functions of duration times $(t - t_{N_{i\cdot}(t^-)})$, this process belongs to the multivariate failure times model in time to event analysis (see chapter 2).

Modulated Alternating Renewal Process

The Markov process is a multi-state generalization of the Poisson process. Its baseline intensities are also functions of time alone, that is, $\lambda_{j0}(t|\mathcal{H}_i(t)) = \lambda_{j0}(t)$. This essentially assumes that time trends exist in all possible transitions. Incorporating this property into a semi-Markov process gives rise to a multi-state version of Cox's modulated renewal process. Here we will consider two-state processes.

Since many chronic diseases consist of relapsing and remitting disease cycles, exhibit seasonality, and show persistent risk, it is desirable to adopt a model which incorporates these behaviors simultaneously. We propose a multiplicative components model as follows:

$$\lambda_{ij}(t|\boldsymbol{u}_i; \mathcal{H}_i(t)) = \exp(\boldsymbol{x}'_{ij}(t)\boldsymbol{\beta}_j + u_{ij}) S_j(c_i(t)) R_j(b_i(t)) T_j(t), \tag{1.2.7}$$

where $S_j(\cdot)$, $R_j(\cdot)$ and $T_j(\cdot)$ are the seasonal component, the semi-Markov component and the Markov (time trend) component for the $j \to 3-j$ transition respectively, t is the time since diagnosis of the disease, $c_i(t) = t + c_i(0)$ is the calendar time, $c_i(0)$ is the date of diagnosis of the disease and $b_i(t) = t - t_{N_{i\cdot}(t^-)}$ is the backward recurrence time for subject i.

This sort of model with parametric specification allows tests for renewal or time trend; see Lawless and Thiagarajah (1996) for the treatment of univariate point processes. Model (1.2.7) will be applied to a data set in chapter 4.

Mixing Distribution and Maximum Likelihood Estimation

A number of methods for constructing a bivariate mixing distribution are discussed in Chapter 2. It is important to choose a distribution so that there is no direct functional relationship between the variances and the correlation of the intensities, i.e., a genuine bivariate distribution is most desirable (Lindeboom and Van Den Berg, 1994). A bivariate normal distribution is an obvious natural candidate. In chapter 4, we will also consider a bivariate non-parametric mixing distribution.

1.2.3 Marginal Models for Mean and Covariance Functions

Marginal models are also popular in Cox regression models (Lin, 1994) and in generalized linear models (Clayton, 1994). In many situations, we are mainly interested in modeling covariate effects on the average number of events occurring over a certain time interval. A marginal approach may be useful for this purpose.

In univariate point processes, Lawless and Nadeau (1995) constructed estimating functions for the cumulative mean function (CMF)

$$\mathrm{E}(N(t)) = \int_0^t \lambda(s)ds$$

using Poisson likelihood estimating equations with robust variance estimates. Nadeau and Lawless (1996) extended this to obtain optimal estimating functions by specifying the covariance process of $\{dN(t)\}$, where $dN(t) = N(t) - N(t^-)$.

The merit of this method is that it is not required to specify the complete probabilistic

structure of the process and yet consistent estimates for the CMF are still available. In chapter 5, we will develop a robust model for bivariate point processes based on this idea.

It is important to note that the assumptions on the censoring processes and covariate processes under this framework are slightly stronger than that in intensity models which require predictability only. Valid inference for the CMF requires that the censoring processes are independent of the point processes and the covariate processes are external in the sense of Kalbfleisch and Prentice (1980, p.123), see Nadeau (1995) for discussion.

1.3 Tests of Homogeneity

In epidemiology, it is often important to know whether the target population is homogeneous with respect to some features of interest. This also relates to testing goodness of fit where the presence of heterogeneity may imply that the available covariates are inadequate to explain the variation of the response.

Homogeneity in random effect models is equivalent to zero variance of the random effects. Thus, the null hypothesis of homogeneity is stated as

$$H_0 : \sigma = 0.$$

Score tests are often employed to test for homogeneity in the context of generalized linear models (Liang, 1987; Dean, 1992). We will construct such tests for mixed univariate point processes in chapter 3. For small sample sizes, the distribution of the score test statistic may not be approximated well by a normal distribution. To investigate this, we carry out a simulation study to investigate the performance of score tests based on their asymptotic distribution for renewal and Poisson processes with small and moderate sample sizes. Focus will be on adjusting the test statistic in order to improve the performance of the test in small samples. Adjusted score statistics for Poisson processes with parametric

and semi-parametric specifications will be derived and their frequency properties will be studied in chapter 3. Possible testing strategies with regard to bivariate processes are discussed as well.

1.4 Applications in Medical Studies

In this section, we describe three medical applications which can be modeled by the methods proposed above. The first study furnish the so-called recurrent event data, the second study belongs to the class of two-state processes, and the last study constitutes an example of bivariate point processes.

1.4.1 Gamma Interferon in CGD

In 1988, the International Chronic Granulomatous Disease (CGD) Cooperative Study Group conducted a randomized and double-blinded controlled clinical trial to study the effect of gamma interferon on reducing the rate of recurrence of serious infections due to CGD. There were 128 eligible patients with CGD of which 65 received placebo and 63 received gamma interferon. These patients were followed for about one year during which dates of the diagnosis of serious infections were recorded. In the placebo group, there were 18, 5, 4, 1, 1 and 1 patients who experienced 1, 2, 3, 4, 5 and 7 serious infections respectively. In the gamma-interferon group, there were only 9, 4 and 1 patients who experienced 1, 2 and 3 serious infections respectively. In addition to the treatment variable, there were eight time-independent covariates measured at the time of randomization (see Table 3.18). A detailed description of this study can be found in Fleming and Harrington (1991).

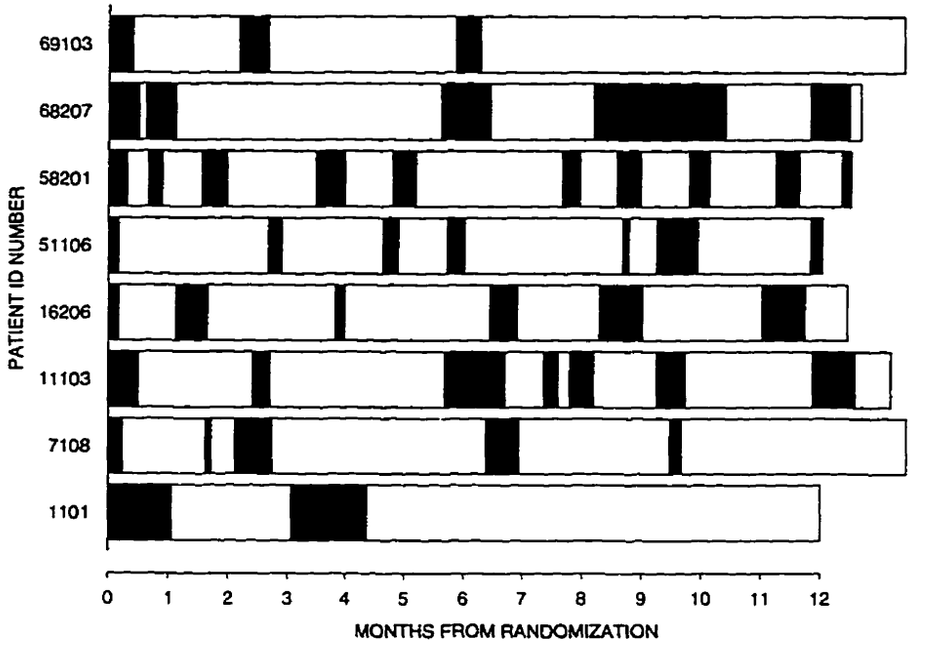

Figure 1.3: Sample Profiles of Exacerbation Patterns from the CHEST Study. Filled boxes: Exacerbation durations; Open boxes: Inter-exacerbation durations.

1.4.2 The CHEST Study

In 1993, Bayer Canada Inc. initiated a large multi-center randomized trial designed to examine the effect of Ciprofloxacin versus standard care on acute exacerbations of chronic bronchitis. One major objective of this trial was to assess the cost effectiveness of Ciprofloxacin and hence the study was called "Ciprofloxacin Health Economic Study" and received the acronym "CHEST". The analyses that we will focus on, however, are directed strictly at assessing the clinical effectiveness (e.g. no economic considerations will be made).

Patients with this disease experience acute exacerbations of chronic bronchitis (AECB) and symptom-free periods in an alternating fashion. Figure 1.3 illustrates typical patterns of chronic bronchitis durations. In this figure, the lengths of the horizontal bars represent

the total durations of follow-up for 9 patients, the filled boxes indicate days during which symptoms were manifested, and the open boxes reflect the symptom-free periods. It is clear that some patients tend to have exacerbations of symptoms of a longer average duration than the others (e.g. compare patient 1101 to patient 51106) and some tend to have these exacerbations more frequently than the others (e.g. compare patient 7108 to patient 58201). Clinically, this means that while a particular patient may have a relatively stable disease process, different patients tend to experience exacerbations of symptoms at varying rates and of varying mean durations. This is the sort of disease process described in section 1.2.2.

To be eligible for the study, patients must have been eighteen years or older, diagnosed with chronic bronchitis, able to maintain a daily diary to record the extent of health resource utilization, able to understand and complete detailed health status questionnaires, and must have been concurrently experiencing an acute exacerbation of symptoms. After randomization, follow-up visits were scheduled to take place at three-month intervals as part of the regular assessment program. In addition, patients were also required to visit their participating clinic when they perceived that a new exacerbation was beginning, or when they determined that an exacerbation was resolved. A consequence was that it was possible to determine the exact transition times corresponding to the onset and resolution of symptoms. Patient follow-up was to continue for 365 days, but early termination could occur if a subject refused to complete their symptom diary, refused to return for further follow-up visits, or died. There was a total of 115 eligible patients randomized to receive Ciprofloxacin and 107 randomized to receive standard care. The average duration of follow-up was 357 and 350 days in the Ciprofloxacin and standard care groups respectively.

In addition to the treatment variable, potential risk factors were also measured. These covariates were

1. factors characterizing the exacerbation at the time of randomization:

number of symptoms present, use of prior antibiotics therapy for the exacerbation and duration of symptoms,

2. factors characterizing the nature of the chronic bronchitis:
   severity as determined by physician assessment, bronchitis duration and the number of exacerbations in the past twelve months,

3. miscellaneous medical abnormalities:
   cardiovascular, gastrointestinal, musculoskelatal, central nervous system, endocrine, hematologic and hepatic,

4. other factors:
   gender, age at randomization and smoking history.

The codings of these covariates are listed in Table 1.1.

It is important to note that this study is somewhat different from most trials in chronic bronchitis in that follow-up was sufficiently long that multiple exacerbations could be observed. Most often, studies are designed to examine the impact of treatment on the resolution of a single exacerbation (Fietta et al., 1992). In addition, the exacerbations were defined in terms of symptoms only, with the onset and resolution dates determined by patients themselves. Often assessments of exacerbations are based on serological tests and laboratory measures of lung function. Finally, rather than having a specified control treatment, patients on the standard care arm received whatever treatment their physicians deemed appropriate. As a result, the medication corresponding to standard care varied considerably across subjects and even within subjects over different exacerbations.

The study was initially designed based on a single patient-level response which was the number of days per year during which symptoms of an acute exacerbation were manifested. During the course of the study, however, it became apparent that there were problems with this analysis strategy. Summarization of each subject's disease process in this way may

Covariate	Level	Value	Covariate	Level	Value
Treatment	Ciprofloxacin	1	Severe	Severe	1
	Standard Care	0	Bronchitis	Mild or Moderate	0
Moderate	Moderate	1	Number of	Three	1
Bronchitis	Mild or Severe	0	AECB Symptoms	Two	0
Use of Prior	Yes	1	Cardiovascular	Abnormal	1
Antibiotics	No	0		Normal	0
Gastrointestinal	Abnormal	1	Musculoskeletal	Abnormal	1
	Normal	0		Normal	0
Central Nervous	Abnormal	1	Endocrine	Abnormal	1
System	Normal	0		Normal	0
Hematologic	Abnormal	1	Hepatic	Abnormal	1
	Normal	0		Normal	0
Smoking	Yes	1	Gender	Female	1
	No	0		Male	0
Duration of Chronic Bronchitis*	Years	C†	Duration of AECB Symptoms*	Days	C
Number of AECBs in Past 12 Months*	Number	C	Age*	Years	C

*Measured at time of randomization.

†Continuous variable.

Table 1.1: Coding of the baseline covariates in the CHEST Study.

be reasonable in some contexts, but such an approach does not provide insight into the nature of the disease activity. For example, a treatment that leads to longer healthy periods between slightly longer disease periods may be of interest but may fail to be identified on the basis of the proportion of time in the diseased state. Also, this summary measure does not distinguish between a subject with many short exacerbations of symptoms and a subject with very few prolonged exacerbations, when the total number of symptom days is the same. This has been recognized in the field of migraine study; the International Headache Society Committee on Clinical Trials (1991) states that "duration cannot be recommended as one of the primary efficacy" responses, where duration here refers to the overall symptom duration. Furthermore, some patients' observation times were censored, requiring the investigators to "annualize" the observed number of exacerbation days. It seems preferable to utilize more standard methods for censored data.

1.4.3 Inhaled Beta-agonist Treatment in Bronchial Asthma

This is a double-blind, placebo-controlled, randomized, crossover study of the effect of regular inhaled bronchodilator therapy using beta-agonist (Sears et al., 1990). There were 89 subjects with stable asthma participating in the study for one year during which the subjects inhaled regularly fenoterol or placebo by a dry powder delivery system. The subjects kept daily records of their symptoms, peak expiratory flow rates and drug use, and were required to return for regular clinic visits every four weeks. There were 64 eligible subjects who completed the trial. Thirty four were in the fenoterol group and thirty were in the placebo group. Dropouts were mainly due to unstable asthma, other lung diseases, concomitant major illness, and inability to reduce bronchodilator aerosol use below 8 puffs daily.

Major symptoms of interest are wheezing and coughing during daytime and night-time periods. The cumulative numbers of these symptoms can be characterized as a bivariate

point process. We will analyze this study using a marginal model for bivariate point processes in chapter 5.

Chapter 2

Frailty Models For Event Time Data

2.1 Overview

In many medical studies involving groups of subjects or multiple measures for each subject, heterogeneity is often observed due to environmental, genetic, experiential, or other differences between subjects. Although introducing covariates into appropriate regression models can help account for subject differences, there is often extra variation between subjects that remains unexplained. There are usually three reasons for this extra variation. First, some important covariates may not have been measured or may not even be possible to measure. Second, some of the covariates may be subject to one or more sources of measurement error. Third, different subjects may react differently to certain conditions or interventions, say a treatment.

Although randomized experiments are frequently used to control for hidden sources of variation, in time-to-event analysis they cannot be considered as part of the experimental random error as they can in the familiar linear regression analysis. This is because events occur over time and thus the composition of the population may change with time. Suppose a population consists of subjects with different risks. Subjects with a higher risk tend to

experience the event earlier and the hazard rate is thus observed to be higher in the early period. Those remaining at risk will be selected to have a lower risk and as a result, the hazard rate decreases in the later period.

Vaupel and Yashin (1985) gave a number of interesting illustrations of this selection effect. They showed that even if there are only two groups of different hazard rates, the population hazard rate can be very different from the group hazard rates. In a long run, the population hazard rate will often approach the hazard of the more robust group (i.e. the lower risk group). For instance, using equations (2) to (4) in Vaupel and Yashin (1985), a population consisting of two groups with constant hazard rates λ_1 and λ_2 where $\lambda_1 > \lambda_2$ gives the population hazard rate

$$\lambda(t) = \lambda_1\pi(t) + \lambda_2(1 - \pi(t)),$$

where $\pi(t)$ is the proportion of the surviving population at time t that is in the first group:

$$\pi(t) = \frac{\pi(0)\exp(-\lambda_1 t)}{\pi(0)\exp(-\lambda_1 t) + (1 - \pi(0))\exp(-\lambda_2 t)},$$

where $\pi(0)$ is the initial proportion. It is readily seen that $\pi(t) \to 0$ and $\lambda(t) \to \lambda_2$ as $t \to \infty$. Thus, the population hazard rate is decreasing towards the smaller hazard rate. This demonstrates that without taking heterogeneity into account, every subject in the population will be wrongly thought of as having a decreasing hazard rate. In general, as a function of time the population hazard rate is the result of both the selection effects due to the variation between subjects and the variation within each subject over time. Therefore, misleading interpretations can follow if heterogeneity is ignored.

In survival analysis, the term "frailty" is often used to describe the unobserved tendency for some subjects to have hazard rates above or below the population average rate. The literature on frailty is vast. Applications can be found in various fields, such as medicine

and epidemiology (Aalen, 1987a; 1987b; 1988; Aalen and Husebye, 1991; Clayton and Cuzick, 1985; Hougaard, 1986a; Klein, 1992), demography (Vaupel et al. 1979; Heckman and Singer, 1982), econometrics (Lancaster and Nickell, 1980; Elbers and Ridder, 1982; Heckman and Singer, 1984) and reliability engineering (Lindley and Singpurwalla, 1986; Whitmore and Lee, 1991). Some good review papers include Aalen (1994), Pickles and Crouchley (1994), Clayton (1994), Hougaard (1995) and Liang et al. (1995).

Since there is a large amount of literature on frailty models, any review is necessarily selective. In this chapter, we provide such a selective review and in particular, focus on applications to medicine and epidemiology. In section 2.2, the frailty model for univariate survival data is introduced and some properties are discussed. In section 2.3, applications in multivariate survival data are considered, and the role of frailty models to measure heterogeneity and correlation is addressed. In section 2.4, some commonly used frailty distributions are presented. Finally in section 2.5, some common methods of estimation are reviewed.

2.2 Frailty For Univariate Event Times

2.2.1 The Model

Consider a sample of m independent subjects. In multiplicative intensity models (e.g. Kalbfleisch and Prentice, 1980; Lawless, 1982; Andersen et al. 1993), the frailty is usually treated as a random effect acting multiplicatively on the hazard rate:

$$\lambda_i(t|v_i) = v_i \exp(\boldsymbol{x}_i'(t)\boldsymbol{\beta})\lambda_0(t), \qquad i = 1, ..., m, \tag{2.2.1}$$

where $\boldsymbol{x}_i(t)$ is a vector of possibly time-dependent observed covariates, $\boldsymbol{\beta}$ is the vector of regression coefficients, $V_i \geq 0$ is the frailty having a probability distribution $G(v)$, and $\lambda_0(t)$

is the baseline hazard rate which is modeled either parametrically or non-parametrically, and is common to all subjects. As a usual assumption in regression analysis, the covariates, the frailty and the baseline hazard function are mutually independent.

The frailty is treated as time-invariant so that it may be thought of as reflecting hidden subject characteristics present at the start of observation, and these characteristics remain unchanged over time. More sophisticated models involving dynamic frailties are possible to formulate, but pose harder problems in terms of identification of the frailty distribution and the baseline hazard, as well as in parameter estimation. Aalen (1994) gave a brief discussion on modeling the frailty as a stochastic process. Model (2.2.1) is only intended to extract part of the unobserved variation between subjects, as suggested by Vaupel et al. (1979).

2.2.2 Effects of Frailty on Regression Coefficients

In the ordinary linear regression model, the omission of an important covariate, which is assumed to be independent of the observed covariates, is not a serious problem. The maximum likelihood method still produces consistent estimates for the regression coefficients, although the precision of the estimators will be over-estimated (Gail et al., 1984). Conversely, the omission of important covariates in a proportional hazard regression model will usually bias the observed covariate effects towards zero and distort the proportionality structure of risk functions (Struthers and Kalbfleisch, 1986; Bretagnolle and Huber-Carol, 1988). Introducing a frailty may thus help in adjusting for this bias. The trade-off is that the interpretation of the regression coefficients must be made at the subject level (Aalen, 1994).

In general, the population hazard rate is a complex function of the regression coefficients

and the baseline hazard, as can be seen from the marginal survivor function,

$$S_i(t) = \mathcal{L}_v(\Lambda_i(t)), \tag{2.2.2}$$

where $\mathcal{L}_v(\cdot)$ is the Laplace transform of V_i and $\Lambda_i(t) = \int_0^t \exp(\boldsymbol{x}_i'(s)\boldsymbol{\beta})\lambda_0(s)ds$.

Furthermore, the population relative risk is no longer time-independent. In a general class of frailty distributions proposed by Aalen (1988; 1992), the population relative risk is shown to be a monotonic function of time. In the presence of a non-susceptible group, i.e. the probability that $V = 0$ is positive, "cross-over" can happen. In other words, if the population relative risk is decreasing, it starts at a value greater than 1 and decreases to below 1 as time increases. Other frailty distributions will produce different behaviors of the population relative risks. For example, the positive stable frailty distribution suggested by Hougaard (1986a) provides constant relative risks, both conditionally and unconditionally. This will be discussed further in section 2.4.

2.2.3 Identifiability of the Frailty Model

It is not always possible to estimate model (2.2.1). In the simplest case where there are no covariates and repeated measurements, if the baseline hazard and the frailty distribution are both unspecified, they cannot be distinguished. This is readily seen from a special case that $\lambda(t|v) = \lambda_0(t)$ and $v \equiv 1$ always satisfy (2.2.2). This is, however, not surprising. In fact, this is comparable to the one-way random effects analysis of variance model in which the variance components can only be identified when there are multiple observations in each group.

In order to make identifiability possible, some extra model structure is necessary. Vaupel et al. (1979) considered specification of a gamma frailty distribution with unit mean and unknown variance. The gamma distribution has received considerable attention because of its mathematical convenience. Some other choices include the inverse Gaussian,

positive stable, log-normal and non-parametric distributions. These distributions will be discussed in section 2.4.

Elbers and Ridder (1982) proved a somewhat surprising result that under the assumptions that the frailty distribution has a finite mean and the model includes at least one covariate, the frailty model (2.2.1) is identifiable from data with only a single, possibly censored, survival time for each subject. This result is also true even for a non-parametric baseline hazard. The argument relies on the variation of the covariate values between subjects. The varying covariates enable the Laplace transform of V to be traced out on some intervals and thus it can be uniquely determined. This suggests that it is important to include covariates with a large variation in order to obtain precise estimate of the Laplace transform.

On the contrary, in the analysis of multiple events if the frailty is the same for different failure times within a cluster, Honoré (1993) proved that it is not necessary to make any assumption about the frailty distribution in order to achieve identifiability.

2.3 Frailty Models For Multiple Event Times

2.3.1 Clustered Event Times

Apart from modeling heterogeneity in univariate failure time data, frailty can also be used to describe the dependence among failure times.

Shared Frailty

Suppose there are m clusters of failure times. A general approach to model the multivariate structure of a cluster is to introduce a latent variable (frailty) into the hazard rate of each cluster member and assume that conditional on this frailty, subject failure times are independent. Specifically, let n_i denote the number of observations in the ith cluster and

$\boldsymbol{T}_i = (T_{i1}, ..., T_{in_i})'$ be the corresponding failure times. The hazard rate for T_{ij} conditional on the cluster-specific frailty v_i is formulated as

$$\lambda_{ij}(t|v_i) = v_i \exp(\boldsymbol{x}'_{ij}(t)\boldsymbol{\beta})\lambda_0(t), \tag{2.3.1}$$

for $j = 1, ..., n_i$ and $i = 1, ..., m$. The association within $\boldsymbol{T}_i$ is induced by the same frailty v_i which is assumed to have distribution function $G(\cdot)$. Further assuming that the V_i are independent and identically distributed, the joint distribution of $\boldsymbol{T}_i$ can be obtained by taking the expectation with respect to v_i of the conditional independent joint distribution of $\boldsymbol{T}_i$ given v_i. This is a reasonable assumption because the clusters are usually defined such that they are unrelated.

Although the ordinary variance component analysis in linear model theory also aims to measure the extra variation due to the correlation between observations in a cluster, it is not always applicable in multivariate survival analysis for two important reasons. First, the event times can be censored. Second, in repeated measurements over time, the $\boldsymbol{T}_i$'s constitute a point process such that there exists a natural time sequence and the cluster size n_i may depend on the event times as well. The ordinary variance component analysis certainly fails to deal with these complications.

Numerous applications can be found in genetic studies. Important examples include family studies (Clayton, 1978), twins studies (Hougaard et al., 1992; Vaupel et al., 1992) and nephrology (Hougaard, 1987).

It should be noted that the cluster-specific frailty induces not only correlation between failure times but also heterogeneity between clusters. This shared characteristic creates a confounding effect that correlation and heterogeneity cannot be measured separately. Furthermore, only positive correlation is allowed though it is common in many situations. We will address this issue in the next section.

Heterogeneity and Correlation

When heterogeneity between clusters and correlation within clusters are both present, the shared frailty model (2.3.1) is certainly not adequate to describe these two sources of extra variation. Additional structure is necessary. Lindeboom and Van Den Berg (1994) summarized a number of approaches for modeling bivariate survival data. Oakes (1989) proposed a class of archimedean distributions (Genest and MacKay, 1986) as a bivariate frailty distribution. Yashin et al. (1995) considered additive frailties in matched pairs study.

For the ease of exposition, we consider bivariate survival data with time-independent covariates. Let T_{i1} and T_{i2} be two failure times of cluster i and v_{i1} and v_{i2} be two frailties acting multiplicatively on the conditional hazards:

$$\lambda_{ij}(t|v_{i1}, v_{i2}) = \lambda_{ij}(t|v_{ij}) = v_{ij} \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_j)\lambda_{0j}(t), \tag{2.3.2}$$

where $\lambda_{0j}(t)$ is the baseline hazard, $j = 1, 2$. It is assumed that given v_{i1} and v_{i2}, T_{i1} and T_{i2} are independent.

The bivariate frailties $\boldsymbol{V}_i = (V_{i1}, V_{i2})'$ are independently and identically distributed having a distribution function $G(\boldsymbol{v})$. Let σ_j be the variance of V_{ij}, $j = 1, 2$ and ρ be the correlation between V_{i1} and V_{i2}. Although the correlation between $\lambda_{i1}(t|v_{i1})$ and $\lambda_{i2}(t|v_{i2})$ coincides with that of V_{i1} and V_{i2}, the correlation between T_{i1} and T_{i2} is difficult to derive in a general model since it depends on the baseline hazards. In the simplest case where there is no covariate and the baseline hazards are constant and equal to 1, the magnitude of the correlation between T_{i1} and T_{i2} is always less than 1/2. Therefore, even if V_{i1} and V_{i2} are perfectly correlated (negatively or positively), T_{i1} and T_{i2} are not. The bivariate failure times are uncorrelated, however, provided that V_{i1} and V_{i2} are uncorrelated.

We describe some frequently used classes of distributions for $\boldsymbol{V}_i$. The first and perhaps the most widely used distribution is a shared frailty distribution, i.e. $V_{i1} = V_{i2}$. However,

as discussed above, the drawbacks are the confounding effect between heterogeneity and correlation, and the correlation of the bivariate failure times must be positive.

A second, more general distribution for V_i that relaxes the restriction of positive correlation is a parameterized univariate specification. By taking $V_{i1} = \exp(\alpha W_i)$ and $V_{i2} = \exp(\beta W_i)$, where W_i is a univariate random variable with finite mean and variance, and α and β are parameters, the correlation between T_{i1} and T_{i2} can be negative (Clayton and Cuzick, 1985). However, ρ never equals 0 unless V_{i1} or V_{i2} is degenerate (e.g. $\alpha = 0$ or $\beta = 0$). Moreover, changing the dependence between V_{i1} and V_{i2} changes the variances of T_{i1} and T_{i2} as well.

A third, still not very satisfactory, approach is to use additive frailties. Analogous with the ordinary linear mixed model, the specification may be taken as $V_{i1} = W_{i0} + W_{i1}$ and $V_{i2} = W_{i0} + W_{i2}$ where W_{i0}, W_{i1} and W_{i2} are mutually independent and their distributions belong to the same family. The variances of W_{i1} and W_{i2} measure the degree of heterogeneity and that of W_{i0} measures the association within cluster. Yashin et al. (1995) studied this bivariate structure under the independent gamma frailty distributions for which the distributions of W_{i1} and W_{i2} are the same. Petersen et al. (1995) generalized this model to any cluster size. Nevertheless, ρ is still restricted to be positive.

The fourth and the most general approach is a genuine bivariate distribution of V_i (Butler et al., 1989; Lineboom and Van Den Berg, 1994). That is, there is no direct relationship between the variances and the correlation of T_{i1} and T_{i2}. An obvious choice is a bivariate log-normal distribution. However, this bivariate frailty distribution does not lead to a closed form expression for the marginal joint distribution of T_{i1} and T_{i2}, and so maximization routines involving numerical quadrature are required. As an alternative, McGilchrist and Aisbett (1991) and McGilchrist (1993) applied the best linear unbiased prediction and the residual maximum likelihood estimation similar to those used in estimating generalized linear mixed models (e.g. Breslow and Clayton, 1993), but bias correction seems to be necessary (e.g. Breslow and Lin, 1995; Kuk, 1995).

2.3.2 Recurrent Events

A special case of clustered event time data arises when a subject generates multiple event times from a point process. There is assumed to be a sequence of events observed from a subject over a period of time. This forms a point process for which the event times for each subject are recorded. Such longitudinal studies are often found in clinical trials.

A point process is governed by its intensity function which is often expressed as a function of its past history. Under the presence of heterogeneity, the conditional intensity of subject i given its history and frailty v_i can be expressed as

$$\lambda_i(t|v_i, \mathcal{H}_i(t)) = v_i \exp(\boldsymbol{x}_i'(t)\boldsymbol{\beta})\lambda_0(t|\mathcal{H}_i(t)), \tag{2.3.3}$$

where $\mathcal{H}_i(t)$ is the history of the process up to but not including time t. Note that the covariates and the baseline intensity can be functions of $\mathcal{H}_i(t)$. Here the subject variation is measured by the term $\exp(\boldsymbol{x}_i'(t)\boldsymbol{\beta})\lambda_0(t|\mathcal{H}_i(t))$ and the difference between subject intensities is measured by the variance of the frailty V_i.

In particular, the Poisson process for which $\lambda_0(t|\mathcal{H}(t)) = \lambda_0(t)$ and the renewal process for which $\lambda_0(t|\mathcal{H}(t)) = \lambda_0(t - t_{N(t-)})$ are widely applied, where $N(t)$ is the number of events occurring in $(0, t]$, $t_{N(t-)}$ is the time of the event occurring just before time t and $t_0 = 0$.

The Poisson process essentially assumes that the intensity process has a time trend, i.e., the baseline intensity does not depend on past events and it is a deterministic function of time alone. Lawless (1987) gave a comprehensive formulation of regression models for fixed-effect or random effect Poisson processes with parametric or non-parametric baseline intensities. On the other hand, given the covariates and the frailty, if the inter-event times are independent and the hazard for an event depends only on the time measured from the occurrence of the last event, the point process is known as a renewal process. This stable cyclical behavior is the principal characteristic of a renewal process and is common in medical studies of biological functions and chronic diseases. For example, Aalen and

Husebye (1991) applied the gamma frailty and a Weibull baseline hazard to model the small bowel motility in healthy adults.

Combining a time trend and renewal behavior in a point process has been suggested by Cox (1972a) who introduced the term "modulated renewal process". Inference procedures for this type of process, using fixed-effect semi-parametric and parametric approaches, have been considered by Oakes and Cui (1994) and Lawless and Thiagarajah (1996) respectively. Extensions to accommodate a frailty are also possible, see Chapter 4.

Further generalizations to incorporate a dynamic frailty process to account for unobserved time-dependent covariates are formidable because the complex structure usually leads to an intractable likelihood function. In addition, it may be necessary to obtain a large number of recurrent times per subject so that model estimation is possible. Recent advances in this direction employ the dynamic Bayesian approach (Gamerman, 1991), a piecewise gamma process (Paik et al., 1994), a gamma random walk process (Yue and Chan, 1997) and a first-order autoregressive log-normal process (Yau and McGilchrist, 1996).

2.3.3 Multivariate Processes

In event history analysis, multivariate processes often arise as Markov processes (e.g. Andersen, 1986), semi-Markov processes (e.g. Fietta et al., 1992), competing risks models (e.g. Butler et al., 1989; Heckman and Taber, 1994) or three-state "illness-death" models (e.g. Andersen, 1988; Kalbfleisch and Lawless, 1988; Lindsey and Ryan, 1993). As an important first step to generalize the frailty model to multi-state processes, Aalen (1987b) discussed various ways to construct multivariate frailty distributions with known Laplace transforms for finite-state and time-homogeneous Markov processes. The marginal likelihood is then obtained by using the Laplace transform of the frailty distribution. Therefore, it is important to choose a frailty distribution whose Laplace transform has a nice closed form. In

particular, the positive stable distribution (Hougaard, 1986b) and a transformation from a multivariate normal distribution are considered. However, numerical problems are likely to arise in maximum likelihood estimation, except probably for processes with small number of states which may give a more numerically tractable likelihood. For example, Xue and Brookmeyer (1996) proposed a random effect model for the alternating renewal process using a bivariate log-normal frailty in which the likelihood is computed by a Gaussian quadrature rule.

2.4 Some Particular Frailty Distributions

2.4.1 Univariate Frailty Distributions

Univariate frailty is usually used in univariate survival data and recurrent events. In this section we describe the features of some widely used univariate frailty distributions. In particular, we will discuss two important classes of distributions proposed by Hougaard (1986a) and Aalen (1988). The well-known gamma, positive stable and inverse Gaussian distributions are special cases of these two classes. Another advantage of the Hougaard and Aalen distributions is that they have explicit Laplace transforms so that no numerical approximation is necessary in constructing the marginal likelihood, see (2.2.2).

Furthermore, the log-normal distribution and the non-parametric distribution are often alternative choices. They do not belong to the above two classes. Features of these distributions will be addressed in this section too.

A summary is given in Table 2.1 to compare different distributions. For simplicity no covariate is included in the hazard function. The population hazard and the population relative risk based on the conditional proportional hazards are given to examine the effect of frailty, where $\lambda_1(t|v) = vr\lambda_0(t)$ and $\lambda_2(t|v) = v\lambda_0(t)$, $r > 1$ is the conditional relative risk and the integrated baseline hazard $\Lambda_0(t)$ is assumed to be an increasing function of t

and $\lim_{t\to\infty} \Lambda_0(t) = \infty$.

The Hougaard Distributions

Hougaard (1986a) extended the positive stable distribution to a class of frailty distributions which is indexed by three parameters (α, δ, θ), where $0 \leq \alpha \leq 1$, $\delta > 0$ and $\theta \geq 0$. The Laplace transform is given in Table 2.1. This class of distributions is shown to be absolutely continuous and unimodal.

The Hougaard distribution contains the following distributions: the inverse Gaussian distribution ($\alpha = 1/2$), the positive stable distribution ($\theta = 0$) and the gamma distribution ($\alpha \to 0$). The distribution is degenerate at δ when $\alpha = 1$ and at 0 when either $\delta = 0$ or $\theta \to \infty$.

For a non-degenerate distribution with non-zero θ, the mean and variance are equal to $\delta\theta^{\alpha-1}$ and $\delta(1-\alpha)\theta^{\alpha-2}$ respectively. It also includes distributions with infinite mean and variance when $\theta = 0$, for example, the positive stable distribution ($\delta = \alpha$). The degree of heterogeneity can be measured by α only. The population is less heterogeneous when α is closer to 1.

The population relative risk is a decreasing function of t for $\theta > 0$. It is equal to r at $t = 0$. As $t \to \infty$, the population relative risk tends to r^α which is still greater than 1 but less than r. Therefore, under specification of a fixed-effect proportional hazard model, the frailty attenuates the covariate effect.

The positive stable distribution is distinct from the other frailty distributions in that it preserves the proportionality of the hazards. The population relative risk is a constant independent of time, r^α. Nevertheless, it also attenuates the covariate effect.

The Aalen Distributions

Aalen (1988) modified the Hougaard distribution based on a compound Poisson distribution. The Aalen distribution is usually taken to have mean 1. It has two parameters $a \geq 0$ and $b \geq 0$. The parameter b is the variance and the parameter a divides the class of distributions into two major categories. For $a \leq 1$, after a reparametrization, it is identical to the Hougaard distribution with unit mean. For $a > 1$, it is a compound Poisson distribution generated by gamma variables. Further properties of this class of distributions were investigated by Aalen (1992). It should be noted that the positive stable distribution does not belong to this class because of its infinite mean and variance.

The compound Poisson distribution (i.e. $a > 1$) is particularly interesting. It allows a non-zero probability at $V = 0$. This corresponds to the situations in which some subjects are not susceptible to the studied disease while the rest have a varying degree of susceptibility. Mathematically the hazard function for the non-susceptible subjects is equal to 0, which is equivalent to $V = 0$. The probability of non-susceptibility can be shown as

$$Pr(V = 0) = \exp(-\frac{a}{b(a-1)}),$$

for $a > 1$. The compound Poisson distribution is not unimodal as opposed to the Hougaard distribution due to the point mass at zero. Aalen (1992) demonstrated that the continuous part of this distribution is not necessarily unimodal either. The multiple modes imply that there exist several subgroups of different risk levels. This usually occurs when the variance is small or a is large.

Furthermore, the compound Poisson distribution leads to the paradox of "crossing" relative risk. As shown in Table 2.1, the population relative risk decreases from r at $t = 0$ to approach r^{1-a} which is less than 1 for $a > 1$. Therefore, this phenomenon may arise because of the presence of a non-susceptible group.

Distribution[1]	Laplace Transform	Variance	Population Hazard[2]	Population Relative Risk[3]
Hougaard (α, δ, θ) $\delta = \theta^{1-\alpha}$ $0 \le \alpha \le 1, \delta \ge 0$	$\exp\left\{-\frac{\delta}{\alpha}[(\theta+s)^\alpha - \theta^\alpha]\right\}$, $0 < \alpha \le 1$ $(1+\theta^{-1}s)^{-\theta}$, $\alpha = 0$	$\frac{1-\alpha}{\theta}$	$\frac{\lambda_0(t)}{(1+\theta^{-1}\Lambda_0(t))^{1-\alpha}}$	$r\left[\frac{\theta+\Lambda_0(t)}{\theta+r\Lambda_0(t)}\right]^{1-\alpha}$
Aalen (a, b) $a \ge 0, b \ge 0$	$\exp\left\{\frac{a}{(1-a)b}[1-(1+\frac{b}{a}s)^{1-a}]\right\}$, $a \ne 1$ $(1+bs)^{-1/b}$, $a = 1$	b	$\frac{\lambda_0(t)}{(1+ba^{-1}\Lambda_0(t))^a}$	$r\left[\frac{a+b\Lambda_0(t)}{a+rb\Lambda_0(t)}\right]^a$
Gamma (δ, δ) $\delta > 0$	$(1+\delta^{-1}s)^{-\delta}$	$\frac{1}{\delta}$	$\frac{\lambda_0(t)}{1+\delta^{-1}\Lambda_0(t)}$	$r\left[\frac{\delta+\Lambda_0(t)}{\delta+r\Lambda_0(t)}\right]$
Positive Stable $0 < \alpha \le 1$	$\exp(-s^\alpha)$	∞	$\frac{\alpha\lambda_0(t)}{\Lambda_0(t)^{1-\alpha}}$	r^α
Inverse Gaussian $\delta \ge 0$	$\exp\left\{-2\delta[(\delta^2+s)^{1/2}-\delta]\right\}$	$\frac{1}{2\delta}$	$\frac{\lambda_0(t)}{(1+\delta^{-2}\Lambda_0(t))^{1/2}}$	$r\left[\frac{\delta^2+\Lambda_0(t)}{\delta^2+r\Lambda_0(t)}\right]^{1/2}$
Lognormal $(0, \delta)$ $\delta > 0$	—	$(e^\delta - 1)e^\delta$	—	—

[1]The means of the Hougaard, Aalen, gamma and inverse Gaussian distributions equal 1. The positive stable distribution does not have a finite mean. The mean of the lognormal distribution is $\exp(\delta/2)$.

[2]The conditional hazard given v is $v\lambda_0(t)$.

[3]The conditional hazards are $\lambda_1(t|v) = vr\lambda_0(t)$ and $\lambda_2(t|v) = v\lambda_0(t)$ for $r > 1$. The population relative risk is defined as $\lambda_1(t)/\lambda_2(t)$ where $\lambda_j(t)$ is the population hazard corresponding to $\lambda_j(t|v)$.

Table 2.1: Some common univariate frailty distributions.

Other Frailty Distributions

Apart from the Hougaard distributions and the Aalen distributions, it is hard to find a frailty distribution whose Laplace transform has an explicit form. Therefore, numerical integration methods, such as the Gaussian quadrature and Monte Carlo approaches, are often needed to compute the Laplace transforms. Heckman and Singer (1984) and Butler et al. (1989) showed the use of Gaussian quadrature in estimating an unspecified frailty distribution, while Clayton (1991) applied the Gibbs sampling approach. On the other hand, the log-normal distribution often serves as a substitute for the gamma distribution (McGilchrist and Aisbett, 1991), especially for multivariate frailty. The log-normal distribution is also included in Table 2.1.

Another important approach is to approximate the unknown continuous frailty distribution by a discrete distribution. Under some weak regularity conditions, including that the frailty has a finite mean and variance, this corresponds to the non-parametric maximum likelihood representation of a continuous distribution (Laird, 1978).

Interestingly, based on some limited simulation studies, the choice of particular parametric (discrete or continuous) frailty distribution is not critical for the estimation and testing of regression coefficients with multiple measurements and univariate frailty (Pickles and Crouchley, 1995).

2.4.2 Multivariate Frailty Distributions

Extension to multivariate frailties with independent components is straightforward and appropriate marginal likelihood may be easily constructed by taking the product of its component terms. However, as measures of missing covariate effects, the frailty components are often likely to be dependent. Some approaches based on transformations of independent random variables to generate bivariate frailties have been discussed in section 2.3.1. The restrictions to positive dependence or possible functional relationships between variances

and correlation of failure times within clusters make these transformations less attractive.

The multivariate log-normal distribution seems to be the simplest genuine multivariate distribution, although there is no closed form expression for the marginal likelihood. Evans and Swartz (1995) pointed out that the multiple quadrature method is preferred to other numerical methods (e.g. Monte Carlo integration and Gibbs sampler) for dimension less than or equal to 6. As in most medical applications the number of frailty components is seldom more than 6, this approach seems to be appropriate.

Alternatively, techniques such as the penalized likelihood and the residual maximum likelihood estimation, used widely in the generalized linear mixed models (Breslow and Clayton, 1993) may be employed. Cluster sizes are usually small in survival analysis, however and this may cause substantial bias making bias corrections necessary (Breslow and Lin, 1995).

The use of a non-parametric multivariate frailty distribution may be also entertained. Butler et al. (1989) showed some success in the bivariate case using Gaussian quadrature.

2.5 Some Methods of Estimation For Frailty Models

Estimation for frailty models is challenging due to the non-linearity of the model and the introduction of the mixing distribution. Closed form marginal likelihoods are not available if the frailty distribution does not have an explicit Laplace transform. Numerical integration techniques, the EM algorithm, and the Gibbs sampler are popular alternative approaches. The penalized quasi-likelihood method widely used in the generalized linear mixing models may be useful as well. Some methods are particularly useful for certain types of processes, for example, models for multi-state processes with a multivariate Gaussian frailty distribution lead to computationally efficient estimates by Gauss-Hermite integration, and a univariate non-parametric frailty distribution is most appropriately estimated by the EM algorithm.

In this section, we review these common methods of estimation with regard to their relative merits and limitations. For the ease of exposition, we will focus on the clustered failure time model (2.3.1) with a single frailty.

2.5.1 The Laplace Transform

Let $t_{i1}, \ldots, t_{in_i}$ be n_i observed times with censoring indicators $d_{i1}, \ldots, d_{in_i}$ in cluster i where $d_{ij} = 1$ if t_{ij} is a failure time and 0 if censoring time. We assume that censorings are non-informative and only right-censoring is allowed. Further assuming that the covariates are time-invariant or constant between consecutive events and independent of the frailty, and using the conditional hazard function (2.3.1), the conditional likelihood for cluster i given v_i can be written as

$$L_i(\boldsymbol{\theta}|v_i) = \prod_{j=1}^{n_i} \lambda_{ij}(t_{ij};\boldsymbol{\theta})^{d_{ij}} v_i^{d_{i\cdot}} \exp\left[-v_i \sum_{j=1}^{m_i} \Lambda_{ij}(t_{ij};\boldsymbol{\theta})\right], \tag{2.5.1}$$

where $\boldsymbol{\theta}$ is the vector of parameters for the covariates and the baseline intensity, $\lambda_0(t)$, is modeled parametrically by γ, $d_{i\cdot} = \sum_{j=1}^{n_i} d_{ij}$, $\lambda_{ij}(t) = \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta})\lambda_0(t)$ and $\Lambda_{ij}(t) = \int_0^t \lambda_{ij}(s)ds$. Then the marginal likelihood for cluster i is given by

$$\begin{aligned} L_i(\boldsymbol{\phi}) &= \int_0^\infty L_i(\boldsymbol{\theta}|v_i)dG(v_i) \\ &= \prod_{j=1}^{n_i} \lambda_{ij}(t_{ij})^{d_{ij}} (-1)^{d_{i\cdot}} \mathcal{L}_v^{(d_{i\cdot})}\Big(\sum_{j=1}^{n_i} \Lambda_{ij}(t_{ij})\Big), \end{aligned} \tag{2.5.2}$$

where $\boldsymbol{\phi} = (\boldsymbol{\theta}', \boldsymbol{\sigma}')'$, $\boldsymbol{\sigma}$ is the vector of the frailty parameters and $\mathcal{L}_v^{(d)}(s)$ is the dth derivative of $\mathcal{L}_v(s)$. The full likelihood based on m independent clusters is obtained as the product of the individual likelihoods.

The log-likelihood can be obtained provided that $\mathcal{L}_v^{(d)}(s)$ is known. For example, the Laplace transform for the gamma density with mean 1 and variance σ is given by $\mathcal{L}_v(s) =$

$(1+\sigma s)^{-1/\sigma}$ and hence,

$$\mathcal{L}_v^{(d)}(s) = (-1)^d \prod_{j=0}^{d-1}(1+(j-1)\sigma)(1+\sigma s)^{-1/\sigma-d},$$

for $d \geq 1$. The log-likelihood for cluster i is then equal to

$$\ell_i(\phi) = \sum_{j=1}^{n_i} d_{ij}\log(\lambda_{ij}(t_{ij})) + \sum_{j=1}^{d_i.-1}\log(1+(j-1)\sigma) - (\frac{1}{\sigma}+d_i.)\log(1+\sigma\sum_{j=1}^{n_i}\Lambda_{ij}(t_{ij})).$$

Therefore, the marginal likelihood may be directly maximized by standard maximization methods (e.g. the Newton-Raphson procedure).

However, it may not be possible to derive a general expression for $\mathcal{L}_v^{(d)}(s)$ for other frailty distributions even when $\mathcal{L}_v(s)$ has a closed form. The Hougaard and Aalen distributions do not have such general expressions as the gamma distribution. The $\mathcal{L}_v^{(d)}(s)$ for the Hougaard distribution can be found by the following recursive relationship,

$$\mathcal{L}_v^{(d)}(s) = \sum_{r=0}^{d-1}\binom{d-1}{r}\mathcal{L}_v^{(r)}(s)h_v^{(d-r)}(s),$$

where $h_v(s) = \log(\mathcal{L}_v(s))$ and $h_v^{(d-r)}(s) = -\delta\prod_{j=1}^{d-1}(\alpha-j)(\theta+s)^{\alpha-r}$, see Table 2.1. Unless the cluster size is small, say < 3, the maximization procedure will be certainly complicated by such structure.

2.5.2 The Numerical Integration

If the Laplace transform of the frailty density has no closed form or higher derivatives are difficult to obtain, the numerical integration method is a natural alternative.

Using the transformation $V = \exp(U)$ for the frailty, the marginal likelihood for the

ith cluster in (2.5.2) can be written as

$$L_i(\phi) = \int_{-\infty}^{\infty} L_i(\theta|u_i)\frac{g(u_i)}{\phi(u_i)}\phi(u_i)du_i,$$

where $g(\cdot)$ is the probability density function of u_i and $\phi(\cdot)$ is the standard normal density function. Combining terms involving u_i into the integrand, we get

$$L_i(\phi) = \prod_{j=1}^{m_i} \lambda_{ij}(t_{ij})^{d_{ij}} \int_{-\infty}^{\infty} h_i(u;\phi)\phi(u)du,$$

where $h_i(u;\phi) = \exp(d_{i\cdot}u - e^u \sum_{j=1}^{n_i} \Lambda_{ij}(t_{ij}))g(u)/\phi(u)$. Applying the Gauss-Hermite rule, the integral can be approximated by

$$\int_{-\infty}^{\infty} h_i(u;\phi)\phi(u)du \approx \frac{1}{\sqrt{\pi}} \sum_{r=1}^{R} c_r h(\sqrt{2}z_r;\phi),$$

where the c_r's are weights, the z_r's are nodes, and R is the number of nodes. Tables for the nodes and weights can be found in Abramowitz and Stegun (1972). This approximation is more accurate if U is normally distributed. In this case, if we let $U = \sqrt{\sigma}Z$ where σ is the variance of U, we have a simpler expression for $h(\cdot)$: $h_i(z;\phi) = \exp(d_{i\cdot}\sqrt{\sigma}z - e^{\sqrt{\sigma}z} \sum_{j=1}^{n_i} \Lambda_{ij}(t_{ij}))$.

Evans and Swartz (1995) surveyed a number of numerical integration methods commonly used in statistics. They recommended that a multiple quadrature rule, such as the Gauss-Hermite rule, should be used for low dimension integration problems. Although it is almost impossible to assess the error as this requires the exact evaluation of the integration, we may use rules with more nodes to check if the results are comparable to those obtained by using fewer nodes.

2.5.3 The EM Algorithm

As the frailty model is a mixture model, the EM algorithm (Dempster et al., 1977) becomes a useful tool in estimation, especially for non-parametric baseline hazards or discrete frailty distributions. In fact, in dealing with the infinite-dimensional nuisance parameter $\Lambda_0(t)$ the integration over frailties destroys the construction of partial likelihood (Cox, 1975) in its usual way. This makes direct maximization of the marginal likelihood impossible. Here we describe briefly how to use the EM algorithm for parametric and semi-parametric models with continuous and discrete frailty distributions.

Continuous Frailties

If the v_i were observed, the complete data log-likelihood for cluster i is given by

$$\begin{aligned} \ell_F(\phi) &= \sum_{i=1}^{m} \ell_i(\theta|v_i) + \sum_{i=1}^{m} \log(g(v_i;\sigma)) \\ &\equiv \ell_C(\theta|v) + \ell_v(\sigma;v), \end{aligned} \tag{2.5.3}$$

where $\ell_i(\theta|v_i) = \log(L_i(\theta|v_i))$ from (2.5.1) and $g(\cdot;\sigma)$ is the density function of V parameterized by σ. This function could be maximized by separately maximizing $\ell_C(\theta|v)$ with respect to θ, and $\ell_v(\sigma;v)$ with respect to σ.

As the frailties v_i are unknown, the EM algorithm considers maximizing the conditional expectation of the complete data likelihood given the observed data and the current estimates of the parameters. Specifically, in the E-step, we compute $\mathrm{E}(\ell_F(\phi)|data;\phi^{(k)}) \equiv Q(\phi|data;\phi^{(k)})$, where $\phi^{(k)}$ is the estimate of ϕ at step k. This can be expressed as, apart from some constants,

$$Q(\phi|data;\phi^{(k)}) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} d_{ij}\left(\log(\lambda_{ij}(t_{ij})) - \hat{v}_i^{(k)}\Lambda_{ij}(t_{ij})\right)$$

$$+\sum_{i=1}^{m} \mathrm{E}(\log(g(v_i;\sigma))|data;\phi^{(k)}),$$

where $\hat{v}_i^{(k)} = \mathrm{E}(v_i|data;\phi^{(k)})$.

The function Q is then maximized with respect to ϕ to get the next estimate. This procedure is iterated until the estimates converge.

Maximizing $\ell_C(\theta|v)$ for given v is equivalent to maximizing the homogeneous or fixed-effect model. If the baseline hazard is unspecified, the estimates of the regression coefficients are obtained by maximizing the partial likelihood (Cox, 1972b) with the offsets $\hat{v}_i^{(k)}$'s. Then the Nelson-Aalen estimator (Andersen et al., 1993) can be used as a non-parametric estimate for $\Lambda_0(t)$ (Nielsen et al., 1992 and Klein, 1992):

$$\hat{\Lambda}_0^{(k+1)}(t) = \sum_{t_{(i)} \le t} \frac{d_{(i)}}{\sum_{j \in R(t_{(i)})} \hat{v}_j^{(k)} \exp(x_j' \hat{\beta}^{(k+1)})},$$

where $t_{(i)}$ is the ith smallest event time, regardless of clusters; $d_{(i)}$ is the number of events at $t_{(i)}$; $R(t_{(i)})$ is the set of individuals at risk at time $t_{(i)}$; $\hat{v}_j^{(k)}$ is the conditional expectation of the frailty given the data and $\phi^{(k)}$, x_j is the covariate value associated with the jth individual in the pooled sample respectively; and $\hat{\beta}^{(k+1)}$ is the estimate obtained from the Cox regression.

The conditional expectation for ℓ_v may require numerical integration. The conjugate property of gamma distribution in multiplicative models provides a convenient way to avoid this complication. If v_i is distributed as gamma with mean 1 and variance σ, it is straightforward to show that the conditional distribution of v_i given the data is also gamma with shape parameter $1/\sigma + d_i$. and scale parameter $1/\sigma + \sum_{j=1}^{n_i} \Lambda_{ij}(t_{ij})$. Hence, functions of v_i in (2.5.3) are imputed by the corresponding conditional expectations with respect to this gamma conditional distribution.

Discrete Frailties

If the frailty distribution is unspecified, we may approximate it by a discrete distribution whose masses and mass points are estimated along with the model parameters. Laird (1978) showed that this is equivalent to the non-parametric maximum likelihood estimation for the unknown frailty distribution. She also provided an EM algorithm for carrying out the maximum likelihood estimation, see also McLachlan and Basford (1987).

Suppose the frailty distribution is given by

$$\Pr(V = \xi_h) = \pi_h, \qquad \text{for} \quad h = 1, ..., H; \qquad \sum_{h=1}^{H} \pi_h = 1,$$

where $\xi_h > 0$. Let $\boldsymbol{\xi} = (\xi_1, ..., \xi_H)'$, $\boldsymbol{\pi} = (\pi_1, ..., \pi_H)'$ and $\boldsymbol{\sigma} = (\boldsymbol{\xi}', \boldsymbol{\pi}')'$. Define $Z_{ih} = I(V_i = \xi_h)$ where $I(\cdot)$ is the indicator function, and let $\boldsymbol{Z}_i = (Z_{i1}, ..., Z_{iH})'$. Note that $\boldsymbol{Z}_1$, ..., $\boldsymbol{Z}_m$ are i.i.d. multinomial with probabilities $\boldsymbol{\pi}$. The complete data likelihood is given by

$$\ell_F(\boldsymbol{\phi}) = \sum_{h=1}^{H} \sum_{i=1}^{m} z_{ih}(\log(\pi_h) + \ell_i(\boldsymbol{\theta}|\xi_h)),$$

where $\boldsymbol{\phi} = (\boldsymbol{\theta}', \boldsymbol{\sigma}')'$. For the sake of identifiability, the covariate vector does not contain the intercept term. We further assume a parametric baseline hazard.

In the E-step, the conditional expectations of Z's given the data and the previous estimates are given by

$$\hat{z}_{ih}^{(k)} = \mathrm{E}(Z_{ih}|data, \boldsymbol{\phi}^{(k)}) = \frac{\pi_h^{(k)} L_i(\boldsymbol{\theta}^{(k)}|\xi_h^{(k)})}{\sum_{h'=1}^{H} \pi_{h'}^{(k)} L_i(\boldsymbol{\theta}^{(k)}|\xi_{h'}^{(k)})}.$$

The update for $\boldsymbol{\pi}$ in the M-step is given by

$$\pi_h^{(k+1)} = \frac{1}{m} \sum_{i=1}^{m} \hat{z}_{ih}^{(k)}.$$

Furthermore, the update for $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are obtained by maximizing

$$\sum_{h=1}^{H}\sum_{i=1}^{m}\hat{z}_{ih}^{(k)}\ell_i(\boldsymbol{\theta}|\xi_h).$$

When the number of masses is unknown, there are usually two approaches to choose H (Laird, 1978). One approach is to start with a small H and gradually increase H until the adjacent mass points become close. Another approach is to start with a large H and combine the adjacent mass points if the distance between them is small. Butler and Louis (1992) began with $H = m$ and combined mass points if their distance was smaller than 10^{-4}.

The Observed Information

If the observed information matrix is difficult to obtain analytically, methods provided by Louis (1982) and Meng and Rubin (1991) may be used. McLachlan and Basford (1987, chapter 1) provides an approximate observed information matrix for the case of discrete frailties:

$$\hat{I}(\hat{\boldsymbol{\phi}}) = \sum_{i=1}^{m}\left(\sum_{h=1}^{H}\hat{z}_{ih}\frac{\partial g_{ih}(\hat{\boldsymbol{\phi}})}{\partial\boldsymbol{\phi}}\right)\left(\sum_{h=1}^{H}\hat{z}_{ih}\frac{\partial g_{ih}(\hat{\boldsymbol{\phi}})}{\partial\boldsymbol{\phi}}\right)', \tag{2.5.4}$$

where $g_{ih}(\boldsymbol{\phi}) = \log(\pi_h) + \ell_i(\boldsymbol{\theta}|\xi_h)$. Caution must be taken when the number of mass points is unknown because the non-parametric approach does not take into account the variability of the estimated number of mass points. This may lead to unrealistic variance estimates for the parameter estimates based on the observed information matrix (2.5.4). Butler and Louis (1992) suggest that appropriate standard errors require the bootstrap or other methods that incorporate more than the mode and curvature of the likelihood.

However, inference is not yet completely resolved in the semi-parametric approach due to the complexity of the non-parametric baseline intensity, although Murphy (1994; 1995) has shown the existence, consistency and asymptotic normality of the estimators in the

special case of no covariates. Based on some empirical applications and simulation studies, Nielsen et al. (1992) suggested that a large sample size is required to obtain reasonable precision in estimating the frailty parameter. As recommended by some authors (Aalen, 1994; Pickles, 1994), a parametric baseline intensity may be more preferable. For instance, the piecewise exponential specification may be used if there is no prior knowledge of the baseline intensity.

2.5.4 The Penalized Partial Likelihood

There are some remarkable similarities between the frailty model and the generalized linear mixed model (GLMM). The multiplicative intensity implies that the frailty is also additive in the linear predictor $\boldsymbol{x}'\boldsymbol{\beta}$. In the case of recurrent events, if the events are frequent per individual and the intensity varies slowly with time, we can divide the continuous time into a number of subintervals and recast the problem in terms of the number of events occurring in each subinterval. Let y_k be the counts in the kth subinterval whose width is t_k. Assuming that given the frailty v, the counts are Poisson variables with

$$\begin{aligned} \mathrm{E}(y_k|v) &= \Lambda(t_k|v), \\ \log(\Lambda(t_k|v)) &= \boldsymbol{x}'\boldsymbol{\beta} + u + \log(\Lambda_0(t_k)), \\ \mathrm{var}(y_k|v) &= \Lambda(t_k|v), \end{aligned}$$

where $u = \log(v)$, then this formulation is a standard mixed Poisson regression model. On the other hand, when events are rare and the intensity varies more rapidly, a finer division of intervals is necessary. In the extreme, the semi-parametric approach takes the division into infinitesimal intervals in which at most one event can occur. Therefore, estimation methods used in the GLMM may also be useful in the frailty model with probably some modifications. Clayton (1994) described the relationship between the frailty model and

the GLMM.

When only the conditional mean and variance of the observations are specified, a commonly used method of estimation in GLMM is the penalized quasi-likelihood. A review of this method and the marginal approach is given by Breslow and Clayton (1993). In survival analysis with an unspecified baseline hazard, the counterpart of the quasi-likelihood is the partial likelihood. Penalized partial likelihood may be used to estimate the regression coefficients (McGilchrist and Aisbett, 1991) and restricted maximum likelihood (REML) to estimate the frailty parameter (McGilchrist, 1993). If the frailty has a log-normal distribution, the penalized partial likelihood is the product of the usual Cox's partial likelihood conditional on the frailties and the joint log-normal distribution of the frailties. Given the frailty parameter, the penalized partial likelihood is maximized with respect to the regression coefficients and the frailties. The frailty parameter is then estimated by a linearized REML procedure given the current estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$.

However, the simulation studies given in McGilchrist (1993) showed a systematic bias for the REML estimate of the frailty parameter, even though the REML aims to adjust for the bias due to the substitution of $\boldsymbol{\beta}$ by its estimate. There are two possible reasons for the cause of bias. First, as the REML estimate is derived from the mixed normal linear model, it may fail to account for the bias in the frailty model which is a non-linear model. Second, which may be the most serious aspect, the estimate of $\boldsymbol{u}$ is obtained essentially by approximating the posterior distribution of $\boldsymbol{u}$ by a normal distribution. This can be seen by viewing the penalized partial likelihood estimation as a posterior mode estimation (Fahrmeir and Tutz, 1994, Chapter 7). The posterior mode estimation is an EM-type procedure in which the posterior expectation and variance of $\boldsymbol{u}$ are approximated by the posterior mode and curvature respectively in the E-step. This approximation is appropriate only if the cluster sizes are large. As the cluster sizes are usually small in survival analysis, the bias may be substantial. McCullagh and Tibshirani (1990), Breslow and Lin (1995) and Kuk (1995) considered methods of bias correction for GLMM. Similar approaches may

be applied in frailty model, although intensive computation is unavoidable.

2.5.5 The Bayesian Approach

Since the frailty model is already a hierarchical model, we only need to specify priors for the parameters $\boldsymbol{\beta}$, $\Lambda_0(t)$ and σ. We assume that the hyperparameter σ for the frailty distribution is univariate.

Clayton (1991) presented an excellent account of the Gibbs sampling approach. Inspired by Kalbfleisch's (1978) Bayesian analysis of proportional hazard models, Clayton considered the priors for $\boldsymbol{\beta}$ and $\Lambda_0(t)$ suggested by Kalbfleisch. Specifically, $\boldsymbol{\beta}$ is assumed to have improper uniform prior on $[-\infty, +\infty]$, and $\Lambda_0(t)$ is a stochastic process with independent increments distributed as

$$d\Lambda_0(t) \sim Ga(cd\Lambda^*(t), c),$$

where $Ga(a, b)$ denotes the gamma distribution with shape parameter a and scale parameter b, $c > 0$ and $\Lambda^*(t)$ is a known non-decreasing function. The mean and the variance of $\Lambda_0(t)$ are given by $\Lambda^*(t)$ and $\Lambda^*(t)/c$ respectively. Moreover, the gamma frailty with variance σ is used. The variance is assumed to have a prior $Ga(\eta, \mu)$ in which we take $\eta = \mu = 0$ to obtain an non-informative prior.

Under this setup, the conditional distributions of any one of the parameters $\boldsymbol{\beta}$, $\Lambda_0(t)$, $\boldsymbol{v}$ and σ given the other 3 parameters and the observed data are known, see Clayton (1991) for the derivations. Denote the conditional distribution of x given y by $[x|y]$. Since the hyperparameter only depends on the frailty, Clayton suggests the following algorithm:
In iteration k,

1. Draw σ^* randomly from the previous B σ's: σ^{k-1}, ..., σ^{k-B}.

2. Put $\sigma = \sigma^*$ and repeat G times:

   (a) Generate $\Lambda_0^i \sim [\Lambda_0|\boldsymbol{\beta}^{i-1}, \boldsymbol{v}^{i-1}, Data]$.

(b) Generate $\boldsymbol{\beta}^i \sim [\boldsymbol{\beta}|\boldsymbol{v}^{i-1}, \Lambda_0^i, Data]$.

(c) Generate $\boldsymbol{v}^i \sim [\boldsymbol{v}|\boldsymbol{\beta}^i, \Lambda_0^i, \sigma^*, Data]$.

3. Generate $\sigma^k \sim [\sigma|\boldsymbol{v}^*]$, where $\boldsymbol{v}^*$ is the current value of $\boldsymbol{v}$.

The number B is called the buffer size which should be gradually increased with the iterations. The number G is the number of steps needed to avoid serious serial correlation in the samples. In the example analyzed by Clayton, he suggested that increasing B from 1 to 100 and choosing G to be 10 should be appropriate.

Nevertheless, other frailty distributions may not have a closed form for the posterior distribution. Numerical integrations, such as the Monte Carlo integration, are necessary in this case. The Bayesian approach is also useful for multi-level hierarchical models and multivariate frailty models.

Chapter 3

Tests of Homogeneity For Point Processes

3.1 Overview

In epidemiological studies, it is often important to know whether the studied population is homogeneous. As will be seen in chapters 4 and 5, estimation methods for random effect and marginal models are complicated whereas fixed-effect models are relatively simpler. It is, therefore, desirable to check whether it is necessary to introduce a random effect to reflect another component of variation in the model.

Tests of homogeneity in GLMMs are often derived via score tests (e.g. Liang, 1987; Dean, 1992; Smith and Heitijan, 1993; Jacqmin-Gadda and Commenges, 1995; Lin, 1997). Commenges and Andersen (1995) and Gray (1995) developed score tests in the context of failure time data based on counting processes. These score tests can be also interpreted as tests of model specification for fixed-effect versus random effect models provided that the baseline intensity and the covariate effects are correctly specified. As a general model specification test, the information matrix (IM) test (White, 1982) is often applied. Chesher

(1984) showed that the implicit alternative hypothesis of the IM test is a model with random parameters. In fact, the test statistics for the random intercept are asymptotically equivalent for the score and the IM tests. However, the finite sample estimators for the variance of the test statistic have different representations. Choice of representation should be determined by the efficiency in terms of size and power of the test. Some simulation studies have been carried out for this purpose (Orme, 1990).

In this chapter, we focus on mixed univariate point processes in order to get some insight for multivariate point processes. We construct a score test and an IM test of homogeneity in sections 3.2 and 3.3, respectively. As there are some common representations of the test arising from different estimates for the variance of the statistic, we assess the finite sample performance of the tests based on these representations via simulations in section 3.4. This simulation study indicates that the test statistic in general has poor performance in small and moderate sample sizes. Although the test statistic is asymptotically unbiased, a non-zero bias is induced in small samples upon substitution of the parameter estimates obtained under the null model of homogeneity. Adjustments for the score statistics via first order Taylor series expansion have been previously suggested in the contexts of Poisson regression models for count data (Dean and Lawless, 1989) and clustered failure time data (Gray, 1995). These adjusted score statistics were shown to have better performance in small samples in terms of the size of the test.

In section 3.5, we derive adjusted score statistics for regression models based on Poisson processes with both parametric and semi-parametric formulations of the baseline intensity. The primary objective is to examine the frequency properties of these statistics in modest sample sizes, and to identify suitable strategies for testing for homogeneity in these contexts. Simulation studies pertaining to the size and power of the tests are also considered. An example involving a clinical study of gamma interferon in chronic granulomatous disease is provided in section 3.5 as well.

Furthermore, we discuss adjustment of the test statistic for other point processes and

possible extensions for bivariate processes with correlated random effects in section 3.6.

3.2 Score Tests of Homogeneity

In random effect models, homogeneity is equivalent to zero variance of the random effect. Thus, the null hypothesis of homogeneity is simply $H_0 : \sigma = 0$ where σ is the variance of the random effect. Score tests are particularly attractive for this purpose for three reasons. First, although the variance is at the boundary of the parameter space under H_0, Moran (1971) showed that under some mild regularity conditions, the usual asymptotic theory still follows for the score test. Second, only the fixed-effect model needs to be fitted. Third, only some weak conditions for the moments of the mixing distribution are required but it is not necessary to specify the entire distribution so that misspecification of the mixing distribution is less influential. This is because the test statistic is evaluated under H_0 in which the process is invariant to any random effect. The mixing distribution is usually assumed to have a finite mean and variance, and the third and higher order moments are of the order $o(\sigma)$ (Liang, 1987; Dean, 1992).

We consider m independent univariate mixed point processes with intensity (1.2.4). In many situations, the covariates are constant between events and we assume that they are so here for simplicity. Thus, the intensity during the $(j-1)$th to the jth events is given by $\lambda_i(t) = \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + u_i)\lambda_0(t;\boldsymbol{\gamma})$ for the ith process, $i = 1, ..., m$. Recall that $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$. The score test statistic can be derived from a slight modification of Dean's (1992) or Liang's (1987) results. We assume that the baseline intensity and the functional form of the covariates are correctly specified.

Let $0 \leq t_{i1} < \cdots < t_{in_i} \leq \tau_i$ be the observed event times for the ith process over a time interval $[0, \tau_i]$ in which τ_i is a right censoring time. Let $Y_i(t) = 1$ if $t \in (0, \tau_i]$ and 0 otherwise for $t \geq 0$. For notational convenience, we let $t_{i0} = 0$ and $t_{i,n_i^c} = \tau_i$ where $n_i^c = N_i(\tau_i)$ if $t_{in_i} = \tau_i$ and $N_i(\tau_i) + 1$ if $t_{in_i} < \tau_i$ is the number of durations including the

possibly censored duration after the last observed event.

As postulated in section 1.2.1, the process $\{Y_i(t)\}$ is predictable and independent of the parameters in the intensity function. Furthermore, we subsequently assume that τ_i is finite and therefore n_i is also finite, $i = 1, ..., m$. The asymptotic setting considered will be for a large number of subjects, i.e., as $m \longrightarrow \infty$, with bounded n_i's.

Let $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i$ be the linear predictor during the $(j-1)$th to the jth events of process i and $\boldsymbol{\eta}_i = (\eta_{i1}, ..., \eta_{in_i^c})'$ for $j = 1, ..., n_i^c$ and $i = 1, ..., m$, where the U_i's are iid with mean zero and variance σ. Let $\eta_{ij}^0 = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$ be the corresponding linear predictor under H_0. We also assume that the third and higher order moments of U_i are of the order $o(\sigma)$. The conditional log-likelihood for process i given u_i is given by

$$\ell_i(\boldsymbol{\theta}|u_i) = \sum_{j=1}^{n_i}[\eta_{ij} + \log\lambda_0(t_{ij})] - \sum_{j=1}^{n_i^c} e^{\eta_{ij}}\Lambda_0(t_{i,j-1}, t_{ij}), \tag{3.2.1}$$

where $\Lambda_0(a,b) = \int_a^b \lambda_0(t)dt$. By applying a Taylor series expansion on $\exp[\ell_i(\boldsymbol{\theta}|u_i)]$ from (3.2.1) at $\boldsymbol{\eta}_i = \boldsymbol{\eta}_i^0$, we get

$$\exp[\ell_i(\boldsymbol{\theta}|u_i)] = \exp[\ell_i(\boldsymbol{\theta}|0)]\left[1 + \left(\frac{\partial\ell_i(\boldsymbol{\theta}|u_i)}{\partial\boldsymbol{\eta}_i'}\bigg|_{\boldsymbol{\eta}_i=\boldsymbol{\eta}_i^0}\right)u_i\mathbf{1}_i + \frac{1}{2}u_i^2\mathbf{1}_i'K_i(\boldsymbol{\theta})\mathbf{1}_i + o(u_i^2)\right],$$

where $\mathbf{1}_i$ is the $n_i^c \times 1$ vector of 1's and

$$K_i(\boldsymbol{\theta}) = \left(\frac{\partial^2\ell_i(\boldsymbol{\theta}|u_i)}{\partial\boldsymbol{\eta}_i\partial\boldsymbol{\eta}_i'} + \frac{\partial\ell_i(\boldsymbol{\theta}|u_i)}{\partial\boldsymbol{\eta}_i}\frac{\partial\ell_i(\boldsymbol{\theta}|u_i)}{\partial\boldsymbol{\eta}_i'}\right)\bigg|_{\boldsymbol{\eta}_i=\boldsymbol{\eta}_i^0},$$

whose (j,j)th element is $K_{ijj}(\boldsymbol{\theta}) = (d_{ij} - \exp(\eta_{ij}^0)\Lambda_0(t_{i,j-1}, t_{ij}))^2 - \exp(\eta_{ij}^0)\Lambda_0(t_{i,j-1}, t_{ij})$ and the (j,k)th element is $K_{ijk}(\boldsymbol{\theta}) = (d_{ij} - \exp(\eta_{ij}^0)\Lambda_0(t_{i,j-1}, t_{ij}))(d_{ik} - \exp(\eta_{ik}^0)\Lambda_0(t_{i,k-1}, t_{ik}))$ for $j \neq k$ and $d_{ij} = 1$ for $j = 1, ..., n_i$ and 0 otherwise. Note that $\ell_i(\boldsymbol{\theta}|0)$ and $K_i(\boldsymbol{\theta})$ are

independent of u_i. For small σ, the marginal log-likelihood can be expressed as

$$\begin{aligned} \ell_i(\phi) &= \log \mathrm{E}[\exp[\ell_i(\theta|u_i)]] \\ &= \ell_i(\theta|0) + \log(1 + \frac{1}{2}\sigma \mathbf{1}_i' K_i(\theta)\mathbf{1}_i + o(\sigma)), \end{aligned} \tag{3.2.2}$$

where $\phi = (\theta', \sigma)'$. Differentiating (3.2.2) with respect to σ and evaluating at $\sigma = 0$, the score function for σ under H_0 based on m independent processes is given by

$$S = \frac{1}{2}\sum_{i=1}^{m} \mathbf{1}_i' K_i(\theta)\mathbf{1}_i. \tag{3.2.3}$$

It is interesting to note that $K_i(\theta)$ is the difference between the cross-product of the score and the observed information about η_i evaluated under H_0, and thus S measures the overall difference. In fact, this relates to the IM test which we will discuss in the next section.

In addition, since given n_i, the integrated intensities $\exp(\eta_{ij}^0)\Lambda_0(t_{i,j-1}, t_{ij})$ are independently and exponentially distributed with unit mean for $j = 1, ..., n_i$, the score function may be thought of comparing the sample variance and model-based variance of the integrated intensity. This interpretation is consistent with that of Dean's (1992) test.

The variance of the score function is obtained by using the Fisher information matrix. We partition it conformably to (θ, σ) as

$$I(\theta, \sigma) = \begin{pmatrix} I_{\theta\theta} & I_{\theta\sigma} \\ I_{\theta\sigma}' & I_{\sigma\sigma} \end{pmatrix}.$$

The components of $I(\theta, \sigma)$ evaluated at $\sigma = 0$ are easily seen as

$$\begin{aligned} I_{\theta\theta} &= -\sum_{i=1}^{m} \mathrm{E}\left(\frac{\partial^2 \ell_i(\theta, 0)}{\partial\theta\partial\theta'}\right), \\ I_{\theta\sigma} &= -\mathrm{E}\left(\frac{\partial S}{\partial\theta}\right), \end{aligned}$$

$$I_{\sigma\sigma} = \frac{1}{4}\sum_{i=1}^{m} \mathrm{E}\left((1_i' K_i(\boldsymbol{\theta}) 1_i)^2\right),$$

where the expectation is taken over the joint density of the $(T_{i1}, ..., T_{in_i}, N_i(\tau_i))$'s. Note that $I_{\theta\theta}$ is simply the information matrix for the fixed-effect model. However, the expected information matrix may not be obtainable in general, and so the observed information matrix has to be used in most cases.

The asymptotic variance of S is equal to $V_S = I_{\sigma\sigma} - I_{\theta\sigma}' I_{\theta\theta}^{-1} I_{\theta\sigma}$ (Pierce, 1982). Hence, the test statistic is given by

$$T_S = \frac{S}{\sqrt{V_S}}, \tag{3.2.4}$$

which is evaluated at the maximum likelihood estimate of $\boldsymbol{\theta}$ using the fixed-effect model. The test statistic is approximately $N(0,1)$ distributed under H_0. As σ is non-negative, this is a one-sided test for which H_0 is rejected if T_S is large. Furthermore, if we are interested in testing against either the under-dispersion or the over-dispersion alternative, we can use T_S^2 as a test statistic which is asymptotically $\chi^2(1)$ distributed.

The score statistic is computed by replacing the unknown parameters by their estimates. For small sample sizes, the distribution of the score statistic may be quite different from its asymptotic distribution. Simulation studies for binomial and Poisson regression models suggested that moderate to large sample sizes are needed for reasonable approximation of the asymptotic distribution (Dean, 1988). Using a first-order Taylor series approximation, Dean provides an adjusted score statistic which converges faster to the asymptotic distribution. O'Hara Hines (1995) carried out a comparative study of score tests for overdispersion in binomial and Poisson regressions. She found that Dean's score test using the observed information to estimate V_S is too liberal and the adjusted statistic only provides a little improvement over the unadjusted statistic. Similar adjusted score tests for Poisson processes will be considered in section 3.5.

Following Smith and Heitjan (1993), extension to multivariate random effects with inde-

pendent components is straightforward. However, the theory does not follow immediately for correlated random effects. We will discuss the bivariate case in section 3.6.

3.3 Information Matrix Tests of Homogeneity

Crouchley and Pickles (1993) illustrated the use of the IM test in parametric univariate and multivariate proportional hazards models. Interestingly, they showed that testing for the specification of the intercept term is equivalent to the score test of homogeneity described in section 3.2. This is not a coincidence. In fact, Chesher (1983, 1984) gave a score test interpretation for the IM test and showed that the implicit alternative of the IM test is a model with random parameter variation.

In this section, we introduce the formal set-up of the IM test with some finite sample representations. We also show how it can be used to test for homogeneity in random effect models for point processes.

3.3.1 The Information Matrix Test

The IM test aims to detect model misspecification. The idea is to compare two estimates of the Fisher information matrix which are consistent under the correct model specification. In an ordinary likelihood setting, suppose $\{Y_1, ..., Y_m\}$ is a random sample from a distribution $F(Y;\boldsymbol{\theta})$. If the distribution and the parameters are correctly specified, the Fisher information matrix can be consistently estimated by either the negative of the Hessian of the log-likelihood, $-\partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$, or the outer product of the gradient, $(\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta})(\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}')$, where $\ell(\boldsymbol{\theta})$ is the log-likelihood, provided that some regularity conditions are satisfied (White, 1982). Therefore, the difference between these two estimates may be used as a test statistic. A significant discrepancy indicates that the model $F(Y;\boldsymbol{\theta})$ is misspecified.

Let $f(y;\boldsymbol{\theta})$ be the probability density function of Y. Define

$$\begin{aligned} A_m(\boldsymbol{\theta}) &= \frac{1}{m}\sum_{i=1}^{m}\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \\ A(\boldsymbol{\theta}) &= \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}\left(\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right), \\ B_m(\boldsymbol{\theta}) &= \frac{1}{m}\sum_{i=1}^{m}\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}, \\ B(\boldsymbol{\theta}) &= \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}\left(\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right), \end{aligned}$$

where $\ell_i(\boldsymbol{\theta})$ is the log-likelihood for Y_i. Let p be the dimension of $\boldsymbol{\theta}$ and $q = p(p+1)/2$. The test statistic is defined as a $q \times 1$ vector

$$D_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{d}_i(\boldsymbol{\theta}),$$

where $\boldsymbol{d}_i(\boldsymbol{\theta})$ is the vector stacking the distinct elements of

$$\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}.$$

If the model is correctly specified, White (1982) showed that $\sqrt{m}D_m(\hat{\boldsymbol{\theta}})$ is asymptotically distributed as a normal random variable with mean zero and covariance matrix $V(\boldsymbol{\theta}_0)$, where $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ are the maximum likelihood estimate and the true value of $\boldsymbol{\theta}$, respectively, under the assumed model and

$$\begin{aligned} V(\boldsymbol{\theta}) &= \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}(\boldsymbol{w}_i(\boldsymbol{\theta})\boldsymbol{w}_i(\boldsymbol{\theta})'), \\ \boldsymbol{w}_i(\boldsymbol{\theta}) &= \boldsymbol{d}_i(\boldsymbol{\theta}) - \nabla D(\boldsymbol{\theta})A(\boldsymbol{\theta})^{-1}\nabla\ell_i(\boldsymbol{\theta})', \\ \nabla D(\boldsymbol{\theta}) &= \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}\left(\frac{\partial \boldsymbol{d}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right), \\ \nabla\ell_i(\boldsymbol{\theta}) &= \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}. \end{aligned} \tag{3.3.1}$$

Hence, the normalized statistic is given by

$$IM = mD_m(\hat{\boldsymbol{\theta}})'V(\hat{\boldsymbol{\theta}})^{-1}D_m(\hat{\boldsymbol{\theta}}), \tag{3.3.2}$$

which is asymptotically $\chi^2(q)$ distributed. This is also true asymptotically if $V(\hat{\boldsymbol{\theta}})$ is replaced by any consistent estimate. Since the variance is obtained from expectations, we call (3.3.2) the efficient score (ES) form of the test.

Computation of the covariance matrix $V(\boldsymbol{\theta})$ could be rather involved due to the third derivative of the log-likelihood and the fact that the expectation is often intractable. It is a common practice that the covariance matrix is estimated by its sample moment,

$$V_m(\hat{\boldsymbol{\theta}}) = \frac{1}{m}\sum_{i=1}^{m}\hat{\boldsymbol{w}}_i(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{w}}_i(\hat{\boldsymbol{\theta}})', \tag{3.3.3}$$

where $\hat{\boldsymbol{w}}_i(\boldsymbol{\theta})$ is equal to $\boldsymbol{w}_i(\boldsymbol{\theta})$ with $\nabla D(\boldsymbol{\theta})$ and $A(\boldsymbol{\theta})$ replaced by their sample moments

$$\nabla D_m(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial \boldsymbol{d}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \qquad \text{and} \qquad A_m(\boldsymbol{\theta}) \tag{3.3.4}$$

respectively. We call the statistic (3.3.3) using (3.3.4) the Hessian form of the IM test. However, this sample moment is not guaranteed to be non-negative definite.

For the purpose of lessening the computation burden, Chesher (1983) and Lancaster (1984) suggested using the outer products $-\sum_{i=1}^{m}\boldsymbol{d}_i(\boldsymbol{\theta})\nabla\ell_i(\boldsymbol{\theta})/m$ and $-B_m(\boldsymbol{\theta})$ to approximate $\nabla D(\boldsymbol{\theta})$ and $A(\boldsymbol{\theta})$ respectively. This is due to the fact that if the model is correct, we have the following relationships:

$$\mathrm{E}\left(\frac{\partial \boldsymbol{d}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right) = -\mathrm{E}(\boldsymbol{d}_i(\boldsymbol{\theta})\nabla\ell_i(\boldsymbol{\theta})) \qquad \text{and} \qquad \mathrm{E}(\nabla\ell_i(\boldsymbol{\theta})'\nabla\ell_i(\boldsymbol{\theta})) = -A(\boldsymbol{\theta}). \tag{3.3.5}$$

We call the statistic (3.3.3) using (3.3.5) the OPG (outer product gradient) form of the IM test. Therefore, the standardized IM test has three common representations in which

the variance can be the expected one or approximated by its corresponding finite sample moments or the outer products of gradients. The ES form should converge faster to the asymptotic distribution than the other forms. Furthermore, Orme (1990) provided empirical evidence based on simulations for certain regression models that the OPG form is inferior to the Hessian form in terms of the size of the test.

3.3.2 Testing for Homogeneity

We consider the same model for univariate point processes in section 3.2. Sub-vectors of $D_m(\boldsymbol{\theta})$ can be considered for tests of particular hypotheses. Since random effect models involve a random intercept, we may apply the IM test to the intercept term.

We assume that, apart from the intercept, the other components of the model are correctly specified. Then , after some simple algebra, the statistic is given by,

$$\begin{aligned} D_m(\beta_0) &= \frac{1}{m}\sum_{i=1}^{m}\left[\frac{\partial^2 \ell_i(\boldsymbol{\theta}|0)}{\partial \beta_0^2} + \left(\frac{\partial \ell_i(\boldsymbol{\theta}|0)}{\partial \beta_0}\right)^2\right] \\ &= \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_i' K_i(\boldsymbol{\theta}) \mathbf{1}_i, \end{aligned}$$

where $\ell_i(\boldsymbol{\theta}|0)$ is the log-likelihood under H_0 and $K_i(\boldsymbol{\theta})$ is defined in section 3.2.1. Hence, the IM statistic is the same as the score statistic apart from a constant $(2/m)$. As $m \longrightarrow \infty$, $D_m(\beta_0)$ converges in distribution to a normal random variable with mean 0 and variance $V(\beta_0)$ which is the $(1,1)$th element of $V(\boldsymbol{\theta})$ in (3.3.1). It can be shown that $V(\beta_0)$ is proportional to the variance of the score statistic where the proportionality constant is equal to $4/m$. We outline the proof as follows. Using (3.3.5) with $\nabla \ell_i(\boldsymbol{\theta}|0)$ in place of $\nabla \ell_i(\boldsymbol{\theta})$, it is straightforward to find that

$$V(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}(\boldsymbol{d}_i(\boldsymbol{\theta})\boldsymbol{d}_i(\boldsymbol{\theta})') + \nabla D(\boldsymbol{\theta}) A(\boldsymbol{\theta})^{-1} \nabla D(\boldsymbol{\theta})'.$$

The first row of $\nabla D(\boldsymbol{\theta})$ is

$$\nabla D(\beta_0) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}\left(\frac{\partial(1_i' K_i(\boldsymbol{\theta}) 1_i)}{\partial \boldsymbol{\theta}'}\right) = -\frac{2}{m} I'_{\theta\sigma},$$

and $A(\boldsymbol{\theta}) = -I_{\theta\theta}/m$, where the I's are defined in section 3.2.1. The $(1,1)$th element of $V(\boldsymbol{\theta})$ is given by

$$\begin{aligned} V(\beta_0) &= \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}\left[(1_i' K_i(\boldsymbol{\theta}) 1_i)^2\right] - \frac{4}{m} I'_{\theta\sigma} I_{\theta\theta}^{-1} I_{\theta\sigma} \\ &= \frac{4}{m} V_S. \end{aligned}$$

Hence, the standardized IM statistic

$$T_E = \frac{\sqrt{m} D_m(\beta_0)}{\sqrt{V(\beta_0)}} \tag{3.3.6}$$

is exactly the same as T_S. This ES test statistic is computed by substituting the maximum likelihood estimates of $\boldsymbol{\theta}$ under H_0.

Although the score test and the IM test of homogeneity are equivalent using the expected variance, in most situations we have to use the sample moment approximations which may vary with respect to the size and power. We will investigate this by some simulation studies in the next section.

3.4 Finite Sample Performance of the Tests

In this section, we study the performance of different representations of the test via simulations based on renewal and Poisson processes with Weibull intensities. Finite sample distributions of the test statistics are estimated and compared with their asymptotic distributions. We also consider the squared statistics which are $\chi^2(1)$ distributed asymptotically.

It is found that the squared statistics seem to approach their asymptotic distributions faster than their unsquared counterparts. However, except for the ES statistic, the other statistics require large sample sizes to achieve reasonable performance.

The power of the test is studied by incorporating a random effect into the intensity. Crouchley and Pickles (1993) provided empirical evidence for failure time data based on simulations that the Hessian form and the OPG form of the IM test have similar size and power for the squared statistic.

We describe the simulation designs and introduce some common representations of the test statistics in sections 3.4.1. The results will be discussed in sections 3.4.2 and 3.4.3.

3.4.1 Design of Simulation Studies

We consider Weibull intensity functions with a time-invariant covariate

$$\lambda(t) = \begin{cases} \exp(\beta_0 + \beta_1 x)\gamma t^{\gamma-1}, & \text{for Poisson processes,} \\ \exp(\beta_0 + \beta_1 x)\gamma (t - t_{N(t-)})^{\gamma-1}, & \text{for renewal processes,} \end{cases} \tag{3.4.1}$$

where t is the time since the process starts, $t_{N(t-)}$ is the event time just before t and $\gamma \geq 1$.

Event times are generated given known values of the parameters over a fixed time interval $[0, 1]$. The designs are listed in Table 3.1. The average number of events per process is calculated from the simulation. Two different values of β are used in order to see if moderate to large number of events has an effect on the distribution. Certainly we would expect a loss in power of the test for a small number of events, say < 2.

Let $0 < t_{i1} < \ldots < t_{in_i} \leq \tau_i$ be the event times of process i and $n_i = N_i(\tau_i)$. The log-likelihood for process i under the null hypothesis is given by

$$\ell_i(\theta) = n_i\eta_i + \sum_{j=1}^{n_i} \log(\lambda_0(t_{ij})) - \exp(\eta_i)\Lambda_0(\tau_i),$$

Design	Process	γ	β_0	β_1	x	Average number of events per process
P1	Poisson	2	1.5	0.5	$Bern(0.5)$	5.8
P2	Poisson	2	1.5	0.5	$N(0,1)$	4.5
P3	Poisson	2	2	0.5	$Bern(0.5)$	9.5
P4	Poisson	2	2	0.5	$N(0,1)$	7.4
R1	Renewal	3	5	2	$Bern(0.5)$	8
R2	Renewal	3	5	1	$N(0,1)$	5.8
R3	Renewal	3	3	2	$Bern(0.5)$	4.1
R4	Renewal	3	3	1	$N(0,1)$	2.8

Table 3.1: Simulation designs.

where $\eta_i = \beta_0 + \beta_1 x_i$, $\boldsymbol{\theta} = (\beta_0, \beta_1, \gamma)'$ and $\Lambda_0(t)$ is the cumulative baseline intensity function such that

$$\Lambda_0(\tau_i) = \begin{cases} \tau_i^\gamma & \text{for Poisson processes,} \\ \sum_{j=1}^{n_i^c} e_{ij}^\gamma & \text{for renewal processes,} \end{cases}$$

where $e_{ij} = t_{ij} - t_{ij-1}$ and n_i^c is the number of durations including the last censored duration. This leads to a simpler expression for $K_i(\boldsymbol{\theta})$ in the test statistic:

$$K_i(\boldsymbol{\theta}) = (n_i - \Lambda_i(\boldsymbol{\theta}))^2 - \Lambda_i(\boldsymbol{\theta}), \tag{3.4.2}$$

where $\Lambda_i(\boldsymbol{\theta}) = \exp(\eta_i)\Lambda_0(\tau_i)$. Maximum likelihood estimates, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$ can be found easily by standard methods. We note that $\sum_{i=1}^m n_i = \sum_{i=1}^m \exp(\hat{\eta}_i)\hat{\Lambda}_0(\tau_i)$ from the score function of β_0. A simpler computational form for $\sum_{i=1}^m K_i(\hat{\boldsymbol{\theta}})$ is thus equal to $\sum_{i=1}^m [(n_i - \Lambda_i(\hat{\boldsymbol{\theta}}))^2 - n_i]$.

As the score statistic and the IM statistic differ only by a scale factor, the test statistic can be expressed as

$$T_a = \frac{\sum_{i=1}^m K_i(\hat{\boldsymbol{\theta}})}{\sqrt{V_a(\hat{\boldsymbol{\theta}})}}, \tag{3.4.3}$$

where $V_a(\hat{\boldsymbol{\theta}})$ is an estimate for the expected variance $V_E(\hat{\boldsymbol{\theta}})$ in which $a = S$, H and O correspond to the score statistic, the Hessian form of the IM test and the OPG form of the IM test respectively. We state these variances as follows:

$$V_E(\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathrm{E}[K_i(\boldsymbol{\theta})^2] + \mathrm{E}(\nabla K_\cdot(\boldsymbol{\theta})) \left(\mathrm{E}(\nabla^2 \ell(\boldsymbol{\theta}))\right)^{-1} \mathrm{E}(\nabla K_\cdot(\boldsymbol{\theta})'), \tag{3.4.4}$$

$$V_S(\boldsymbol{\theta}) = \sum_{i=1}^{m} K_i(\boldsymbol{\theta})^2 + \nabla K_\cdot(\boldsymbol{\theta}) \left(\nabla^2 \ell(\boldsymbol{\theta})\right)^{-1} \nabla K_\cdot(\boldsymbol{\theta})', \tag{3.4.5}$$

$$V_H(\boldsymbol{\theta}) = \sum_{i=1}^{m} w_i^H(\boldsymbol{\theta}) w_i^H(\boldsymbol{\theta})', \tag{3.4.6}$$

$$V_O(\boldsymbol{\theta}) = \sum_{i=1}^{m} w_i^O(\boldsymbol{\theta}) w_i^O(\boldsymbol{\theta})' \tag{3.4.7}$$

where

$$K_\cdot(\boldsymbol{\theta}) = \sum_{i=1}^{m} K_i(\boldsymbol{\theta}),\ \nabla K_\cdot(\boldsymbol{\theta}) = \frac{\partial K_\cdot(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'},\ \ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \ell_i(\boldsymbol{\theta}),\ \nabla^2 \ell(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'},$$
$$w_i^H(\boldsymbol{\theta}) = K_i(\boldsymbol{\theta}) - \nabla K_\cdot(\boldsymbol{\theta}) \left(\nabla^2 \ell(\boldsymbol{\theta})\right)^{-1} \nabla \ell_i(\boldsymbol{\theta})',$$
$$w_i^O(\boldsymbol{\theta}) = K_i(\boldsymbol{\theta}) - \left(\sum_{i=1}^{m} K_i(\boldsymbol{\theta}) \nabla \ell_i(\boldsymbol{\theta})\right) \left(\sum_{i=1}^{m} \nabla \ell_i(\boldsymbol{\theta})' \nabla \ell_i(\boldsymbol{\theta})\right)^{-1} \nabla \ell_i(\boldsymbol{\theta})'.$$

Here the variance of the score statistic is estimated by replacing the expectations in (3.4.4) by their corresponding sample moments. Note that these formulations are valid for any point process with time-invariant covariate intensity function (1.2.4). In particular, the expected variance of the test statistic for Poisson processes with Weibull intensity can be found from the following expectations:

$$\mathrm{E}[K_i(\boldsymbol{\theta})^2] = \Lambda_i(\boldsymbol{\theta}) + 2\Lambda_i(\boldsymbol{\theta})^2,$$
$$\mathrm{E}(\nabla K_\cdot(\boldsymbol{\theta})) = -\sum_{i=1}^{m} \Lambda_i(\boldsymbol{\theta})(\boldsymbol{x}_i' \ \log \tau_i),$$
$$\mathrm{E}(\nabla^2 \ell(\boldsymbol{\theta})) = \sum_{i=1}^{m} \begin{pmatrix} \Lambda_i(\boldsymbol{\theta}) \boldsymbol{x}_i \boldsymbol{x}_i' & \Lambda_i(\boldsymbol{\theta})(\log \tau_i) \boldsymbol{x}_i \\ \Lambda_i(\boldsymbol{\theta})(\log \tau_i) \boldsymbol{x}_i' & \gamma^{-2} \Lambda_i(\boldsymbol{\theta}) + \Lambda_i(\boldsymbol{\theta})(\log \tau_i)^2 \end{pmatrix},$$

where $\Lambda_i(\boldsymbol{\theta}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\tau_i^{\gamma}$ and $\boldsymbol{x}_i = (1, x_i)'$. However, there is no closed form expression for the expected variance of the test statistic for renewal processes.

We carry out the simulation study as follows. The sample sizes are chosen as 20, 50 (small), 100, 200 (moderate), and 300 (large). We generate a process from each design and compute the test statistic. This procedure is repeated 2,000 times, based on which an empirical distribution of the test statistic is obtained and compared to its asymptotic distribution. It suffices to compare their tail probabilities as the test is either a one-sided test or a Chi-square test for which the squared statistic is used in the latter. The result will be discussed in the next section.

On the other hand, we introduce a random effect $\log(v)$ with different variances into the linear predictor in order to study the power of the test. The mixing distributions considered are Bernoulli and gamma. As will be seen in the next section, the finite sample distributions do not differ much whether the covariate is distributed as Bernoulli(0.5) or standard normal. Therefore, we will consider only the designs with normal distributed covariates and the sample size is taken as 100. The rates of rejections at 5% significant level are computed for different variances. The result is given in section 3.4.3.

3.4.2 Finite Sample Distributions

Poisson Process

The empirical tail probabilities of the test statistics T_a for $a = E, S, H$ and O are tabulated for designs P1 and P2 along with the tail probabilities of $N(0,1)$ distribution in Tables 3.2 and 3.3. For small sample sizes (20 and 50), none of the finite sample distributions of T_E can be approximated by $N(0,1)$. The empirical tail probabilities of T_E for the 1% to 10% significance levels appear reasonably well approximated by $N(0,1)$ for sample size of at least 100. The distributions of other statistics are far away from $N(0,1)$ even for large sample sizes in which they shift to the left of $N(0,1)$. As a result, the type I error is

m	mean	variance	Upper tail probabilities of $N(0,1)$							
	0.0	1.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
20	-0.317	0.877	0.185	0.152	0.118	0.088	0.057	0.029	0.020	0.009
20	-0.656	1.489	0.168	0.128	0.090	0.053	0.030	0.006	0.002	0.000
20	-0.801	1.872	0.181	0.145	0.111	0.078	0.046	0.021	0.007	0.001
20	-0.845	2.113	0.181	0.137	0.102	0.064	0.030	0.009	0.002	0.000
50	-0.210	0.925	0.208	0.172	0.142	0.102	0.069	0.038	0.022	0.012
50	-0.412	1.420	0.231	0.193	0.145	0.096	0.052	0.018	0.004	0.001
50	-0.388	1.409	0.236	0.188	0.142	0.098	0.061	0.020	0.006	0.001
50	-0.479	1.557	0.217	0.172	0.125	0.085	0.044	0.012	0.005	0.002
100	-0.106	1.007	0.249	0.210	0.174	0.135	0.095	0.049	0.026	0.011
100	-0.354	1.216	0.218	0.173	0.128	0.094	0.049	0.018	0.006	0.002
100	-0.306	1.271	0.240	0.197	0.156	0.104	0.064	0.022	0.008	0.002
100	-0.337	1.233	0.230	0.178	0.133	0.085	0.049	0.015	0.004	0.001
200	-0.054	1.000	0.273	0.226	0.183	0.133	0.089	0.046	0.026	0.015
200	-0.161	1.101	0.259	0.210	0.164	0.112	0.073	0.028	0.012	0.003
200	-0.204	1.197	0.261	0.210	0.157	0.112	0.073	0.032	0.014	0.003
200	-0.238	1.133	0.248	0.199	0.140	0.093	0.056	0.021	0.010	0.003
300	-0.045	1.018	0.282	0.240	0.198	0.150	0.103	0.053	0.026	0.012
300	-0.219	1.100	0.251	0.205	0.157	0.111	0.066	0.026	0.011	0.004
300	-0.145	1.154	0.283	0.234	0.178	0.122	0.073	0.035	0.021	0.006
300	-0.201	1.053	0.257	0.207	0.153	0.104	0.066	0.023	0.011	0.005

Table 3.2: Upper tail probabilities, sample mean and sample variance of the test statistics for design P1. First row: The efficient score statistic. Second row: The Hessian form of the IM test. Third row: The OPG form of the IM test. Fourth row: The score statistic.

m	mean	variance	Upper tail probabilities of $N(0,1)$							
	0.0	1.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
20	-0.367	0.783	0.131	0.108	0.091	0.070	0.052	0.032	0.020	0.010
20	-0.709	1.257	0.150	0.118	0.086	0.048	0.024	0.005	0.000	0.000
20	-0.857	1.846	0.175	0.139	0.110	0.080	0.050	0.018	0.008	0.001
20	-0.911	2.290	0.180	0.146	0.109	0.064	0.028	0.011	0.004	0.002
50	-0.252	0.854	0.191	0.149	0.121	0.092	0.060	0.033	0.020	0.008
50	-0.521	1.325	0.194	0.161	0.120	0.070	0.038	0.012	0.003	0.002
50	-0.552	1.605	0.211	0.173	0.134	0.098	0.060	0.022	0.008	0.000
50	-0.611	1.786	0.199	0.154	0.119	0.074	0.040	0.011	0.001	0.000
100	-0.165	0.989	0.219	0.186	0.156	0.118	0.082	0.054	0.033	0.016
100	-0.390	1.271	0.222	0.175	0.138	0.098	0.053	0.015	0.005	0.001
100	-0.373	1.383	0.242	0.196	0.148	0.098	0.063	0.020	0.008	0.001
100	-0.432	1.407	0.220	0.166	0.125	0.081	0.044	0.012	0.005	0.001
200	-0.116	1.036	0.246	0.208	0.170	0.128	0.088	0.051	0.030	0.014
200	-0.282	1.093	0.226	0.180	0.136	0.096	0.055	0.016	0.007	0.002
200	-0.271	1.242	0.246	0.198	0.152	0.101	0.065	0.023	0.012	0.002
200	-0.258	1.261	0.257	0.214	0.163	0.109	0.063	0.016	0.005	0.001
300	-0.090	0.983	0.267	0.218	0.173	0.123	0.084	0.044	0.024	0.011
300	-0.224	1.112	0.247	0.206	0.161	0.108	0.061	0.024	0.008	0.002
300	-0.245	1.217	0.256	0.208	0.156	0.114	0.070	0.026	0.010	0.001
300	-0.248	1.159	0.245	0.194	0.144	0.099	0.057	0.021	0.008	0.003

Table 3.3: Upper tail probabilities, sample mean and sample variance of the test statistics for design P2.

m	mean	variance	Upper tail probabilities of $\chi^2(1)$							
	1.0	2.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
20	0.976	1.588	0.321	0.259	0.199	0.136	0.073	0.033	0.012	0.007
20	1.918	8.092	0.407	0.371	0.326	0.274	0.223	0.169	0.125	0.080
20	2.512	11.49	0.485	0.442	0.403	0.355	0.303	0.220	0.174	0.131
20	2.826	19.46	0.469	0.420	0.382	0.337	0.286	0.224	0.179	0.146
50	0.969	1.962	0.300	0.249	0.187	0.137	0.086	0.045	0.025	0.008
50	1.589	6.407	0.376	0.328	0.279	0.231	0.179	0.113	0.081	0.050
50	1.559	6.198	0.375	0.327	0.277	0.221	0.169	0.110	0.077	0.048
50	1.785	11.16	0.369	0.323	0.265	0.224	0.175	0.125	0.095	0.066
100	1.017	1.860	0.314	0.268	0.221	0.164	0.099	0.046	0.022	0.007
100	1.341	5.019	0.349	0.294	0.242	0.196	0.146	0.089	0.059	0.034
100	1.364	5.179	0.344	0.297	0.244	0.200	0.148	0.091	0.060	0.032
100	1.346	5.699	0.327	0.280	0.234	0.179	0.131	0.085	0.060	0.038
200	1.003	2.145	0.288	0.250	0.202	0.149	0.103	0.050	0.029	0.011
200	1.127	3.489	0.300	0.259	0.211	0.159	0.116	0.065	0.038	0.022
200	1.238	4.023	0.318	0.276	0.229	0.181	0.134	0.081	0.047	0.028
200	1.189	3.969	0.314	0.269	0.215	0.170	0.125	0.071	0.045	0.024
300	1.020	1.910	0.318	0.266	0.209	0.150	0.104	0.049	0.022	0.011
300	1.148	3.489	0.314	0.266	0.215	0.173	0.122	0.071	0.038	0.017
300	1.175	3.052	0.328	0.271	0.222	0.171	0.123	0.076	0.041	0.019
300	1.092	2.546	0.315	0.270	0.221	0.162	0.107	0.063	0.035	0.014

Table 3.4: Upper tail probabilities, sample mean and sample variance of the squared statistics for design P1.

m	mean	variance	Upper tail probabilities of $\chi^2(1)$							
	1.0	2.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
20	0.917	1.629	0.302	0.240	0.182	0.119	0.069	0.030	0.012	0.007
20	1.759	5.304	0.429	0.386	0.343	0.287	0.237	0.152	0.099	0.055
20	2.580	10.40	0.529	0.478	0.433	0.375	0.317	0.243	0.184	0.131
20	3.118	26.48	0.490	0.443	0.399	0.359	0.314	0.257	0.204	0.156
50	0.917	1.664	0.291	0.237	0.176	0.124	0.074	0.036	0.015	0.007
50	1.596	5.926	0.376	0.326	0.287	0.231	0.182	0.121	0.086	0.051
50	1.909	8.646	0.435	0.380	0.336	0.274	0.217	0.146	0.104	0.068
50	2.159	15.38	0.405	0.362	0.316	0.261	0.217	0.157	0.120	0.088
100	1.016	2.138	0.305	0.255	0.205	0.156	0.107	0.051	0.024	0.011
100	1.422	4.845	0.364	0.310	0.261	0.212	0.153	0.100	0.068	0.040
100	1.521	6.088	0.368	0.324	0.276	0.223	0.166	0.104	0.070	0.045
100	1.593	8.765	0.355	0.308	0.262	0.208	0.154	0.107	0.081	0.053
200	1.049	2.431	0.308	0.262	0.214	0.161	0.111	0.054	0.025	0.013
200	1.172	3.213	0.326	0.280	0.237	0.180	0.125	0.071	0.046	0.021
200	1.315	4.493	0.330	0.286	0.240	0.198	0.144	0.094	0.052	0.033
200	1.327	5.481	0.334	0.287	0.239	0.184	0.132	0.079	0.054	0.035
300	0.991	2.057	0.290	0.243	0.194	0.147	0.094	0.050	0.024	0.011
300	1.162	3.656	0.311	0.257	0.215	0.171	0.126	0.070	0.040	0.020
300	1.276	4.105	0.337	0.284	0.243	0.192	0.134	0.087	0.051	0.026
300	1.221	4.483	0.306	0.261	0.216	0.169	0.120	0.077	0.055	0.034

Table 3.5: Upper tail probabilities, sample mean and sample variance of the squared statistics for design P2.

m	mean	variance	Upper tail probabilities of $\chi^2(1)$							
	1.0	2.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
100	1.019	2.015	0.305	0.251	0.206	0.154	0.100	0.053	0.023	0.010
100	1.359	5.147	0.338	0.293	0.250	0.202	0.145	0.094	0.059	0.035
100	1.330	4.650	0.339	0.288	0.240	0.196	0.146	0.087	0.057	0.035
100	1.382	5.511	0.350	0.296	0.249	0.198	0.141	0.091	0.054	0.039
200	0.962	1.867	0.295	0.243	0.201	0.144	0.098	0.042	0.017	0.008
200	1.218	3.469	0.337	0.282	0.236	0.181	0.130	0.078	0.047	0.022
200	1.090	2.937	0.307	0.258	0.203	0.157	0.105	0.062	0.035	0.020
200	1.164	3.729	0.303	0.262	0.212	0.168	0.115	0.074	0.046	0.026
300	1.012	2.139	0.302	0.253	0.198	0.141	0.099	0.049	0.024	0.011
300	1.135	3.110	0.318	0.268	0.220	0.166	0.114	0.069	0.041	0.022
300	1.134	3.103	0.321	0.269	0.223	0.161	0.115	0.063	0.040	0.020
300	1.116	3.078	0.314	0.269	0.217	0.164	0.109	0.058	0.038	0.022

Table 3.6: Upper tail probabilities, sample mean and sample variance of the squared statistics for design P3.

deflated too much.

On the other hand, the squared statistics seem to converge a little faster to the asymptotic distribution ($\chi^2(1)$) (Tables 3.4 and 3.5), although the distributions for the sample moment variance forms now shift to the right of $\chi^2(1)$, implying that the type I error is inflated. The ES form converges relatively quickly; a sample size of 100 is enough to get a reasonable asymptotic approximation. However, a sample size of at least 200 is required for the other statistics.

The ES form certainly out-performs the other statistics. This may be due to the fact that the expected variance of the statistic is less variable than its finite sample approximations. On the other hand, the score statistic, the Hessian form and the OPG form of the IM test have a similar finite sample distribution for all sample sizes. It is difficult to specify a preferred test.

m	mean	variance	Upper tail probabilities of $\chi^2(1)$							
	1.0	2.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
100	0.942	1.944	0.282	0.234	0.189	0.142	0.090	0.039	0.019	0.008
100	1.333	5.052	0.341	0.296	0.245	0.193	0.137	0.087	0.057	0.030
100	1.515	5.700	0.385	0.330	0.277	0.226	0.167	0.109	0.074	0.046
100	1.536	6.856	0.364	0.318	0.272	0.219	0.168	0.110	0.074	0.048
200	0.964	1.994	0.282	0.230	0.185	0.144	0.092	0.047	0.026	0.012
200	1.173	3.456	0.314	0.265	0.221	0.175	0.128	0.076	0.045	0.025
200	1.296	4.424	0.348	0.298	0.239	0.193	0.139	0.077	0.048	0.028
200	1.257	4.535	0.327	0.279	0.231	0.187	0.131	0.073	0.045	0.029
300	1.035	2.114	0.312	0.271	0.218	0.159	0.100	0.054	0.026	0.009
300	1.154	3.135	0.316	0.269	0.219	0.175	0.121	0.075	0.047	0.020
300	1.254	4.099	0.344	0.292	0.240	0.178	0.128	0.077	0.050	0.027
300	1.216	3.958	0.323	0.280	0.239	0.191	0.135	0.075	0.046	0.025

Table 3.7: Upper tail probabilities, sample mean and sample variance of the squared statistics for design P4.

Furthermore, the distributions of the test statistics seem not to be affected by whether the covariate is distributed as Bernoulli or normal, or the number of events per process provided the number is not too small (Tables 3.6 and 3.7).

Renewal Process

The finite sample distributions are tabulated in Tables 3.8 to 3.13 for T_S, T_H and T_O. Recall that the ES form for renewal process is not available. We start with moderate sample sizes because the finite sample distributions are far away from the asymptotic distributions for small sample sizes.

The results are very similar to the Poisson process.

m	mean	variance	Upper tail probabilities of $N(0,1)$							
	0.0	1.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
100	-0.241	1.204	0.263	0.212	0.153	0.106	0.061	0.021	0.006	0.003
100	-0.316	1.393	0.242	0.200	0.162	0.118	0.071	0.026	0.009	0.003
100	-0.340	1.290	0.244	0.184	0.142	0.103	0.056	0.021	0.007	0.002
200	-0.133	1.072	0.268	0.222	0.175	0.124	0.076	0.026	0.014	0.003
200	-0.233	1.189	0.251	0.201	0.159	0.122	0.067	0.030	0.012	0.004
200	-0.208	1.156	0.262	0.217	0.166	0.114	0.068	0.025	0.013	0.003
300	-0.177	1.056	0.258	0.204	0.156	0.106	0.069	0.025	0.011	0.004
300	-0.195	1.118	0.258	0.211	0.164	0.116	0.074	0.029	0.008	0.001
300	-0.137	1.049	0.265	0.215	0.169	0.114	0.071	0.029	0.016	0.007

Table 3.8: Upper tail probabilities, sample mean and sample variance of the test statistics for design R1. First row: The Hessian form of the IM test. Second row: The OPG form of the IM test. Third row: The score statistic.

m	mean	variance	Upper tail probabilities of $N(0,1)$							
	0.0	1.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
100	-0.325	1.202	0.229	0.188	0.146	0.104	0.058	0.021	0.007	0.002
100	-0.300	1.264	0.238	0.193	0.149	0.108	0.061	0.023	0.008	0.003
100	-0.375	1.292	0.224	0.180	0.137	0.088	0.049	0.018	0.006	0.001
200	-0.249	1.163	0.254	0.208	0.156	0.109	0.067	0.022	0.011	0.004
200	-0.218	1.181	0.264	0.204	0.157	0.109	0.069	0.030	0.010	0.003
200	-0.246	1.173	0.254	0.203	0.162	0.108	0.066	0.024	0.009	0.002
300	-0.137	1.111	0.283	0.231	0.184	0.135	0.082	0.029	0.010	0.004
300	-0.168	1.149	0.262	0.222	0.166	0.124	0.076	0.030	0.013	0.004
300	-0.201	1.123	0.250	0.195	0.153	0.113	0.066	0.028	0.014	0.004

Table 3.9: Upper tail probabilities, sample mean and sample variance of the test statistics for design R2.

m	mean	variance	Upper tail probabilities of $\chi^2(1)$							
	1.0	2.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
100	1.261	4.040	0.330	0.284	0.233	0.182	0.133	0.080	0.054	0.034
100	1.493	6.296	0.366	0.320	0.275	0.213	0.159	0.100	0.068	0.042
100	1.406	5.591	0.356	0.301	0.251	0.204	0.150	0.093	0.060	0.040
200	1.089	2.964	0.310	0.253	0.203	0.147	0.102	0.063	0.032	0.021
200	1.243	3.919	0.347	0.290	0.235	0.190	0.131	0.075	0.042	0.024
200	1.199	3.192	0.338	0.283	0.236	0.185	0.126	0.075	0.037	0.020
300	1.087	2.812	0.304	0.256	0.207	0.159	0.113	0.059	0.033	0.018
300	1.156	3.032	0.321	0.275	0.232	0.181	0.132	0.068	0.039	0.019
300	1.054	2.596	0.300	0.255	0.200	0.150	0.102	0.059	0.030	0.017

Table 3.10: Upper tail probabilities, sample mean and sample variance of the squared statistics for design R1.

m	mean	variance	Upper tail probabilities of $\chi^2(1)$							
	1.0	2.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
100	1.307	3.991	0.352	0.307	0.263	0.210	0.148	0.085	0.049	0.026
100	1.353	4.735	0.351	0.301	0.249	0.200	0.150	0.092	0.058	0.035
100	1.432	6.127	0.346	0.302	0.252	0.208	0.156	0.094	0.061	0.037
200	1.224	3.744	0.338	0.283	0.235	0.175	0.120	0.072	0.043	0.024
200	1.228	3.845	0.324	0.273	0.231	0.181	0.132	0.078	0.049	0.027
200	1.233	3.697	0.332	0.285	0.240	0.180	0.126	0.080	0.043	0.027
300	1.129	2.854	0.329	0.280	0.226	0.169	0.109	0.061	0.034	0.019
300	1.176	3.304	0.329	0.282	0.233	0.184	0.127	0.071	0.045	0.022
300	1.163	3.252	0.323	0.275	0.219	0.173	0.122	0.073	0.045	0.021

Table 3.11: Upper tail probabilities, sample mean and sample variance of the squared statistics for design R2.

m	mean	variance	Upper tail probabilities of $\chi^2(1)$							
	1.0	2.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
100	1.197	3.531	0.325	0.274	0.224	0.173	0.124	0.073	0.042	0.023
100	1.355	4.578	0.356	0.303	0.257	0.200	0.151	0.095	0.059	0.031
100	1.534	7.006	0.357	0.305	0.259	0.214	0.166	0.106	0.074	0.048
200	1.123	3.046	0.318	0.276	0.224	0.167	0.112	0.062	0.036	0.017
200	1.150	3.041	0.318	0.275	0.223	0.174	0.122	0.068	0.040	0.019
200	1.182	3.594	0.313	0.265	0.209	0.164	0.125	0.073	0.048	0.027
300	1.077	2.584	0.314	0.263	0.208	0.159	0.103	0.057	0.031	0.016
300	1.199	3.448	0.323	0.273	0.227	0.176	0.126	0.073	0.049	0.028
300	1.139	3.256	0.316	0.264	0.215	0.169	0.121	0.062	0.034	0.020

Table 3.12: Upper tail probabilities, sample mean and sample variance of the squared statistics for design R3.

m	mean	variance	Upper tail probabilities of $\chi^2(1)$							
	1.0	2.0	0.30	0.25	0.20	0.15	0.10	0.05	0.025	0.01
100	1.232	3.867	0.320	0.272	0.223	0.184	0.130	0.078	0.050	0.030
100	1.409	5.342	0.364	0.303	0.252	0.198	0.148	0.097	0.061	0.036
100	1.626	7.410	0.380	0.332	0.280	0.231	0.181	0.122	0.080	0.050
200	1.164	2.999	0.322	0.274	0.228	0.180	0.130	0.080	0.042	0.018
200	1.331	3.764	0.360	0.314	0.264	0.208	0.152	0.086	0.052	0.028
200	1.358	5.080	0.338	0.293	0.247	0.199	0.143	0.087	0.060	0.037
300	1.135	2.600	0.327	0.280	0.231	0.185	0.126	0.062	0.033	0.013
300	1.208	3.223	0.338	0.285	0.243	0.192	0.130	0.064	0.041	0.020
300	1.240	4.217	0.327	0.274	0.226	0.182	0.130	0.078	0.045	0.025

Table 3.13: Upper tail probabilities, sample mean and sample variance of the squared statistics for design R4.

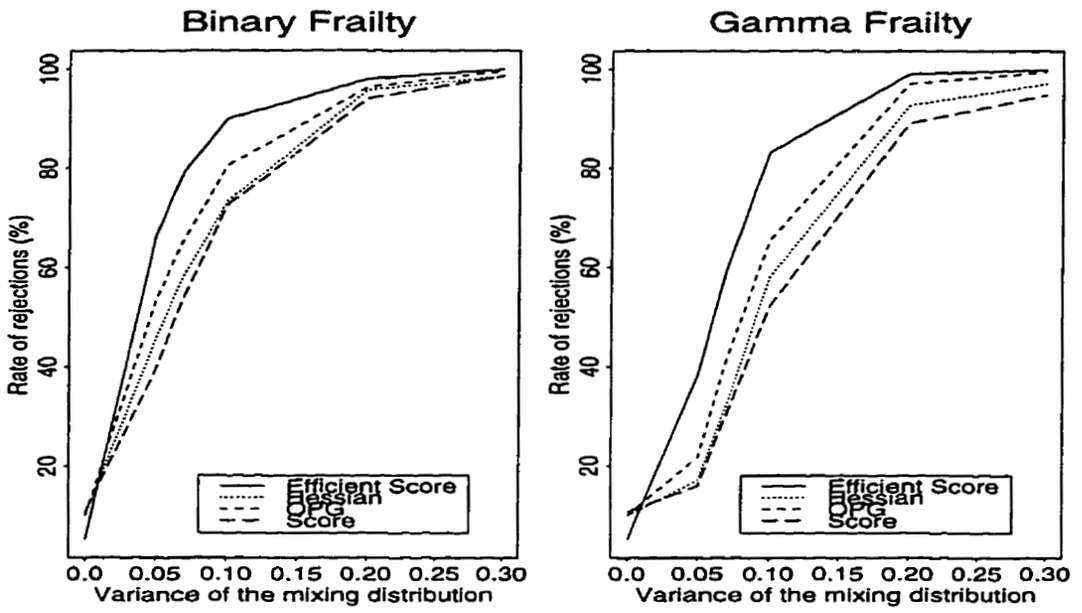

Figure 3.1: Power curves for Poisson process, design P2

3.4.3 Power of the Tests

Power curves at 5% significance level for different squared statistics under designs P2 and R2 with sample size 100 are displayed in Figures 3.1 to 3.2 for binary and gamma frailties.

Strictly speaking as type I errors for the statistics are not the same, the power curves are not directly comparable. Despite this, since the test using the ES form has good type I error rate and its power curve is definitely higher than the other power curves, this should indicate that it has higher power than the other tests. On the other hand, since the type I errors for T_H, T_O and T_S are very close, this also makes the comparison between them reasonable. In both Poisson and renewal processes, the powers of T_H, T_O and T_S are very close for binary frailty. For gamma frailty, we observe the following order: $T_O > T_H > T_S$, although the difference is quite small.

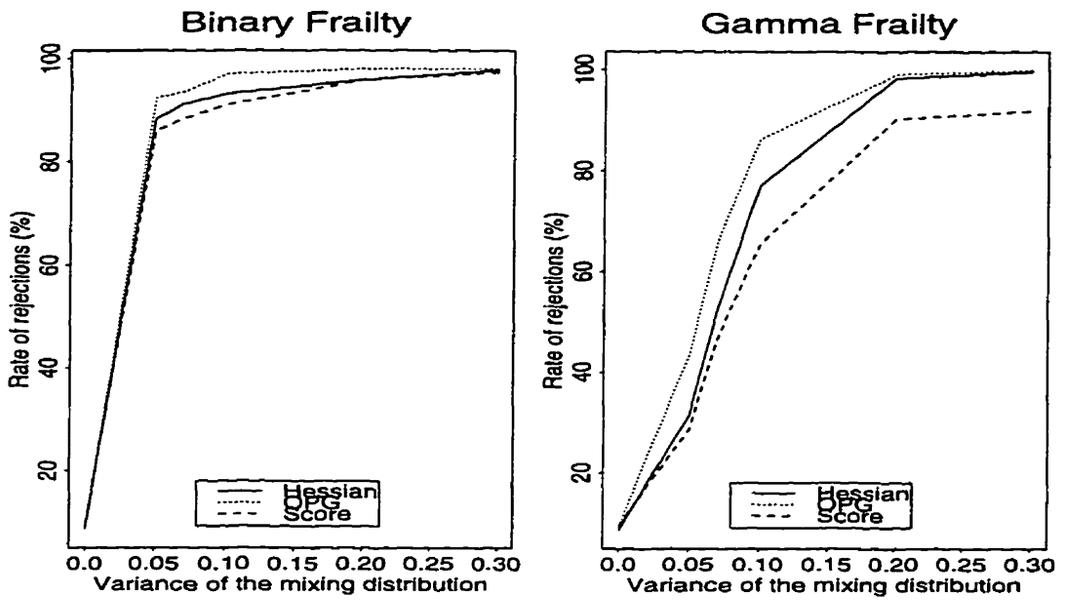

Figure 3.2: Power curves for renewal process, design R2

3.4.4 Summary

We summarize the results of the simulation study:

1. The ES statistic performs the best in terms of size and power. A sample size of about 100 is enough to achieve reasonable asymptotic approximation. It is recommended that the ES statistic should be used whenever it is available.

2. The sample moment statistics have similar performance. This is consistent with Crouchley and Pickles (1993)'s results. This makes the OPG form more attractive because of its simpler computation.

3. Large sample sizes are required for the statistics based on sample moment estimates of the variance in order to get a reasonable type I error rate.

4. Apart from the ES statistic, the statistics have smaller type I error rates than the nominal type I error while the squared statistics have higher type I error rates. As

a result, the one-sided and two-sided tests tend to reject too infrequently and too frequently, respectively, when the null hypothesis is true.

3.5 Adjusted Score Tests for Poisson Processes

The simulation study in the last section clearly shows that the score and IM tests of homogeneity have poor finite sample performance. Adjustments for the test statistics are definitely desirable. In this section, we examine the frequency properties of score and adjusted score statistics for testing the hypothesis of homogeneity in continuous time non-homogeneous Poisson processes subject to non-informative censoring. Specifically, given the random effect u, we express the conditional intensity for events at time t as

$$\lambda(t|u) = \exp(\boldsymbol{x}'\boldsymbol{\beta} + \sigma u)\lambda_0(t), \tag{3.5.1}$$

where $\boldsymbol{x} = (x_1, ..., x_p)'$ is a $p \times 1$ vector of time independent covariates, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ is a $p \times 1$ vector of the corresponding regression coefficients, $\lambda_0(t)$ is the baseline intensity function, u is the subject-specific random effect with mean 0, variance 1 and probability density function $g(u)$, and σ measures the extent of heterogeneity. Based on this specification, it is clear that we assume the random effect to be distributed independently of the covariates. A test of homogeneity is equivalent to testing $H_0 : \sigma^2 = 0$.

We first consider an adjustment for parametric models in section 3.5.1. We then extend the test to semi-parametric models based on counting process methodology (Fleming and Harrington, 1991) in section 3.5.2. Simulation studies pertaining to the size and power of the tests are described in section 3.5.3. These simulation studies also compare the performance of the tests using a weakly parametric model based on a piecewise constant baseline intensity and the semi-parametric model. An example involving a clinical study of gamma interferon in chronic granulomatous disease is provided in section 3.5.4.

3.5.1 Adjustment for Parametric Models

Here we assume that the setting given in section 3.2 is satisfied. We also assume that the conditional baseline intensity, $\lambda_0(t)$, in (3.5.1) is completely specified by γ, a $q \times 1$ vector of parameters such that $\lambda_0(t)$ is at least twice differentiable with respect to γ, and satisfies the condition

$$\Lambda_0(t) = \int_0^t \lambda_0(s)ds < \infty, \qquad 0 \leq t < \infty.$$

We have shown in section 3.4 that the score statistic under this specification is given by

$$T_p(\hat{\theta}) = \frac{1}{2}\sum_{i=1}^{m}\left[\left(n_i - \Lambda_i(\hat{\theta})\right)^2 - \Lambda_i(\hat{\theta})\right], \tag{3.5.2}$$

where $\theta = (\beta', \gamma')'$, $\Lambda_i(\theta) = \exp(x_i'\beta)\Lambda_0(\tau_i)$ and $\hat{\theta}$ is the m.l.e. of θ under the null model. The asymptotic variance of $T_p(\hat{\theta})$ is given by

$$V_p(\theta) = I_\sigma(\theta) - I'_{\theta\sigma}(\theta)I_\theta^{-1}(\theta)I_{\theta\sigma}(\theta), \tag{3.5.3}$$

where $I_\sigma(\theta) = \frac{1}{4}\sum_{i=1}^{m}(\Lambda_i(\theta) + 2\Lambda_i^2(\theta))$, $I_{\theta\sigma}(\theta) = \mathrm{E}(-\partial T_p(\theta)/\partial\theta)$ and $I_\theta(\theta)$ is the expected information matrix for the null model. The expressions for $I_{\theta\sigma}(\theta)$ and $I_\theta(\theta)$ are given in appendix A.1. We carry out the test by using the standardized statistic

$$\hat{Z}_p = \frac{T_p(\hat{\theta})}{\sqrt{V_p(\hat{\theta})}},$$

which is asymptotically distributed as $N(0,1)$ under H_0. Since σ^2 is non-negative, we consider a one-sided test in which the null hypothesis is rejected if $\hat{Z}_p$ is large.

We are now going to derive an approximate expectation of $T_p(\hat{\theta})$. Consider a first-order

Taylor series expansion for $n_i - \Lambda_i(\hat{\boldsymbol{\theta}})$ about the true parameter value $\boldsymbol{\theta}$,

$$\begin{aligned} n_i - \Lambda_i(\hat{\boldsymbol{\theta}}) &= n_i - \Lambda_i(\boldsymbol{\theta}) - \frac{\partial \Lambda_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + O_p(m^{-1}) \\ &= n_i - \Lambda_i(\boldsymbol{\theta}) - \frac{\partial \Lambda_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} I_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}) U(\boldsymbol{\theta}) + O_p(m^{-1}), \end{aligned} \tag{3.5.4}$$

where $U(\boldsymbol{\theta})$ is the score function for $\boldsymbol{\theta}$ under the null model. Since $U(\boldsymbol{\theta})$ has mean zero and variance $I_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, the left-hand side of (3.5.4) has expectation approximately equal to 0 and variance given by

$$\Lambda_i(\boldsymbol{\theta}) + \frac{\partial \Lambda_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} I_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}) \frac{\partial \Lambda_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - 2 \frac{\partial \Lambda_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} I_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}) \mathrm{cov}(n_i, U(\boldsymbol{\theta})).$$

We show in appendix A.1 that $\mathrm{cov}(n_i, U(\boldsymbol{\theta})) = \partial \Lambda_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ and hence, the expectation of the score statistic evaluated at $\hat{\boldsymbol{\theta}}$ is approximately equal to

$$\mathrm{E}(T_p(\hat{\boldsymbol{\theta}})) \approx \frac{1}{2} \sum_{i=1}^{m} \Lambda_i(\boldsymbol{\theta}) - b_p(\boldsymbol{\theta}) - \frac{1}{2} \sum_{i=1}^{m} \Lambda_i(\boldsymbol{\theta}),$$

as $m^{-1/2}(\Lambda_i(\hat{\boldsymbol{\theta}}) - \Lambda_i(\boldsymbol{\theta}))$ converges to 0 in probability, where the approximate expected bias is given by

$$b_p(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{m} \frac{\partial \Lambda_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} I_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}) \frac{\partial \Lambda_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{3.5.5}$$

Since $I_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ has order $O(m)$ and $\partial \Lambda_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ has order $O(1)$, the normalized expected bias, $m^{-1/2} b_p(\boldsymbol{\theta})$, will vanish as $m \longrightarrow \infty$, giving $m^{-1/2} T_p(\hat{\boldsymbol{\theta}})$ and $m^{-1/2} T_p(\boldsymbol{\theta})$ the same asymptotic distribution. For small to moderate sample sizes, this bias may be appreciable, so we define the adjusted score statistic as

$$T_p^A(\hat{\boldsymbol{\theta}}) = T_p(\hat{\boldsymbol{\theta}}) + b_p(\hat{\boldsymbol{\theta}}), \tag{3.5.6}$$

and carry out the test based on the standardized version of the statistic,

$$\hat{Z}_p^A = \frac{T_p^A(\hat{\boldsymbol{\theta}})}{\sqrt{V_p(\hat{\boldsymbol{\theta}})}}, \tag{3.5.7}$$

which is asymptotically distributed as $N(0,1)$. We anticipate that the distribution of $\hat{Z}_p^A$ converges faster to $N(0,1)$ than that of $\hat{Z}_p$ under the null hypothesis.

We also note that $b_p(\hat{\boldsymbol{\theta}})$ is non-negative since $I_\theta(\hat{\boldsymbol{\theta}})$ is a semi-positive definite matrix with probability 1. As a consequence, the adjusted score statistic is at least as large as the unadjusted score statistic with probability 1. This implies that the finite sample distribution of the unadjusted score statistic is left-shifted compared to its asymptotic distribution. The simulation study in section 3.5.3 provides empirical evidence that the type I error for the unadjusted score statistic is indeed smaller than the nominal level.

3.5.2 Semi-parametric Models

Model Specification

In this section, we employ the counting process approach (Andersen and Gill, 1982; Fleming and Harrington, 1991) for Poisson processes in which $\lambda_0(t)$ is completely unspecified. Using the notation in section 3.5.1, we express the cumulative number of events for subject i at time t, given u_i as,

$$N_i(t) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma u_i) \int_0^t Y_i(s)\lambda_0(s)ds + M_i(t), \tag{3.5.8}$$

where $M_i(t)$ is a zero-mean martingale.

Without loss of generality, we assume that the lengths of follow-up are in ascending

order, $\tau_1 \leq \cdots \leq \tau_m$. Given $\boldsymbol{u} = (u_1, ..., u_m)'$, the conditional partial likelihood is given by

$$L_p(\boldsymbol{\beta}, \sigma|\boldsymbol{u}) = \exp\left[\sum_{i=1}^{m} \int_0^{\infty} \left(\boldsymbol{x}_i'\boldsymbol{\beta} + \sigma u_i - \log\left[\sum_{j=1}^{m} Y_j(s)\exp(\boldsymbol{x}_j'\boldsymbol{\beta} + \sigma u_j)\right]\right) dN_i(s)\right], \tag{3.5.9}$$

$i = 1, ..., m$ (Fleming and Harrington, 1991 Chapter 4). Hence, the marginal partial likelihood is obtained by integrating (3.5.9) with respect to the random effects (Commenges and Andersen, 1995):

$$L_p(\boldsymbol{\beta}, \sigma) = \int_{-\infty}^{\infty} L_p(\boldsymbol{\beta}, \sigma|\boldsymbol{u}) g(\boldsymbol{u}) d\boldsymbol{u}. \tag{3.5.10}$$

Before proceeding, we define some further notation. For $t \geq 0$, let

$$S^{(0)}(t) = \sum_{i=1}^{m} Y_i(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \quad S^{(1)}(t) = \sum_{i=1}^{m} Y_i(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i,$$

$$S^{(2)}(t) = \sum_{i=1}^{m} Y_i(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i\boldsymbol{x}_i', \quad E(t) = \frac{S^{(1)}(t)}{S^{(0)}(t)}, \quad V(t) = \frac{S^{(2)}(t)}{S^{(0)}(t)} - E(t)E(t)',$$

$$\text{and} \qquad w_i(t) = \frac{Y_i(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{S^{(0)}(t)}, \qquad i = 1, ..., m.$$

We assume that the regularity conditions of Fleming and Harrington (1991; chapter 8) hold for the asymptotic arguments below.

Score Statistic

Let $N.(t) = \sum_{i=1}^{m} N_i(t)$ be the total number of events in the sample occurring in $(0, t]$. Denote the martingale residual at t under $H_0 : \sigma^2 = 0$ by

$$\hat{M}_i(t, \boldsymbol{\beta}) = N_i(t) - \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \int_0^t Y_i(s) d\hat{\Lambda}_0(s), \qquad i = 1, ..., m, \tag{3.5.11}$$

where

$$\hat{\Lambda}_0(t) = \int_0^t \frac{dN.(s)}{S^{(0)}(s)}$$

is the Nelson-Aalen estimator for $\Lambda_0(t)$ (Andersen et al., 1993).

As suggested by Commenges and Andersen (1995), the score statistic for testing $\sigma^2 = 0$ has the same form as (3.5.2) except that the partial likelihood (3.5.9) is used in place of (3.2.1). The resulting score statistic is then given by

$$T_{sp}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{m}\left[\hat{M}_i^2(\boldsymbol{\beta}) - \int_0^{\infty}(w_i(s) - w_i^2(s))dN_{\cdot}(s)\right], \tag{3.5.12}$$

where $\hat{M}_i(\boldsymbol{\beta}) = \hat{M}_i(\infty, \boldsymbol{\beta})$ (see appendix A.2 for derivation). The score statistic (3.5.12) has a similar interpretation to its parametric counterpart (3.5.2). By expressing (3.5.11) as $\hat{M}_i(t, \boldsymbol{\beta}) = M_i(t) - \int_0^t w_i(s)dM_{\cdot}(s)$, since the $M_i(\cdot)$'s are independent zero-mean martingales, the predictable variation process of $\hat{M}_i(t, \boldsymbol{\beta})$ is equal to

$$< \hat{M}_i(\cdot, \boldsymbol{\beta}) > (t) = \int_0^t (S^{(0)}(s) - \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))w_i(s)d\Lambda_0(s)$$

which can be estimated by $\int_0^t (w_i(s) - w_i^2(s))dN_{\cdot}(s)$. Therefore, the expectation of $\hat{M}_i^2(\boldsymbol{\beta})$, which is equal to $\mathrm{E}[< \hat{M}_i(\cdot, \boldsymbol{\beta}) > (\infty)]$ (Fleming and Harrington, 1991 section 2.4), can be estimated by the second term on the right-hand side of (3.5.12).

The asymptotic distribution of $T_{sp}(\boldsymbol{\beta})$ can be obtained by noting that $T_{sp}(\boldsymbol{\beta})$ is a martingale transform of the $M_i(\cdot)$'s at ∞. Specifically, we can show that the process

$$T_{sp}(t, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{m}\left[\hat{M}_i^2(t, \boldsymbol{\beta}) - \int_0^t (w_i(s) - w_i^2(s))dN_{\cdot}(s)\right]$$

can be expressed as

$$T_{sp}(t, \boldsymbol{\beta}) = \sum_{i=1}^{m}\int_0^t H_i(s, \boldsymbol{\beta})dM_i(s), \tag{3.5.13}$$

where

$$H_i(t, \boldsymbol{\beta}) = \hat{M}_i(t-, \boldsymbol{\beta}) - w_i(t) - \sum_{j=1}^{m}[\hat{M}_j(t-, \boldsymbol{\beta}) - w_j(t)]w_j(t)$$

is predictable. The proof of (3.5.13) closely follows Commenges and Andersen (1995); the details are given in appendix A.2. The predictable variation process of $T_{sp}(t,\beta)$ is then equal to

$$< T_{sp}(\cdot,\beta) > (t) = \sum_{i=1}^{m} \int_0^t H_i^2(s,\beta) Y_i(s) \exp(x_i'\beta) d\Lambda_0(s),$$

whose expectation is equal to the variance of $T_{sp}(t,\beta)$. Hence, $T_{sp}(\beta) = T_{sp}(\infty,\beta)$ has mean 0 and variance $\mathrm{E}[< T_{sp}(\cdot,\beta) > (\infty)]$ which can be estimated by

$$\hat{I}_\sigma(\beta) = \sum_{i=1}^{m} \int_0^\infty H_i^2(s,\beta) w_i(s) dN_\cdot(s).$$

Under suitable regularity conditions (Gray, 1995) and by the martingale central limit theorem (Fleming and Harrington, 1991 Chapter 5), $T_{sp}(\beta)/\sqrt{\hat{I}_\sigma(\beta)}$ converges to $N(0,1)$ in distribution.

The maximum partial likelihood estimates, $\hat{\beta}$, under H_0 can be obtained using the procedure suggested by Lawless (1987). When the regression parameters are replaced by $\hat{\beta}$, the asymptotic variance of $T_{sp}(\hat{\beta})$ has the same adjustment as in the parametric case (3.5.3). Based on the idea of Theorem 8.3.3 of Fleming and Harrington (1991) with some modifications, one may show that the asymptotic variance of $T_{sp}(\hat{\beta})$ can be estimated by

$$\hat{V}_{sp}(\beta) = \hat{I}_\sigma(\beta) - \hat{I}'_{\beta\sigma}(\beta) \hat{I}_\beta^{-1}(\beta) \hat{I}_{\beta\sigma}(\beta), \tag{3.5.14}$$

where

$$\hat{I}_{\beta\sigma}(\beta) = \sum_{i=1}^{m} \int_0^\infty H_i(s,\beta) w_i(s) x_i dN_\cdot(s)$$

estimates $\mathrm{E}(-\partial T_{sp}(\beta)/\partial\beta)$ and

$$\hat{I}_\beta(\beta) = \int_0^\infty V(s) dN_\cdot(s)$$

is the observed information matrix for β. Hence, the test is based on the standardized

statistic

$$\hat{Z}_{sp} = \frac{T_{sp}(\hat{\beta})}{\sqrt{\hat{V}_{sp}(\hat{\beta})}},$$

which converges to $N(0,1)$ in distribution.

Adjusted Score Statistic

As in the parametric case, we consider a first-order Taylor series expansion for the martingale residuals around the true β, which is equal to

$$\hat{M}_i(\hat{\beta}) = \hat{M}_i(\beta) + J_i'(\beta)(\hat{\beta} - \beta) + O_p(m^{-1}), \qquad i = 1, ..., m,$$

where

$$J_i(\beta) = \frac{\partial \hat{M}_i(\beta)}{\partial \beta} = -\int_0^\infty (x_i - E(s))w_i(s)dN_\cdot(s),$$

which can be decomposed into two terms

$$J_i(\beta) = -\int_0^\infty (x_i - E(s))w_i(s)dM_\cdot(s) - \int_0^\infty (x_i - E(s))Y_i(s)\exp(x_i'\beta)d\Lambda_0(s). \quad (3.5.15)$$

The first term of (3.5.15) is a $p \times 1$ vector of zero-mean martingales with predictable variation $\int_0^\infty (x_i - E(s))(x_i - E(s))'w_i^2(s)S^{(0)}(s)d\Lambda_0(s)$ which, subject to the regularity conditions in Chapter 8 of Fleming and Harrington (1991), converges to 0 in probability. The second term of (3.5.15) is of order $O_p(1)$ and thus $J_i(\beta)$ converges to $\mathrm{E}[J_i(\beta)] = -\int_0^\infty \mathrm{E}[(x_i - E(s))Y_i(s)\exp(x_i'\beta)]d\Lambda_0(s)$ in probability. Furthermore, using the relation $\hat{\beta} - \beta = V_\beta U(\beta) + O_p(m^{-1})$, we have the following approximation:

$$\mathrm{E}[\hat{M}_i^2(\hat{\beta})] \approx \mathrm{E}[\hat{M}_i^2(\beta)] + \mathrm{E}[J_i'(\beta)]V_\beta \mathrm{E}[J_i(\beta)] + 2\mathrm{E}[J_i'(\beta)]V_\beta \mathrm{cov}(\hat{M}_i(\beta), U(\beta)),$$

where $V_\beta = [\mathrm{E}(\hat{I}_\beta(\beta))]^{-1}$ and $U(\beta)$ is the score function for β under H_0. Similar to the parametric case, one can prove that $\mathrm{cov}(\hat{M}_i(\beta), U(\beta)) = -\mathrm{E}[J_i(\beta)]$ (see appendix A.2). It is straightforward to see that the bias induced in the second term of $T_{sp}(\hat{\beta})$ is negligible. Consequently, the expectation of $T_{sp}(\hat{\beta})$ is approximately equal to $-b_{sp}(\beta)$ where

$$b_{sp}(\beta) = \frac{1}{2}\sum_{i=1}^{m} \mathrm{E}[J_i'(\beta)]V_\beta \mathrm{E}[J_i(\beta)],$$

which is estimated by

$$\hat{b}_{sp}(\beta) = \frac{1}{2}\sum_{i=1}^{m} J_i'(\beta)\hat{I}_\beta^{-1}(\beta)J_i(\beta), \tag{3.5.16}$$

for true β. We therefore define the adjusted score statistic as

$$T_{sp}^A(\hat{\beta}) = T_{sp}(\hat{\beta}) + \hat{b}_{sp}(\hat{\beta}). \tag{3.5.17}$$

We note that the estimated bias $\hat{b}_{sp}(\hat{\beta})$ has the same properties as its parametric counterpart, namely it is non-negative and $m^{-1/2}\hat{b}_{sp}(\hat{\beta})$ converges to 0 in probability. Hence, the unadjusted and adjusted score statistics are asymptotically equivalent.

We carry out the test using the standardized statistic

$$\hat{Z}_{sp}^A = \frac{T_{sp}^A(\hat{\beta})}{\sqrt{\hat{V}_{sp}(\hat{\beta})}},$$

which is asymptotically distributed as $N(0,1)$.

3.5.3 Simulation Studies

Preliminary Remarks

In this section, we compare the performance of the score and adjusted score statistics with respect to the size and power via simulations in both the parametric and semi-parametric

frameworks. First we make some general remarks to aid in the interpretation of the results.

The score test based on parametric models may be sensitive to the misspecification of the baseline intensity. Although the semi-parametric model is robust in this regard, the performance of the corresponding score test may be adversely affected by the use of an empirical estimator for $\text{var}(T_{sp}(\hat{\beta}))$ based on sample moments. The simulation study in section 3.4 provides empirical evidence of this regard. Together with the estimation of β, (3.5.14) can be highly unstable in small samples (Breslow, 1989). As a result, the distributions of $\hat{Z}_{sp}$ and $\hat{Z}_{sp}^{A}$ may require large sample sizes to approach $N(0,1)$. In contrast, in parametric models, the asymptotic variance of $T_p(\hat{\beta})$ (3.5.3) has an exact form for given β, and therefore $\hat{Z}_p$ and $\hat{Z}_p^A$ will be expected to converge faster to $N(0,1)$ than $\hat{Z}_{sp}$ and $\hat{Z}_{sp}^{A}$, provided the parametric model is correctly specified. As a compromise between a particular parametric model and the semi-parametric model, we suggest a piecewise exponential specification for the baseline intensity. As the number of pieces increases the model becomes weakly parametric, and so will exhibit robustness.

The objectives of this section are two-fold. First, we intend to compare the frequency properties of the unadjusted and adjusted score statistics. Second, we plan to compare the performance of the tests based on the piecewise exponential and the semi-parametric formulations.

Type I Error Rates

To investigate the type I error rates, we generate m independent Poisson processes under H_0 according to the following Weibull intensity:

$$\lambda_i(t) = \exp(\beta_0 + \beta_1 x_i)\gamma t^{\gamma-1}, \qquad i = 1, ..., m, \tag{3.5.18}$$

where $\exp(\beta_0)$ and $\gamma > 0$ are scale and shape parameters respectively. The x_i's are taken to be independently and identically distributed Bernoulli random variables such that the

probability of $x_i = 1$ is 0.5 and the probability of $x_i = 0$ is 0.5. This specification was chosen to mimic treatment assignment in a randomized trial.

We set the target length of follow-up to $\tau = 1$ (year) and model loss to follow-up by simulating censoring times with an exponential distribution with mean $\log(0.5)$; this generates about 50% censored follow-ups. In other words, the length of follow-up for subject i is equal to $\tau_i = \min(C_i, 1)$, where the C_i's are independently and identically distributed as exponential with mean $\log(0.5)$. We further specified $\beta_1 = 1$ to represent a treatment effect and $\gamma = 2$ to induce a trend in the conditional intensity. The parameter β_0 is determined by the expected number of events over $[0, 1]$ based on the baseline model, i.e., $\beta_0 = \log(K)$, where $K = \mathrm{E}(N_i(1)|x_i = 0)$. We consider $K = 2$ and 10 to represent small and moderately large numbers of events per subject. Our primary interest lies in the performance for small to moderate sample sizes and so we consider sample sizes: $m = 10, 20, 50$ and 100.

Having generated the processes, we compute the score and adjusted score statistics according to the following specifications of the baseline intensity: (*i*) Weibull, (*ii*) piecewise exponential with 5 equally spaced sub-intervals over $[0, 1]$, denoted by PE-equal, (*iii*) piecewise exponential with 5 cut-points determined by the 20th, 40th, 60th and 80th percentiles of the observed event times, denoted by PE-percentile, and (*iv*) semi-parametric specification. All simulations were replicated 2,000 times.

The Weibull model is the correct model specification in this setting and thus serves as the basis for comparison. We consider two ways of selecting the cut-points for the piecewise exponential model in (*ii*) and (*iii*) above. The equally-spaced division is attractive on the grounds of clinical interpretation of, for example, monthly rates of disease recurrence. The division based on the percentiles of the observed event times groups the ordered event times into a pre-determined number of strata. In the extreme case that the number of strata is equal to the number of observed events, the piecewise exponential model and the semi-parametric model are equivalent (Clayton, 1994).

The empirical type I errors at 10%, 5% and 1% nominal levels are reported in Tables

3.14 and 3.15 for $K = 2$ and $K = 10$ respectively. At the nominal 10% and 5% levels, all of the unadjusted score tests have conservative empirical type I error rates for the sample sizes considered. In other words, the unadjusted tests tend to reject the null hypothesis less frequently than anticipated. This feature is also exhibited at the 1% nominal level, although to a lesser degree.

In contrast, the adjusted score tests generally have well-controlled type I error rates. The adjusted test based on the Weibull specification is substantially less conservative than its unadjusted version, and is satisfactory for sample sizes as small as $m = 20$. The unadjusted test for the PE-equal specification is conservative to approximately the same degree as the Weibull counterpart, but the adjustment appears to over-compensate, leading to inflated empirical type I error rates for small to moderate sample sizes. The unadjusted test based on the PE-percentile method is again conservative , but the adjustment performs extremely well in this context. In fact the adjusted statistic competes very favorably with the adjusted Weibull statistic, and has the attractive property of being somewhat more robust to misspecification of the baseline intensity. The semi-parametric specification again leads to a conservative unadjusted test statistic, as one might expect, but the adjustment does not lead to such good improvements in the empirical type I error rates as are observed in the Weibull and PE-percentile specifications. In fact, at the 5% nominal level, the adjusted statistic for the semi-parametric specification leads to an unacceptably conservative test even for $m = 100$.

This confirms the argument made above regarding the estimator of (3.5.14). To get further evidence, we also carried out simulation studies for the Weibull and PE-percentile specifications with $m = 100$ in which sample moment estimators were used for (3.5.3), i.e., $I_\sigma(\boldsymbol{\theta})$, $I_{\theta\sigma}(\boldsymbol{\theta})$ and $I_\theta(\boldsymbol{\theta})$ in (3.5.3) were computed using their sample moments. The empirical type I errors were severely deflated, even the adjusted statistics failed to provide sufficient improvement (Table 3.16). Therefore, we conclude that the expected form of (3.5.3) should be used, and therefore a weakly parametric model based on a piecewise

	Weibull			PE-equal			PE-percentile			Semi-parametric		
m	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
10	0.030	0.015	0.004	0.031	0.018	0.004	0.031	0.015	0.004	0.056	0.030	0.009
	0.062	0.036	0.010	0.118	0.077	0.030	0.067	0.040	0.012	0.080	0.044	0.010
20	0.045	0.025	0.009	0.048	0.029	0.009	0.048	0.027	0.009	0.065	0.032	0.008
	0.086	0.044	0.017	0.122	0.082	0.027	0.088	0.049	0.018	0.086	0.044	0.009
50	0.059	0.029	0.012	0.059	0.029	0.013	0.060	0.030	0.012	0.064	0.028	0.007
	0.086	0.047	0.017	0.113	0.063	0.023	0.089	0.045	0.019	0.074	0.030	0.009
100	0.072	0.034	0.009	0.071	0.035	0.009	0.071	0.036	0.010	0.072	0.034	0.006
	0.089	0.050	0.015	0.106	0.061	0.017	0.090	0.052	0.016	0.081	0.038	0.007

Table 3.14: Empirical type I errors for the unadjusted score statistic (first row) and the adjusted score statistic (second row) in which $\mathrm{E}(N_i(1)|x_i = 0) = 2$ and the number of cut-points for the PE models is 5.

	Weibull			PE-equal			PE-percentile			Semi-parametric		
m	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
10	0.037	0.025	0.005	0.036	0.024	0.005	0.037	0.024	0.005	0.059	0.034	0.008
	0.076	0.041	0.019	0.147	0.090	0.036	0.075	0.041	0.018	0.076	0.044	0.012
20	0.050	0.029	0.012	0.049	0.032	0.013	0.051	0.032	0.013	0.060	0.029	0.008
	0.078	0.049	0.018	0.127	0.070	0.031	0.083	0.053	0.020	0.073	0.038	0.012
50	0.072	0.032	0.007	0.070	0.033	0.007	0.073	0.035	0.008	0.073	0.030	0.002
	0.101	0.055	0.015	0.131	0.075	0.021	0.108	0.057	0.016	0.083	0.036	0.004
100	0.076	0.043	0.017	0.074	0.043	0.017	0.081	0.046	0.017	0.076	0.036	0.009
	0.091	0.054	0.020	0.107	0.067	0.023	0.095	0.057	0.021	0.082	0.040	0.010

Table 3.15: Empirical type I errors for the unadjusted score statistic (first row) and the adjusted score statistic (second row) in which $\mathrm{E}(N_i(1)|x_i = 0) = 10$ and the number of cut-points is 5.

	Weibull			PE-percentile		
m	0.10	0.05	0.01	0.10	0.05	0.01
100	0.047	0.015	0.001	0.046	0.016	0.001
	0.061	0.024	0.002	0.063	0.026	0.002

Table 3.16: Empirical type I errors for the unadjusted score statistic (first row) and the adjusted score statistic (second row) using the sample moment estimates of the variance in which $\mathrm{E}(N_i(1)|x_i = 0) = 2$ and the number of cut-points is 5.

exponential specification of the intensity is preferred.

Finally we remark that as one would expect, the simulations based on more expected number of events per subject (Table 3.15) had slightly better performance over the statistics with less expected number of events per subject (Table 3.14).

Power of the Tests

We now turn to assess the power of the test based on the above four specifications. We consider gamma random effects in which for fixed z_i, the conditional intensity is given by

$$\lambda_i(t|z_i) = z_i \exp(\beta_0 + \beta_1 x_i)\gamma t^{\gamma-1}, \qquad i = 1, ..., m,$$

where we consider sample size $m = 20$. The z_i's are then assumed to be independently and identically distributed gamma random variables with mean 1 and variance δ, to generate subject-to-subject variation in the intensities. With this model specification, the ratio of the marginal variance to the marginal mean of $N_i(\tau_i)$ is equal to $1 + \delta \exp(\beta_0 + \beta_1 x_i)\tau_i^{\gamma-1}$. We consider the case of small n_i and set the parameters as $\beta_0 = \log(2)$, $\beta_1 = 1$, $\gamma = 2$ and δ ranging from 0.0 to 0.8 corresponding to mild to moderate heterogeneity. The empirical power curves for the unadjusted and adjusted statistics at the 5% nominal significance level are displayed in Figures 3.3 and 3.4.

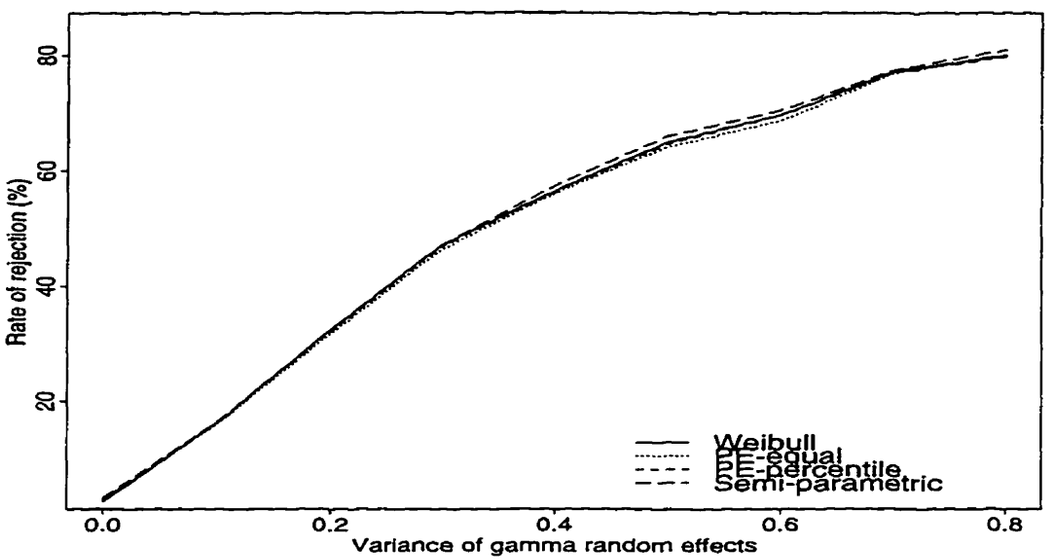

Figure 3.3: Power curves for the unadjusted score statistics at 5% nominal significance level, where $m = 20$.

Since the results in Tables 3.14 and 3.15 suggest that the empirical type I error rates of the unadjusted and adjusted tests are different, it is not appropriate to compare the power of these tests. Nevertheless, the unadjusted tests based on the four model specifications in this section are roughly comparable, suggesting that it is reasonable to compare their relative powers (see Figure 3.3). The power curves were almost indistinguishable, indicating that the unadjusted tests based on these four specifications had similar performance.

The power curves for the adjusted tests are given in Figure 3.4. Here we must bear in mind the inflated type I error rates of the adjusted PE-equal statistic and the conservative nature of the adjusted semi-parametric specification. Since the empirical sizes of the adjusted tests based on the PE-equal and semi-parametric specifications were different from each other and from the other two specifications, we did not attempt to compare their

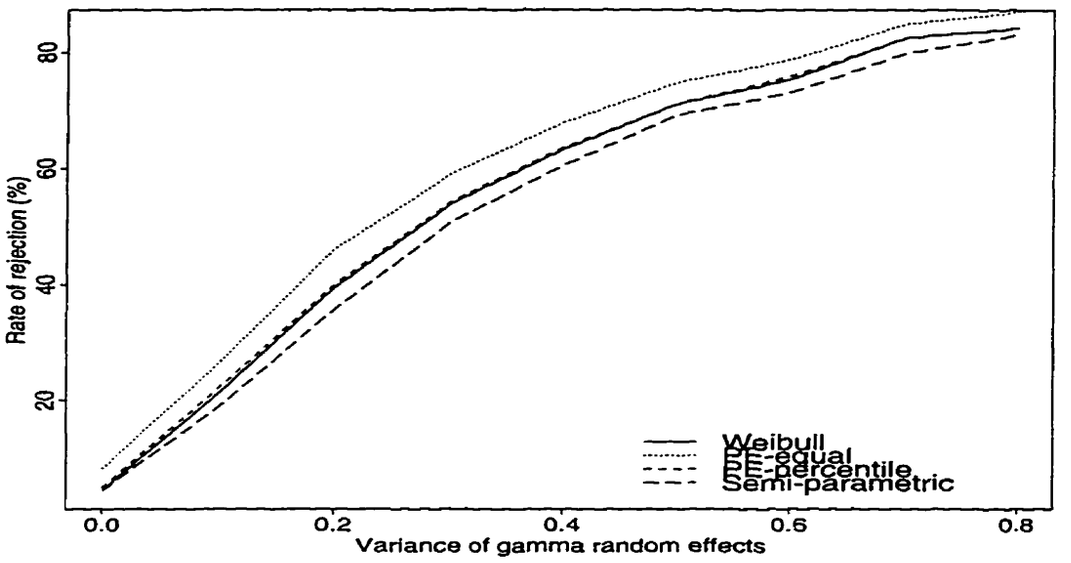

Figure 3.4: Power curves for the adjusted score statistics at 5% nominal significance level, where $m = 20$.

powers. As the adjusted Weibull and adjusted PE-percentile statistic have comparable empirical type I error rates, comparisons between them are most relevant. It is reassuring that there is no loss in power in adopting the robust PE-percentile approach.

On the whole, the adjusted tests seem to have reasonable power to detect heterogeneity and the PE-percentile model seems to out-perform the semi-parametric model in terms of the type I error rate.

Comparison to Pearson Chi-squared Statistic

The Pearson chi-squared statistic is often used to assess the goodness-of-fit in Poisson regression models (McCullagh and Nelder, 1989) and so we compare the performance of the score statistics with this statistic as well. Since the Pearson chi-squared statistic is usually defined for parametric models, we consider the parametric specifications (*i*), (*ii*) and (*iii*) given in above. The usual Pearson chi-squared statistic is defined in the present

context as

$$P = \sum_{i=1}^{m} \frac{(n_i - \Lambda_i(\hat{\boldsymbol{\theta}}))^2}{\Lambda_i(\hat{\boldsymbol{\theta}})},$$

where $\Lambda_i(\boldsymbol{\theta}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\Lambda_0(\tau_i;\boldsymbol{\gamma})$ with $\boldsymbol{x}_i = (1, x_i)'$ and $\boldsymbol{x}_i = x_i$ for Weibull and piecewise exponential specifications respectively, $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients, $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters specifying the baseline intensity, and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma})'$. Under H_0, P is approximately χ^2 distributed with degrees of freedom $m - k$, where k is the dimension of $\boldsymbol{\theta}$.

Farrington (1996) recently constructed a first-order modification to P which improved substantially the performance of the goodness-of-fit test in generalized linear models with overdispersion manifested by a multiplicative variance inflation factor. The modification is directed at inducing approximate orthogonality between the adjusted statistic and the estimates of the regression parameters in the sense that the distribution of the adjusted statistic conditional on $\hat{\boldsymbol{\theta}}$ depends only weakly on the true value of $\boldsymbol{\theta}$. The modified Pearson statistic for Poisson regression models is defined by

$$P^A = P - \sum_{i=1}^{m} \frac{(n_i - \Lambda_i(\hat{\boldsymbol{\theta}}))}{\Lambda_i(\hat{\boldsymbol{\theta}})},$$

which is distributed approximately as $\chi^2(m-k)$. Farrington (1996) also suggested using the standardized version of P^A,

$$ZP^A = \frac{P^A - (m-k)}{\sqrt{2(m-k)}},$$

which is distributed approximately as $N(0,1)$.

Here we study the type I error rate of the tests of homogeneity based on P, P^A and ZP^A. The data were generated according to (3.5.18). Since it is well-known that the Pearson chi-squared test has poor performance for small n_i (McCullagh and Nelder, 1989),

	Weibull			PE-equal			PE-percentile		
m	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05	0.01
10	0.135	0.084	0.026	0.370	0.245	0.075	0.369	0.233	0.063
	0.125	0.062	0.014	0.447	0.282	0.079	0.493	0.322	0.093
	0.136	0.087	0.036	0.460	0.364	0.201	0.509	0.399	0.225
20	0.134	0.079	0.034	0.190	0.110	0.030	0.183	0.097	0.019
	0.113	0.052	0.013	0.241	0.129	0.032	0.301	0.167	0.043
	0.117	0.070	0.024	0.258	0.167	0.077	0.316	0.211	0.090
50	0.128	0.082	0.038	0.119	0.064	0.013	0.101	0.050	0.011
	0.109	0.051	0.011	0.174	0.097	0.022	0.234	0.133	0.028
	0.113	0.060	0.021	0.182	0.113	0.038	0.247	0.152	0.047
100	0.133	0.092	0.043	0.076	0.041	0.011	0.068	0.031	0.007
	0.106	0.058	0.012	0.158	0.076	0.017	0.229	0.125	0.032
	0.112	0.067	0.018	0.166	0.087	0.027	0.236	0.146	0.046

Table 3.17: Empirical type I errors for the Pearson statistic (first row), Farrington's modified Pearson statistic (second row) and standardized modified Pearson statistic (third row) in which $\mathrm{E}(N_i(1)|x_i = 0) = 10$ and the number of cut-points is 5.

we consider it only for the case $K = 10$. Note that these tests detect both under-dispersion and over-dispersion.

The empirical type I error rates of the tests based on P, P^A and ZP^A are reported in Table 3.17. The ordinary statistic P for all specifications has grossly inflated type I error rates. The modified statistic P^A based on the correct specification (Weibull), however, performs very well, although its standardized version ZP^A has slightly higher type I error rates than the nominal levels. In contrast to our findings regarding the adjusted score tests, P^A and ZP^A based on the piecewise exponential specifications exhibit unacceptably inflated type I error rates. This may indicate that the Pearson statistic is very sensitive to the specification of the baseline intensity and thus serves more as an omnibus goodness-of-fit test, rather than a test directed at detecting extra-Poisson variation alone. Provided that

the model is correctly specified, P^A is an attractive statistic favored over Z_p^A because of its simpler computation. However, correct model specification is rather difficult to achieve and some sort of robustness is definitely preferable. Therefore, when testing for homogeneity, the adjusted score statistic using the PE-percentile specification is recommended due to its robustness to model specification and its satisfactory performance.

3.5.4 Gamma Interferon in CGD

We consider the CGD study described in section 1.4.1 as an illustration to the testing procedure in this section.

We first fit a semi-parametric model involving all available covariates under the assumption of homogeneity in which log-transforms were applied to the three continuous covariates: age, height and weight. Plots for the martingale residuals from this model versus log(age), log(height) and log(weight) are shown in Figure 3.5 with a LOWESS smoother to help elucidate any patterns. These residual plots indicate that the functional forms of the covariates are quite reasonable under the assumption of homogeneity.

We then fit the same regression models using the Weibull and piecewise exponential (with 5 cut-points based on the percentiles of the event times) specifications. The results of fitting these models are given in the left panel of Table 3.18 along with the corresponding tests of homogeneity. All the tests indicate that there is significant evidence against the hypothesis of homogeneity. For all specifications of the intensity, the adjusted score statistic is at least as large as the unadjusted one, as expected, and hence gives stronger evidence against $H_0 : \sigma^2 = 0$. Note that the estimates and, in particular, the standard errors from the PE-percentile and semi-parametric specifications are generally in very close agreement suggesting that the PE-percentile model with as few as five pieces is sufficiently

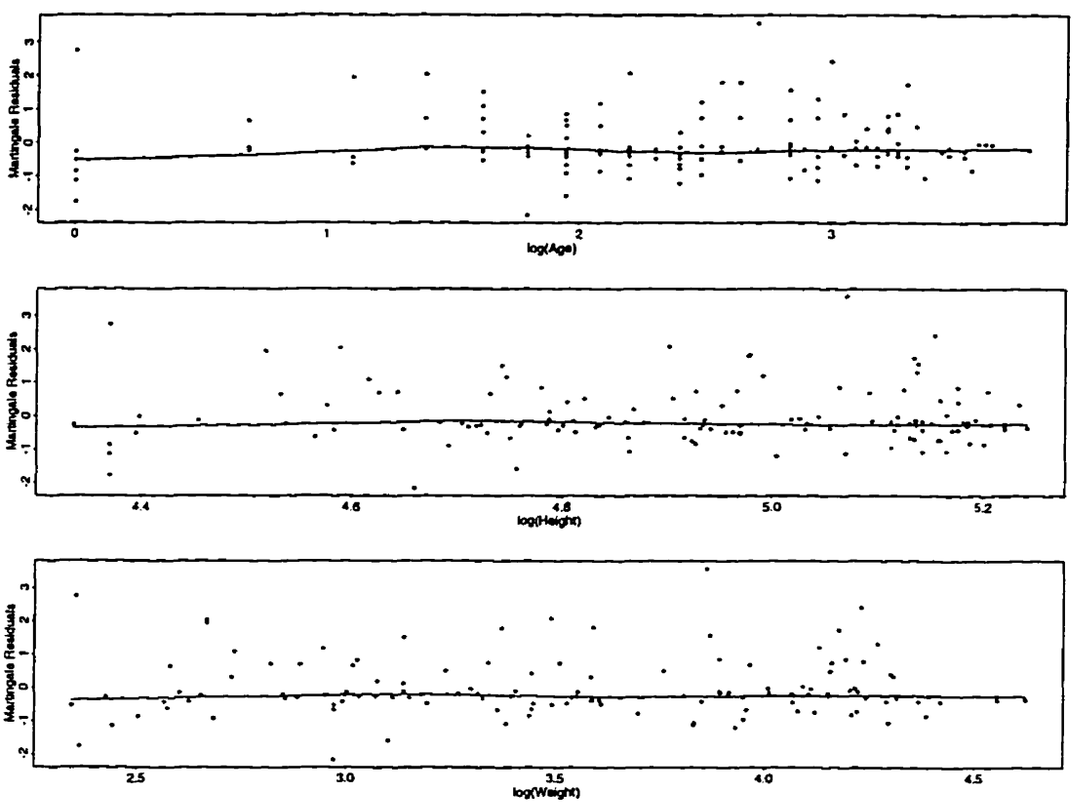

Figure 3.5: Martingale residuals from the fixed-effect semi-parametric model for the CGD data with LOWESS smoothers.

robust. Based on the simulation studies in the previous section, we favor inference using the adjusted PE-percentile statistic and so claim that there is very strong evidence against the hypothesis of homogeneity ($p = 0.001$).

Given this very strong evidence, we fit this data using a random effect model with gamma frailties, i.e., we assume that $z = \exp(\sigma u)$ in (3.5.1) follows a gamma distribution with mean 1 and variance δ. Since parameter estimation for mixed semi-parametric models is very complicated, we elected to adopt the PE-percentile specification with 5 cut-points. The maximum likelihood estimates of the regression coefficients are given in the right panel of Table 3.18. Using the log-transform for δ, we also obtained an approximate 95%

	Fixed-effect Models						Random-effect Model	
	Weibull		PE-percentile		Semi-parametric		PE-percentile	
Covariate*	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
Treatment	-1.060	0.272	-1.063	0.272	-1.070	0.271	-0.987	0.298
Inheritance	-0.884	0.295	-0.924	0.298	-0.874	0.298	-0.857	0.343
log(Age)	-1.040	0.458	-1.056	0.461	-1.032	0.462	-0.995	0.524
log(Height)	4.179	2.788	4.413	2.836	4.229	2.820	4.072	3.170
log(Weight)	-0.548	0.818	-0.603	0.830	-0.592	0.822	-0.607	0.939
Corticosteroids	2.258	0.654	2.270	0.655	2.231	0.656	2.368	0.863
Antibiotics	-0.782	0.348	-0.743	0.347	-0.763	0.349	-0.789	0.420
Gender	0.968	0.394	0.944	0.394	0.907	0.395	0.947	0.468
Hospital:								
US-other	-0.112	0.335	-0.064	0.337	-0.017	0.338	-0.173	0.376
Amsterdam	-1.176	0.502	-1.135	0.503	-1.051	0.509	-1.165	0.584
Other	-0.742	0.497	-0.639	0.501	-0.587	0.504	-0.757	0.549
Test	Statistic	p-value	Statistic	p-value	Statistic	p-value	Estimate of δ	
Unadjusted	1.999	0.023	2.118	0.017	1.638	0.051	0.347	
Adjusted	2.937	0.002	3.069	0.001	2.369	0.009		

*Treatment: 0=placebo, 1=gamma interferon; Inheritance: 0=autosomal recessive, 1=X-linked; Age in years; Height in cm; Weight in kg; Corticosteroids: 0=not used, 1=used at time of study entry; Antibiotics: 0= not used, 1=used at time of study entry; Gender: 0=female, 1=male; Hospital category: baseline is US-NIH.

Table 3.18: Parameter estimates for the fixed-effect and random effect models and the unadjusted and adjusted statistics for the CGD data.

confidence interval for δ as $(0.071, 1.705)$, providing further evidence that heterogeneity was present. Note that the effect of age and antibiotic use is no longer statistically significant under the frailty model, underscoring the importance of testing for homogeneity before making inferences.

3.6 Concluding Remarks and Discussion

3.6.1 General Remarks

In this chapter, we have derived score and IM statistics with different variance estimates for univariate point processes. The performance of the tests are investigated using simulations based on Poisson and renewal processes. The simulation study demonstrates that the test statistics other than the ES form generally have poor finite sample properties, suggesting that adjustment for the test is necessary.

Since the Poisson process is widely used in many applications, we construct adjusted score tests for parametric and semi-parametric regression models arising from Poisson processes. We have shown that the bias induced by the substitution of parameter estimates in the score statistic is non-negative and tends to zero when normalized by $\sqrt{m}$. Although the bias is asymptotically negligible, we demonstrated based on simulations that the score tests performed poorly without adjustment in small to moderately large samples. The adjusted statistics, on the contrary, have reasonable performance in small samples. In particular, the PE-percentile model not only performs extremely well, but also is less restrictive than any parametric model. Although the semi-parametric model is robust to the specification of the baseline intensity, the variance of the score statistic cannot be estimated as efficiently as its parametric counterpart and thereby the performance of the test will compromise.

The proposed tests can also be utilized as tests of model specification. Since the random

effects may arise due to missing important covariates, misspecification of the functional form for the covariates, or misspecification of the baseline intensity, one should consider alternative model specifications when the null hypothesis is rejected.

We have not considered the important case of time-dependent covariates here. For the semi-parametric model in section 3.5.2, we can simply replace $\boldsymbol{x}$ by $\boldsymbol{x}(t)$ in the test statistics provided that $\boldsymbol{x}(t)$ is predictable (Gray, 1995). For the parametric model, we can modify the test statistics easily for the case in which covariates are constant between consecutive events. However, numerical integration may be required to compute the cumulative intensity for general time-dependent covariates.

We remark that the development of the score statistics for the semi-parametric model in section 3.5.2 closely follows that of Gray (1995) and Commenges and Andersen (1995). Gray (1995) and Commenges and Andersen (1995) focused on clustered failure times data in which the cluster sizes were fixed, while we focus on Poisson processes in which the number of events during the follow-up is random, but the arguments are largely the same and the resulting statistics have the same form. We have provided empirical evidence concerning the relative performance of the unadjusted and adjusted tests, and thus provided insight into the practical use of different formulations.

The remainder of this chapter involves some discussion about further research pertaining to a test of homogeneity in stochastic processes. Section 3.6.1 suggests an adjusted score test and a parametric bootstrap method for general parametric point processes. Furthermore, tests of homogeneity for multivariate processes may not be simple extensions of their univariate counterpart. We discuss the underlying difficulties and propose an IM test for this purpose in section 3.6.2.

3.6.2 General Point Processes

Adjusted Score Tests

We consider m independent point processes with parametric conditional intensity (1.2.4). Since $N_i(t)$ is a sub-martingale, the Doob-Meyer decomposition gives the following representation,

$$dN_i(t) = \lambda_i(t|\mathcal{H}_i(t))dt + dM_i(t),$$

where $M_i(t)$ is a zero-mean right-continuous martingale (Fleming and Harrington, 1991, chapter 2; Karatzas and Shreve, 1991). Using the argument for semi-parametric Poisson processes (section 3.5.2), one can show that an approximate bias of the score statistic $T_S(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{m}[(n_i - \Lambda_i(\hat{\boldsymbol{\theta}}))^2 - \Lambda_i(\hat{\boldsymbol{\theta}})]$ is given by

$$b_s(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{m} \frac{\partial \Lambda_i(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'} \hat{I}_{\theta}^{-1}(\hat{\boldsymbol{\theta}}) \frac{\partial \Lambda_i(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}},$$

where $\Lambda_i(\boldsymbol{\theta}) = \int_0^{\tau_i} \lambda_i(t|\mathcal{H}_i(t))dt$, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}')'$ in which $\boldsymbol{\beta}$ is the vector of regression coefficients and $\boldsymbol{\gamma}$ is the vector of parameters for the baseline intensity, and $\hat{I}_{\theta}(\boldsymbol{\theta})$ is the observed information matrix for $\hat{\boldsymbol{\theta}}$ under the null model. Since the expected information matrix is not tractable in general, we have to use the observed version. The variance of the adjusted statistic can be estimated by an empirical estimate discussed in sections 3.2 and 3.3.

Further investigation of this adjusted statistic is desirable. Since renewal processes are popular in many applications, a simulation study of the performance of the adjusted statistic focusing on such processes can be considered.

Another adjustment approach was suggested by Chesher and Spady (1991) for a general IM test. They considered a more complicated second-order approximation to the distribution of the IM statistic. An Edgeworth-type expansion is applied to approximate the moment generating function to the order $O(m^{-1})$ of the series expansion of the statistic. This procedure reduces the bias due to small sample size to the order of m^{-1}. However, it

requires the derivatives of the log-likelihood up to order 3 and the cumulants up to order 4 of the first and second derivatives of the log-likelihood. The computation is extremely complex even for a simple normal linear regression model considered by Chesher and Spady (1991). In most situations, high-order cumulants are difficult to find. This complication, together with the implementation of the method, makes this approach less attractive than the bias-adjustment approach we considered, even though Chesher and Spady's adjustment seems to be more efficient.

Parametric Bootstrap

Given a data set, the finite sample distribution of the test statistic can be estimated by the parametric bootstrap (Efron and Tibshirani, 1993). We outline the procedure here.

First, the parameters are estimated under the fixed-effect model and the test statistic is computed. Denote this estimate by $\hat{\boldsymbol{\theta}}$ and the observed test statistic by $\hat{T}$.

Second, we generate a sample from the assumed fixed-effect model with parameter $\hat{\boldsymbol{\theta}}$ based on simulations. A test statistic, T^*, is then computed from this sample. This simulation procedure is repeated B times, say 1,000, and an empirical distribution of the test statistic is obtained. The ith point process with a general intensity (1.2.4) may be generated as follows, keeping the covariate process $x_i(t)$ and the length of follow-up τ_i unchanged:

1. Generate u_1 from $Unif(0,1)$. The first event time is found by solving $\Lambda(t_1) = -\log(1-u_1)$, where $\Lambda(t) = \int_0^t \lambda(s|\mathcal{H}(s))ds$. A numerical procedure may be required here.

2. Let t_k be the time for the kth event. Given $\mathcal{H}(t_k)$, the survivor function for the next duration $D_{k+1} = T_{k+1} - T_k$ is equal to

$$\Pr(D_{k+1} > d|\mathcal{H}(t_k)) = \prod_{s\in[t_k,t_k+d)} (1-\lambda(s|\mathcal{H}(s))ds)$$

$$= \exp\left(-\int_{t_k}^{t_k+d} \lambda(s|\mathcal{H}(s))ds\right).$$

Thus, generate u_{k+1} independently of the previous u's from $Unif(0,1)$. The $(k+1)$th event time is found by solving $\Lambda(t_{k+1}) = \Lambda(t_k) - \log(1-u_{k+1})$.

3. The process stops when $t_{k+1} > \tau_i$.

Third, the p-value of the test is approximately equal to $\{\#\text{ of } T^* > \hat{T}\}/B$.

As only the fixed-effect model needs to be estimated, this parametric bootstrap should not be a computationally intensive procedure.

3.6.3 Bivariate Processes

We consider m independent bivariate processes with intensities (1.2.5). Let

$$\Sigma = \begin{pmatrix} \sigma_1 & \sigma_{12} \\ \sigma_{12} & \sigma_2 \end{pmatrix}$$

be the variance-covariance matrix of the bivariate random effect $u_i = (u_{i1}, u_{i2})'$. The null hypothesis of homogeneity is equivalent to $H_0 : \Sigma = 0$. Since a zero variance term implies zero covariance, the score function for $(\sigma_1, \sigma_2, \sigma_{12})$ is invariant to the correlation (ρ) under H_0 and thus the score statistic will not asymptotically follow the $\chi^2_{(3)}$ distribution.

One way to tackle this problem is to impose certain additional assumptions. As the main purpose is to detect heterogeneity, we may merely focus on the variances by assuming that the correlation is equal to zero. This reduces to the independent components test of Smith and Heitijan (1993). We will demonstrate this approach using the CHEST study in chapter 4.

Alternatively, since the bivariate random effects usually arise from some unmeasured shared covariates, we may assume that their variances are proportional, say $\sigma_2 = c\sigma_1$.

For given c and ρ, we need to test $H_0 : \sigma_1 = 0$ only (c.f. Sutradhar and Das, 1995 for autocorrelated random effects in generalized linear models, where $c = 1$ there). However, the test statistic depends on the values of c and ρ which are usually unknown. Sutradhar and Das (1995) argued that since the score function for ρ evaluated under H_0 is identically equal to zero, ρ can be taken as any convenient value. They considered the score functions for parameters other than ρ and the test statistic is evaluated at $\sigma = 0$, $\rho = 0$ and the maximum likelihood estimates for the other parameters under H_0. Based on this argument, it is not difficult to show that for a bivariate point process, the score function for c evaluated under H_0 is also identically equal to zero. We may then assume that in the neighborhood of zero covariance matrix, the random effects have common variance and are uncorrelated. In this case, the score test is not difficult to derive and it is asymptotically normally distributed.

Another approach is to use the IM test. As pointed out by Chesher (1984), the alternative to $H_0 : \Sigma = 0$ is a model with both intercepts random. The IM test should focus on testing whether the intercept terms in both intensity functions (1.2.5) are constant. The IM statistic is obtained from

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_0 \partial \boldsymbol{\beta}_0'} + \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_0} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_0'},$$

where $\boldsymbol{\beta}_0 = (\beta_{10}, \beta_{20})'$ is the vector of intercepts.

For bivariate processes, the log-likelihood of the fixed-effect model consists of two parts arising from each component/transition:

$$\ell(\boldsymbol{\theta}) = \ell_1(\boldsymbol{\theta}_1) + \ell_2(\boldsymbol{\theta}_2).$$

If $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are not functionally related and we usually assume so, maximum likelihood estimates for them under H_0 can be found by maximizing $\ell_1(\boldsymbol{\theta}_1)$ and $\ell_2(\boldsymbol{\theta}_2)$ individually.

It is readily seen that for any $\boldsymbol{\theta}$, $\partial^2\ell(\boldsymbol{\theta})/\partial\beta_{10}\partial\beta_{20} = 0$ and evaluated at the maximum likelihood estimates, $\partial\ell(\hat{\boldsymbol{\theta}})/\partial\beta_{10}\partial\ell(\hat{\boldsymbol{\theta}})/\partial\beta_{20} = 0$. Therefore, we may consider the IM statistic as

$$\begin{aligned} D_m(\boldsymbol{\beta}_0) &= \frac{1}{m}\sum_{i=1}^{m}\begin{pmatrix} d_{i1}(\boldsymbol{\theta}_1) \\ d_{i2}(\boldsymbol{\theta}_2)\end{pmatrix}, \\ d_{ij}(\boldsymbol{\theta}_j) &= \frac{\partial^2\ell_{ij}(\boldsymbol{\theta}_j)}{\partial\beta_{j0}^2} + \left(\frac{\partial\ell_{ij}(\boldsymbol{\theta}_j)}{\partial\beta_{j0}}\right)^2, \end{aligned}$$

for $j = 1, 2$. The variance of $D_m(\boldsymbol{\beta}_0)$ may be obtained accordingly. The standardized statistic is asymptotically $\chi^2(2)$ distributed. This IM test turns out to be the score test under the assumption of zero correlation described above.

It is important to investigate the properties of these tests, possibly by means of simulation. It is also interesting to consider an adjusted version of the test in order to improve the finite sample performance. Further research in this direction is recommended.

Chapter 4

Inference for Random Effect Models

4.1 Overview

In chapter 2, we reviewed some common methods of estimation for random effect models in the context of clustered failure time data, and contrasted their merits and limitations. The objective of this chapter is to adapt and investigate the use of these methods for univariate and bivariate processes. In particular, in view of adopting genuine mixing distributions for multivariate processes, we will focus on Gauss-Hermite integration with log-normal random effects and the EM algorithm with non-parametric (discrete) random effects.

Some methods of estimation for random effect models have been investigated for bivariate survival data (Pickles and Crouchley, 1995) and mixed linear and logistic regression models (Butler and Louis, 1992). These studies provide empirical evidence that estimates of regression coefficients typically have negligible bias regardless of the assumed mixing distributions and certain methods of estimation. Neuhaus et al. (1992) showed that parameter estimates in mixed logistic regression models are inconsistent when the mixing distribution is misspecified, but notes that the magnitude of the bias in the estimated covariate effects is small in general and the variance estimates obtained from the misspecified likelihood are

also valid for practical use. The estimate for the variance of the random effect, however, may have large bias if the assumed mixing distribution has a very different shape from the true mixing distribution. We anticipate similar findings for random effect models arising from point processes and intend to investigate this phenomenon systematically in this chapter.

Simulation studies for comparing the performance of Gauss-Hermite integration and the EM algorithm are carried out based on mixed Poisson and renewal processes in section 4.2. The comparison is made in terms of bias, coverage probability, and efficiency of the estimators, as well as robustness to misspecification of the mixing distribution. A similar simulation study is also carried out to examine the use of piecewise constant baseline functions when the functional form of the baseline intensity is unknown. The CGD study described in section 1.4.1 is analyzed using these two methods in section 4.3. Extensions of these methods of estimation for bivariate processes are discussed in section 4.4. In section 4.5, the CHEST study (section 1.4.2) is modeled using the mixed two-state processes proposed in section 1.2.2.

It should be noted that observed heterogeneity may be mainly due to misspecification of the model. Proper specification of the intensity functions is thus more important than specifying the mixing distribution for the purpose of better understanding of the process, and avoiding "spurious" heterogeneity. Without model diagnostic tools, it may not be possible to obtain a proper specification of model. We therefore suggest that one should start with a more comprehensive model, e.g., using multiple time scales as in (1.2.7), then carry out a test of homogeneity to check if a random effect component is necessary, fit a random effect model if the test provides evidence for heterogeneity, and finally select the significant components by likelihood ratio test or Wald test. The analysis of the CHEST study demonstrates this modeling strategy.

4.2 Comparison of Methods of Estimation

4.2.1 Design of Simulation Studies

In this section, we consider log-normal and non-parametric (discrete) random effects and compare the performance of the estimators obtained from Gauss-Hermite integration for the former and the EM algorithm for the latter, via simulations.

We generate mixed Poisson and renewal processes with conditional intensities of the Weibull form, given v_i,

$$\lambda_i(t|v_i) = v_i \exp(\beta_0 + \beta_1 x_i)\alpha t^{\alpha-1},$$

where t is the calendar time and backward recurrence time for Poisson and renewal processes respectively, $x_i \sim Bin(1, 0.5)$ which mimics a random treatment assignment and $\alpha = \exp(\gamma)$. The random effects v_i have mean 1 and variance σ from one of the following distributions:

1. log-normal;
2. binary: $V = 1 - \sqrt{\sigma}$ with probability $1/2$ and $V = 1 + \sqrt{\sigma}$ with probability $1/2$;
3. mixture of two log-normals: $V = V_1$ with probability $1/2$ and $V = V_2$ with probability $1/2$, where V_1 and V_2 are independent log-normal random variables with $\mathrm{E}(V_1) = 3/2$, $\mathrm{E}(V_2) = 1/2$, $\mathrm{var}(V_1) = 9(4\sigma - 1)/20$ and $\mathrm{var}(V_2) = (4\sigma - 1)/20$.

The probability density functions of log-normal and mixture of two log-normal distributions are plotted in Figure 4.1 for $\sigma = 0.3$ and 0.5. We utilize these forms to examine the robustness of the estimators to misspecification of mixing distribution.

We simulated $m = 100$ independent processes according to the above scheme over a unit interval $[0, 1]$. The true values of the parameters are: $\beta_0 = \log(2)$, $\beta_1 = 1$, $\gamma = \log(2)$ and $\sigma = 0.3$ and 0.5 for Poisson processes; and $\beta_0 = \log(6)$, $\beta_1 = 1$, $\gamma = \log(2)$ and

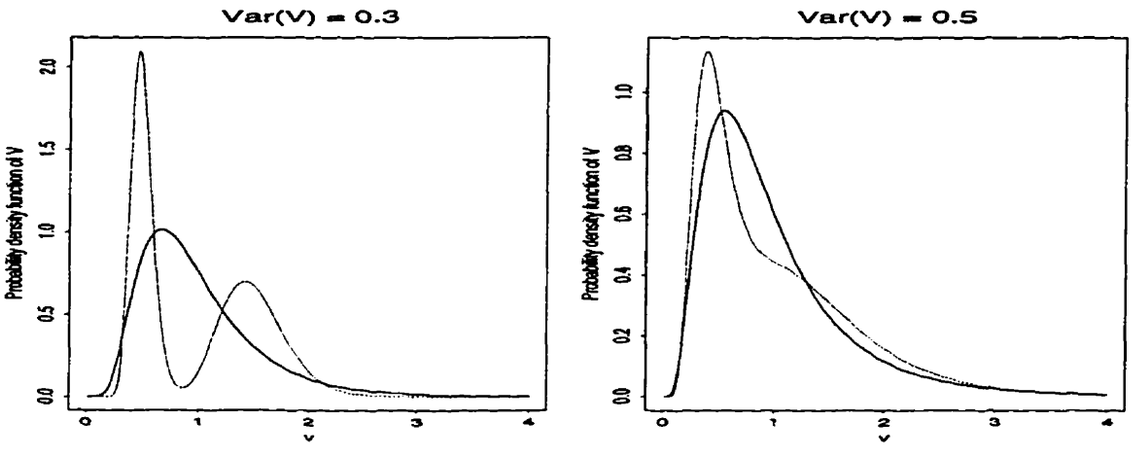

Figure 4.1: Probability density functions for log-normal (solid line) and mixture of two log-normal (dotted line) distributions with mean 1 and variance $\sigma = 0.3$ (left) and $\sigma = 0.5$ (right) considered in the design of the simulation.

$\sigma = 0.3$ and 0.5 for renewal processes. We simulated 1,000 times for each configuration. Under these settings, the number of events per subject is small to moderately large, see Table 4.1. As expected, the standard deviation of the average number of events increases as σ increases. The choice of the values for σ corresponds to moderately large and large overdispersion.

The generated data were fit by assuming log-normal random effects or discrete random effects from an unspecified mixing distribution. Estimation for the former was carried out by the Gauss-Hermite rule with 5, 10 and 20 nodes (section 2.5.2), and the latter by the EM algorithm (section 2.5.3). The range of these nodes was specified to examine the effect of the number of nodes on the properties of the resulting estimation. For the non-parametric random effects, since the number of mass points required is usually quite small, the initial number of mass points is set to 10. During the iteration of the EM algorithm, we combine adjacent mass points if their difference divided by the estimated standard deviation of the mixing distribution is less than 0.05, where the estimated standard deviation is calculated from the current estimates of the mass points and their respective masses. For notational

		Log-normal		Binary		Log-normal Mixture	
σ	Process	$x=0$	$x=1$	$x=0$	$x=1$	$x=0$	$x=1$
0.0	Poisson	2.00	5.44				
		(1.40)	(2.32)				
	Renewal	2.40	4.19				
		(0.93)	(1.16)				
0.3	Poisson	2.00	5.43	2.00	5.43	2.00	5.43
		(1.78)	(3.72)	(1.77)	(3.76)	(1.77)	(3.76)
	Renewal	2.31	4.05	2.28	4.00	2.30	4.03
		(1.15)	(1.62)	(1.20)	(1.73)	(1.19)	(1.68)
0.5	Poisson	1.99	5.47	2.01	5.42	1.99	5.44
		(1.95)	(4.45)	(1.99)	(4.47)	(1.97)	(4.45)
	Renewal	2.26	3.98	2.19	3.84	2.24	3.94
		(1.25)	(1.81)	(1.39)	(2.07)	(1.28)	(1.87)

Table 4.1: Average numbers of events per subject with standard deviations in parentheses.

convenience, we label the Gauss-Hermite rule with k nodes as GHk and the EM algorithm for non-parametric random effects as NP.

Since the baseline intensity is often of unknown form, it is desirable to examine the effect of using a piecewise constant function as an approximation to the baseline intensity. For the purpose of parameter parsimony, we use a 5-point piecewise constant function whose cut-points are determined by the percentiles of the observed event times. The results in chapter 3 also provide empirical evidence for the adequacy of using a small number of cut-points. As will be seen below, the performance of the GH and NP estimates is quite similar. Therefore, we consider the GH estimation only.

We computed the standardized biases, 95% coverage probabilities, averages of the model-based standard errors of the estimates and the simulation-based standard errors. We also included the results for the logarithmic transform for σ because the distribution of the estimate of $\log(\sigma)$ may be less skewed than that of the estimate of σ. The stan-

dardized bias is the average bias divided by the simulation-based standard error and it is approximately normally distributed, except possibly for the estimate of σ when the true value of σ is small. The 95% coverage probability is the proportion of the model-based 95% confidence intervals that cover the true value of the parameter. The model-based standard error is computed from the observed information matrix. The simulation-based standard error is the standard deviation of the estimates computed from the simulations.

Remarks on the EM Algorithm

Here we give more details on the EM algorithm. The convergence rate of the EM is usually very slow. Although there are a number of algorithms for speeding it up (see Meng and van Dyk, 1997 for a review), many of them are quite difficult to implement. We do not intend to investigate different EM algorithms, but look for simple-to-implement and yet "fast" algorithms in terms of computational time.

For discrete mixing distributions, the E-step is very simple. However, the M-step involves maximizing a function of possibly a large number of parameters ($\boldsymbol{\theta}$ and $\boldsymbol{\xi}$). The computational time required in the M-step is likely to be high. Rai and Matthews (1993) suggested a one-step Newton-Raphson iteration in the M-step to modify the original EM algorithm, which is called the EM1 algorithm. They established self-consistency of the EM1 algorithm and demonstrated that although larger number of EM cycles may be required, the overall computational time is shorter than the EM algorithm.

Jamshidian and Jennrich (1997) recently proposed two accelerated EM algorithm using quasi-Newton methods: QN1 and QN2 algorithms. The QN1 algorithm only requires a few more steps in the EM algorithm, and the QN2 requires more effort in implementation as an additional maximization for the observed data log-likelihood is needed. Jamshidian and Jennrich demonstrated that the QN1 and QN2 algorithms can improve the computational time dramatically over the EM algorithm. Since the QN1 algorithm is simpler to implement

and also leads to fast convergence, we state this algorithm here.

Given an initial value $\phi^{(0)}$, obtain an update $\phi_{EM}^{(0)}$ from an EM cycle. Set $g^{(0)} = \phi_{EM}^{(0)} - \phi^{(0)}$ and $A^{(0)} = -I$, where I is the identity matrix of dimension equaling the number of parameters in ϕ. Iterate the following steps until convergence:

1. Compute $\triangle\phi^{(k)} = -A^{(k)}g^{(k)}$ and set $\phi^{(k+1)} = \phi^{(k)} + \triangle\phi^{(k)}$.

2. Given $\phi^{(k+1)}$, obtain an update $\phi_{EM}^{(k+1)}$ from an EM cycle. Compute $g^{(k+1)} = \phi_{EM}^{(k+1)} - \phi^{(k+1)}$ and $\triangle g^{(k)} = g^{(k+1)} - g^{(k)}$. The update for A is obtained from the quasi-Newton method:

$$A^{(k+1)} = A^{(k)} + \frac{(\triangle\phi^{(k)} - A^{(k)}\triangle g^{(k)})(\triangle\phi^{(k)})'A^{(k)}}{(\triangle\phi^{(k)})'A^{(k)}\triangle g^{(k)}}.$$

Nevertheless, the QN1 is not globally convergent and the adjustment for ϕ in step 1 may lead to a point outside the parameter space. Given these considerations, we chose the EM1 algorithm in the simulation study and illustrate the QN1 algorithm in the examples.

4.2.2 Simulation Results

First we summarize the results for mixed Poisson processes (Tables 4.2 and 4.3). Let $\hat{\phi}$ be the estimate of ϕ. For $\sigma = 0.3$, the GH and NP estimates of the regression coefficients (β_0 and β_1), the shape parameter (γ) and the dispersion parameter (σ) have negligible bias regardless of the underlying mixing distributions The magnitudes of the bias are of comparable sizes between the GH and NP methods for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\gamma}$. Although the GH estimates of σ have larger bias than the NP estimate of σ, the bias is still very small for practical purposes. For $\sigma = 0.5$, the GH and NP estimates for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\gamma}$ also have negligible bias, but $\hat{\sigma}$ obtained from the GH has small positive bias whereas the NP still provides an unbiased estimate for σ when the mixing distribution is highly discrete (binary).

	Method[6]	Log-normal					Binary					Mixture of 2 Log-normal				
		β_0	β_1	γ	σ	$\log\sigma$	β_0	β_1	γ	σ	$\log\sigma$	β_0	β_1	γ	σ	$\log\sigma$
Std	GH5	-0.056	0.012	0.046	-0.083	-0.262	0.010	-0.147	0.018	0.753	0.692	-0.024	-0.036	0.027	0.563	0.442
Bias[1]	GH10	-0.057	0.016	0.043	-0.055	-0.238	-0.011	-0.008	0.018	0.701	0.614	-0.045	0.028	0.025	0.493	0.358
	GH20	-0.057	0.016	0.042	-0.065	-0.244	-0.014	0.001	0.018	0.686	0.600	-0.047	0.028	0.025	0.485	0.348
	NP	-0.054	-0.002	0.038	-0.016	-0.196	-0.104	0.075	0.016	0.081	-0.025	-0.095	0.060	0.017	-0.046	-0.082
Cover	GH5	0.952	0.951	0.956	0.904	0.967	0.966	0.948	0.953	0.978	0.876	0.963	0.946	0.947	0.970	0.909
Prob[2]	GH10	0.955	0.959	0.956	0.905	0.971	0.964	0.949	0.953	0.990	0.941	0.968	0.957	0.946	0.980	0.956
	GH20	0.955	0.960	0.956	0.908	0.978	0.963	0.947	0.953	0.992	0.951	0.968	0.956	0.946	0.981	0.964
	NP	0.956	0.955	0.965	0.915	0.971	0.960	0.966	0.960	0.974	0.984	0.964	0.954	0.964	0.963	0.977
Model	GH5	0.131	0.158	0.052	0.106	0.386	0.137	0.165	0.052	0.136	0.350	0.136	0.163	0.052	0.130	0.361
SE[3]	GH10	0.131	0.159	0.052	0.108	0.390	0.137	0.167	0.052	0.136	0.365	0.136	0.166	0.052	0.129	0.372
	GH20	0.131	0.159	0.052	0.108	0.392	0.137	0.168	0.052	0.136	0.368	0.136	0.166	0.052	0.128	0.371
	NP	1.546	0.174	0.057	1.799	5.730	0.830	0.172	0.056	1.004	3.104	2.120	0.175	0.056	1.567	6.469
Sim	GH5	0.127	0.163	0.050	0.110	0.406	0.127	0.164	0.052	0.130	0.330	0.130	0.163	0.052	0.129	0.353
SE[4]	GH10	0.127	0.159	0.050	0.111	0.410	0.127	0.165	0.052	0.113	0.308	0.129	0.163	0.052	0.117	0.338
	GH20	0.127	0.158	0.050	0.109	0.407	0.127	0.166	0.052	0.112	0.306	0.129	0.162	0.052	0.113	0.332
	NP	0.128	0.163	0.050	0.119	0.410	0.123	0.153	0.052	0.071	0.239	0.129	0.161	0.052	0.086	0.284
Mass points[5]		3.0 (0.77)					2.55 (0.72)					2.84 (0.80)				

[1]Standardized bias = bias divided by simulation-based standard error. [2]95% coverage probability. [3]Average of the model-based standard error.
[4]Simulation-based standard error. [5]Average number of mass points used in NP with standard deviation in parentheses.
[6]GH5 = 5-node Gauss-Hermite; GH10 = 10-node Gauss-Hermite; GH20 = 20-node Gauss-Hermite; NP = Non-parametric.

Table 4.2: Comparison of the methods of estimation for mixed Poisson processes based on log-normal, binary and mixture of two log-normal mixing distributions with $\sigma = 0.3$.

	Method	Log-normal				Binary				Mixture of 2 Log-normal				
		β_0	β_1	γ	σ	$\log\sigma$	β_0	β_1	γ	σ	$\log\sigma$	β_0	β_1	γ	σ	$\log\sigma$
Std	GH5	-0.105	0.058	0.074	-0.088	-0.271	0.093	-0.835	0.049	1.211	1.687	-0.036	0.021	0.023	0.422	0.290
Bias	GH10	-0.088	0.048	0.070	-0.017	-0.195	0.167	-0.279	0.049	1.288	1.725	-0.038	-0.003	0.017	0.436	0.316
	GH20	-0.089	0.053	0.071	-0.028	-0.201	0.190	-0.049	0.047	1.338	1.847	-0.050	0.023	0.015	0.433	0.320
	NP	-0.100	0.071	0.056	-0.060	-0.252	-0.090	0.063	0.046	0.105	0.005	-0.103	0.048	0.016	-0.050	-0.198
Cover	GH5	0.936	0.905	0.956	0.866	0.922	0.980	0.845	0.962	0.632	0.414	0.957	0.938	0.947	0.921	0.895
Prob	GH10	0.950	0.921	0.949	0.892	0.935	0.992	0.922	0.961	0.873	0.601	0.966	0.955	0.946	0.956	0.915
	GH20	0.949	0.921	0.950	0.903	0.946	0.987	0.935	0.962	0.968	0.654	0.967	0.959	0.947	0.963	0.942
	NP	0.945	0.933	0.967	0.886	0.928	0.965	0.965	0.964	0.959	0.972	0.950	0.958	0.966	0.919	0.952
Model	GH5	0.143	0.173	0.052	0.146	0.315	0.172	0.181	0.052	0.333	0.302	0.140	0.178	0.052	0.174	0.305
SE	GH10	0.145	0.177	0.052	0.157	0.327	0.164	0.204	0.052	0.275	0.307	0.151	0.184	0.052	0.185	0.322
	GH20	0.146	0.179	0.052	0.162	0.337	0.170	0.212	0.052	0.344	0.353	0.151	0.187	0.052	0.191	0.332
	NP	4.8E5	0.201	0.058	1.3E7	4.6E6	0.607	0.178	0.056	0.306	0.732	0.840	0.202	0.058	1.107	2.183
Sim	GH5	0.154	0.196	0.051	0.184	0.373	0.140	0.183	0.050	0.495	0.415	0.147	0.189	0.053	0.205	0.346
SE	GH10	0.151	0.191	0.052	0.183	0.368	0.153	0.222	0.051	0.307	0.309	0.143	0.185	0.053	0.195	0.328
	GH20	0.151	0.192	0.052	0.176	0.363	0.153	0.220	0.050	0.340	0.321	0.143	0.187	0.053	0.191	0.320
	NP	0.154	0.201	0.052	0.199	0.380	0.135	0.162	0.051	0.104	0.204	0.145	0.192	0.053	0.139	0.284
Mass points		3.4 (0.84)					2.6 (0.70)					3.4 (0.79)				

Table 4.3: Comparison of the methods of estimation for mixed Poisson processes based on log-normal, binary and mixture of two log-normal mixing distributions with $\sigma = 0.5$.

	Method	Log-normal				Binary				Mixture of 2 Log-normal				
		β_0	β_1	γ	σ	$\log\sigma$	β_0	β_1	γ	σ	$\log\sigma$	β_0	β_1	γ	σ	$\log\sigma$
Std	GH5	0.031	0.059	0.085	0.013	-0.204	0.000	-0.025	-0.017	0.889	0.877	0.041	0.021	0.015	0.571	0.439
Bias	GH10	0.031	0.059	0.086	0.013	-0.203	-0.006	0.006	-0.027	0.881	0.859	0.034	0.020	0.010	0.568	0.429
	GH20	0.031	0.058	0.085	0.013	-0.204	-0.007	0.006	-0.028	0.885	0.860	0.034	0.020	0.009	0.568	0.428
	NP	0.139	0.114	0.228	0.190	-0.032	0.166	0.100	0.273	0.323	0.212	0.186	0.105	0.251	0.208	0.030
Cover	GH5	0.917	0.956	0.940	0.953	0.990	0.921	0.900	0.956	0.998	0.922	0.930	0.952	0.950	0.992	0.955
Prob	GH10	0.915	0.957	0.936	0.953	0.993	0.917	0.955	0.957	1.000	0.940	0.930	0.950	0.950	0.992	0.964
	GH20	0.915	0.957	0.937	0.953	0.993	0.916	0.956	0.958	1.000	0.947	0.928	0.951	0.951	0.992	0.963
	NP	0.950	0.963	0.942	0.932	0.972	0.952	0.963	0.953	0.971	0.967	0.956	0.952	0.947	0.941	0.969
Model	GH5	0.142	0.163	0.050	0.158	0.576	0.149	0.175	0.051	0.201	0.470	0.147	0.171	0.051	0.187	0.509
SE	GH10	0.142	0.163	0.050	0.158	0.576	0.150	0.176	0.051	0.205	0.488	0.147	0.171	0.051	0.180	0.512
	GH20	0.142	0.163	0.050	0.158	0.576	0.150	0.176	0.051	0.204	0.488	0.147	0.171	0.051	0.180	0.511
	NP	0.480	0.176	0.053	1.422	2.751	0.365	0.187	0.052	0.418	1.026	0.240	0.182	0.052	0.362	0.913
Sim	GH5	0.159	0.157	0.051	0.131	0.482	0.170	0.172	0.050	0.154	0.358	0.165	0.172	0.051	0.150	0.399
SE	GH10	0.159	0.158	0.051	0.132	0.482	0.170	0.173	0.050	0.152	0.358	0.165	0.173	0.051	0.145	0.394
	GH20	0.159	0.158	0.051	0.132	0.482	0.169	0.173	0.050	0.151	0.357	0.164	0.172	0.051	0.145	0.394
	NP	0.163	0.168	0.052	0.194	0.543	0.169	0.172	0.050	0.132	0.341	0.167	0.178	0.051	0.120	0.389
Mass points		2.96 (0.68)					2.94 (0.64)					2.74 (0.66)				

Table 4.4: Comparison of the methods of estimation for mixed renewal processes based on log-normal, binary and mixture of two log-normal mixing distributions with $\sigma = 0.3$.

	Method	Log-normal				Binary				Mixture of 2 Log-normal			
		β_0	β_1	γ	σ	$\log\sigma$	β_0	β_1	γ	σ	$\log\sigma$	β_0	β_1	γ	σ	$\log\sigma$
Std	GH5	0.013	0.028	0.051	-0.020	-0.207	-0.016	-0.307	-0.066	1.439	2.057	0.009	0.052	0.035	0.509	0.386
Bias	GH10	0.020	0.044	0.060	-0.002	-0.188	0.091	-0.014	-0.047	1.564	2.266	0.019	0.060	0.049	0.520	0.396
	GH20	0.015	0.043	0.056	-0.011	-0.197	0.108	0.016	-0.041	1.558	2.258	0.017	0.060	0.047	0.519	0.395
	NP	0.107	0.095	0.181	0.129	-0.094	0.181	0.153	0.325	0.345	0.257	0.099	0.110	0.200	0.159	-0.012
Cover	GH5	0.922	0.953	0.954	0.945	0.985	0.929	0.946	0.941	0.919	0.523	0.928	0.948	0.941	0.982	0.951
Prob	GH10	0.928	0.955	0.954	0.955	0.991	0.923	0.949	0.938	0.970	0.557	0.925	0.946	0.945	0.983	0.972
	GH20	0.929	0.954	0.953	0.953	0.995	0.920	0.948	0.939	0.996	0.579	0.926	0.946	0.946	0.985	0.970
	NP	0.957	0.942	0.957	0.902	0.942	0.943	0.961	0.936	0.968	0.963	0.959	0.954	0.947	0.931	0.904
Model	GH5	0.153	0.180	0.051	0.220	0.404	0.171	0.205	0.052	0.399	0.371	0.157	0.189	0.051	0.262	0.437
SE	GH10	0.153	0.181	0.051	0.224	0.406	0.175	0.218	0.053	0.447	0.405	0.158	0.190	0.052	0.268	0.445
	GH20	0.153	0.181	0.051	0.223	0.408	0.176	0.220	0.053	0.470	0.418	0.158	0.190	0.052	0.268	0.445
	NP	2.3E2	0.192	0.054	6.4E3	1.9E3	0.229	0.196	0.052	0.223	0.371	0.859	0.202	0.054	11.143	6.287
Sim	GH5	0.169	0.180	0.049	0.191	0.393	0.194	0.204	0.054	0.406	0.346	0.173	0.192	0.052	0.233	0.374
SE	GH10	0.168	0.179	0.049	0.191	0.386	0.204	0.218	0.055	0.389	0.328	0.176	0.192	0.052	0.229	0.370
	GH20	0.168	0.179	0.049	0.189	0.393	0.205	0.219	0.055	0.402	0.334	0.176	0.192	0.052	0.227	0.368
	NP	0.173	0.194	0.051	0.303	0.482	0.187	0.185	0.053	0.141	0.243	0.179	0.202	0.053	0.236	0.383
Mass points		3.3 (0.72)					2.69 (0.65)					3.28 (0.70)				

Table 4.5: Comparison of the methods of estimation for mixed renewal processes based on log-normal, binary and mixture of two log-normal mixing distributions with $\sigma = 0.5$.

When $\sigma = 0.3$, the 95% coverage probabilities for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\gamma}$ obtained from the GH and the NP agree very closely to the nominal level regardless of the mixing distributions, thereby indicating that inference based on the asymptotic normality of maximum likelihood estimation is appropriate for these parameters. There are some discrepancies for the 95% coverage probabilities for the GH and NP estimates for σ probably due to the skewness of the distribution of $\hat{\sigma}$. The empirical coverage is smaller for the log-normal mixing but larger for the binary and mixture of log-normals mixings than the nominal level. The log-transformed σ improves slightly if GH10 and GH20 are used. In contrast, for a larger σ, the 95% coverage probabilities for the GH estimates of β_1 is smaller than 0.95 in log-normal and binary mixtures, whereas the NP estimates can generally maintain a reasonable coverage probability. The 95% coverage probabilities for $\hat{\sigma}$ are in general quite different from the nominal level. For the GH method, the discrepancy gets larger for the binary mixing distribution, and using a larger number of nodes helps in narrowing the discrepancy. There are also slight disagreements in the NP method, but this approach gives coverage which is quite close to the nominal level for the binary mixing distribution.

We also included the model-based standard errors to examine the relative efficiencies of the estimators. For $\sigma = 0.3$, the GH model-based and simulation-based standard errors for all parameters agree very closely, indicating that the parameters are estimated quite efficiently by the GH estimation. Although the NP model-based and simulation-based standard errors for β_1 and γ also agree closely, the NP model-based standard errors for β_0 and σ are much larger than their corresponding simulation-based standard errors. This is because $\hat{\beta}_0$ and $\hat{\sigma}$ obtained from the NP are computed based on the estimated mass points and masses. This finding is consistent with the general estimation of non-parametric distribution which usually ignores the fact that estimation is carried out conditionally on the number of mass points, which is unknown (see section 2.5.3 for a brief discussion). Moreover, similar findings are obtained for $\sigma = 0.5$.

Next we look at the simulation results obtained from the mixed renewal processes

(Tables 4.4 and 4.5). The results in general are in good agreement with the findings for mixed Poisson processes. The only major difference is that for large σ, the 95% coverage probability for $\hat{\beta}_1$ stays very close to the nominal level but the corresponding values for $\hat{\beta}_0$ are slightly different from the nominal level. In the simulation for mixed Poisson processes, an opposite result was observed.

On the whole, the simulation results agree with the previous studies (Pickles and Crouchley, 1995; Butler and Louis, 1992 and Neuhaus et al., 1992). The parameter estimates for the regression parameters and the shape parameter are typically unbiased with valid and quite efficient variance estimates. The major difference between the GH and NP methods of estimation occurs in the estimation of parameters in the mixing distribution. The variance estimate of σ is more stable in the GH than in the NP, although the GH estimate of σ tends to have slight positive bias when the true mixing distribution is highly discrete and the NP estimate of σ generally has negligible bias. Therefore, inference for σ can be carried out in the usual way for GH estimation, but bootstrapped variance estimate may be necessary for NP estimation. Nevertheless, testing for $\sigma = 0$ should be done by a score test described in chapter 3. Unless the underlying mixing distribution is highly discrete, the performance of the GH and NP is very comparable. Furthermore, the number of nodes used in the GH estimation has minimal effect in terms of bias and efficiency of the parameters. Using 10 nodes should be sufficient in most applications involving univariate random effects.

Finally, we examine the effect of using a piecewise constant function for the baseline intensity. The standardized biases, 95% empirical coverage probabilities, model-based standard errors and simulation-based standard errors for β_1, σ and $\log(\sigma)$ are given in Table 4.6. The results in general agree quite closely to the results of specifying a correct baseline intensity. Therefore, the use of piecewise constant baseline function has minimal effect on the performance of the estimators and is also robust to the specification of the baseline intensity.

		Log-normal			Binary			Mixture of Log-normals		
	Process	β_1	σ	$\log\sigma$	β_1	σ	$\log\sigma$	β_1	σ	$\log\sigma$
Std	Poisson	0.048	-0.019	-0.197	-0.273	1.286	1.722	-0.003	0.433	0.313
Bias	Renewal	-0.384	-0.961	-0.955	-0.536	0.825	0.860	-0.376	-0.361	-0.522
Cover	Poisson	0.921	0.882	0.927	0.922	0.915	0.609	0.954	0.958	0.915
Prob	Renewal	0.936	0.713	0.951	0.930	0.992	0.915	0.935	0.869	0.978
Model	Poisson	0.177	0.153	0.316	0.204	0.284	0.314	0.184	0.185	0.320
SE	Renewal	0.171	0.150	0.450	0.199	0.293	0.408	0.178	0.179	0.421
Sim	Poisson	0.191	0.183	0.368	0.223	0.308	0.310	0.185	0.195	0.328
SE	Renewal	0.167	0.148	0.446	0.189	0.271	0.357	0.178	0.171	0.396

Table 4.6: Maximum likelihood estimation for mixed Poisson and renewal processes with 5-point piecewise constant baseline intensity using 10-node Gauss-Hermite integration, where $\sigma = 0.5$.

4.3 Gamma Interferon in CGD

The CGD study described in section 1.4.1 was found to have substantial heterogeneity using the adjusted score test based on a Poisson assumption in section 3.5.4. A mixed Poisson process with gamma random effects was also fit. Here we fit two mixed Poisson processes to the CGD data with log-normal and non-parametric random effects. Specifically, given the random effect v_i, the conditional intensity is given by

$$\lambda_i(t|v_i) = v_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\lambda_0(t),$$

where v_i has mean 1 and variance σ and is distributed either as gamma, log-normal or discrete, and $\lambda_0(t)$ is a piecewise constant function with 5 cut-points determined by the empirical percentiles of the observed event times. Gauss-Hermite integration with 10 nodes was used for the log-normal random effects. The EM1 algorithm was employed for the discrete random effects in which an initial number of mass points is equal to 10 and we combine adjacent mass points if their difference divided by the standard deviation of the

	Gamma		Log-normal		Non-parametric	
Covariate*	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
Treatment	-0.987	0.298	-1.009	0.299	-0.941	0.396
Inheritance	-0.857	0.343	-0.888	0.348	-1.031	0.389
log(Age)	-0.995	0.524	-1.013	0.528	-1.145	0.898
log(Height)	4.072	3.170	4.168	3.205	4.727	4.914
log(Weight)	-0.607	0.939	-0.605	0.947	-0.649	1.204
Corticosteroids	2.368	0.863	2.341	0.825	2.696	1.010
Antibiotics	-0.789	0.420	-0.796	0.423	-0.735	0.668
Gender	0.947	0.468	0.948	0.466	1.096	0.624
Hospital:						
US-other	-0.173	0.376	-0.176	0.382	-0.101	0.495
Amsterdam	-1.165	0.584	-1.190	0.578	-1.228	0.823
Other	-0.757	0.549	-0.774	0.557	-0.751	0.699
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
σ	0.347	0.282	0.390	0.347	0.453	0.396

*See Table 3.18.

	Non-parametric	
Mass Point	0.272	1.622
Mass	0.461	0.539

Table 4.7: Parameter estimates for the random effect models with gamma, log-normal and non-parametric random effects for the CGD data.

mixing distribution is less than 0.05. The estimation result is presented in Table 4.7.

The estimates for the covariate effects are quite similar across the assumed mixing distributions. The variance estimates in the non-parametric random effects are larger than the corresponding values for the gamma and log-normal random effects. The gamma and log-normal estimates for σ are quite close to each other, and the variance estimate for σ is slightly larger in the latter. The NP estimate for σ is larger than for the other two random effects. It is interesting to see that the estimated number of mass points is two in which the masses are quite close. This may imply that there are two hidden subgroups in the sample.

4.4 Random Effect Models for Bivariate Processes

The Gauss-Hermite rule and the EM algorithm can be utilized directly to estimate mixed bivariate processes with intensity functions (1.2.5). In this section, we describe how these two methods of estimation can be implemented for two-state processes. Estimation for bivariate point processes can be obtained similarly.

Using the setup postulated in section 1.2.2, given the random effect $\boldsymbol{u}_i = (u_{i1}, u_{i2})'$, the conditional intensity for the $j \to 3-j$ transition is formulated as

$$\lambda_{ij}(t; \boldsymbol{\theta}_j | \mathcal{H}_i(t), \boldsymbol{u}_i) = \exp(\boldsymbol{x}'_{ij}(t)\boldsymbol{\beta}_j + u_{ij})\ \lambda_{j0}(t; \boldsymbol{\psi}_j | \mathcal{H}_i(t))\ , \tag{4.4.1}$$

where $\lambda_{j0}(t; \boldsymbol{\psi}_j | \mathcal{H}_i(t))$ is the baseline $j \to 3-j$ transition intensity common to all subjects and is completely specified by $\boldsymbol{\psi}_j$, a $q_j \times 1$ vector of parameters, $\boldsymbol{x}_{ij}(t)$ is a $p_j \times 1$ vector of covariates, $\boldsymbol{\beta}_j$ is the corresponding $p_j \times 1$ vector of regression parameters, $\boldsymbol{\theta}_j = (\boldsymbol{\beta}'_j, \boldsymbol{\psi}'_j)'$ and $\boldsymbol{u}_i$ is the subject-specific bivariate random effect which is independent of the covariates, $j = 1, 2$, $i = 1, \ldots m$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)'$.

Recall that $Y_{ij}(t)$ is the indicator for the jth state (section 1.2.2). Let $t_{i1} < \cdots < t_{in_i}$ be the observed transition times for subject i occurring during the course of follow-up. Let $\mathcal{D}_{ij} = \{k | Y_{ij}(t) = 1 \quad \text{for} \quad t_{i,k-1} \leq t < t_{ik}, k = 1, ..., n_i\}$ denote the set of indices for the inter-event intervals over which subject i is observed to be in state j, where $t_{i0} = \tau_{i1}$ and $t_{i,n_i+1} = \tau_{i2}$. Furthermore, let $\mathcal{D}_{ijc} = \mathcal{D}_{ij} \cup \{n_i + 1 | Y_{ij}(t) = 1 \quad \text{for} \quad t_{in_i} \leq t < \tau_{i2}\}$ denote the set $\mathcal{D}_{ij}$ augmented to include the index of the interval between the last observed transition and the censoring time. For simplicity, we assume subsequently that covariates are independent of time.

Let $\Lambda_{j0}(a, b | \mathcal{H}_i(b)) = \int_a^b \lambda_{j0}(t | \mathcal{H}_i(t)) dt$, $j = 1, 2$, $i = 1, ..., m$. Assuming that τ_{i1} is a transition time (or the time origin if it is zero), then it follows that the likelihood for

subject i, conditional on $\boldsymbol{u}_i$, is given by

$$L_i(\boldsymbol{\theta}|\boldsymbol{u}_i) = \prod_{j=1}^{2}\left[\prod_{k\in\mathcal{D}_{ij}} \lambda_{ij}(t_{ik}|\mathcal{H}_i(t_{ik}), u_{ij}) \cdot \exp\left(-\sum_{k\in\mathcal{D}_{ijc}} \Lambda_{ij}(t_{ik-1}, t_{ik}|\mathcal{H}_i(t_{ik}), u_{ij})\right)\right], \tag{4.4.2}$$

and $\Lambda_{ij}(a, b|\mathcal{H}_i(b), u_{ij}) = \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_j + u_{ij})\Lambda_{j0}(a, b|\mathcal{H}_i(b))$. As in many clinical trials, the time of randomization very often does not coincide with a transition time or the time origin. In such cases, one should model the intensity in the first observed duration differently. We will illustrate this modification in next section.

4.4.1 Gauss-Hermite Rule

To facilitate the parameter estimation, we reparameterize the overdispersion parameters as follows. By Cholesky decomposition, the covariance matrix of the mixing distribution can be written as $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}'$ where

$$\boldsymbol{\Omega} = \begin{pmatrix} \omega_1 & 0 \\ \omega_3 & \omega_2 \end{pmatrix},$$

and ω_1, ω_2 and ω_3 are real numbers. Let $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3)'$. The random effects are reparametrized as $\boldsymbol{z}_i = \boldsymbol{\Omega}^{-1}\boldsymbol{u}_i$ where $\boldsymbol{z}_i = (z_{i1}, z_{i2})' \sim N(\boldsymbol{0}, \boldsymbol{I}_2)$ and $\boldsymbol{I}_2$ is the 2×2 identity matrix. Then the marginal likelihood for subject i is given by

$$L_i(\boldsymbol{\phi}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L_i(\boldsymbol{\phi}|\boldsymbol{z}_i)d\Phi(z_{i1})d\Phi(z_{i2}), \tag{4.4.3}$$

where $\boldsymbol{\phi} = (\boldsymbol{\theta}', \boldsymbol{\omega}')'$, $L_i(\boldsymbol{\phi}|\boldsymbol{z}_i)$ is the conditional likelihood (4.4.2) in which the transition intensity is expressed as $\lambda_{ij}(t|\mathcal{H}_i(t), \boldsymbol{z}_i) = \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_j + \boldsymbol{\Omega}_j\boldsymbol{z}_i)\lambda_{j0}(t|\mathcal{H}_i(t))$ with $\boldsymbol{\Omega}_j$ the jth row of $\boldsymbol{\Omega}$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

This specification ensures that all parameters lie on the whole real line so that no restrictions on the scalar parameters are required as part of the estimation procedure. Estimates of σ_j, $j = 1, 2$ and ρ may be obtained by noting that $\sigma_1 = \omega_1^2$, $\sigma_2 = \omega_2^2 + \omega_3^2$ and $\rho = \omega_3/\sqrt{\omega_2^2 + \omega_3^2}$.

The integrations appearing in the likelihood (4.4.3) can be approximated by Gauss-Hermite integration. The bivariate Gauss-Hermite rule approximates integrals of the form

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-z_1^2)\exp(-z_2^2)g(z_1, z_2)dz_1dz_2,$$

by a double sum

$$\sum_{l_1=1}^{R}\sum_{l_2=1}^{R} c_{l_1}c_{l_2}g(z_{l_1}, z_{l_2}),$$

where c_l's are weights, z_l's are nodes, and R represents the number of nodes. Tables for the nodes and weights can be found in Abramowitz and Stegun (1972).

The unconditional likelihood for subject i (4.4.3) can be re-written as

$$L_i(\phi) = \exp(\sum_{j=1}^{2} n_{ij}\boldsymbol{x}_{ij}'\beta_j) \cdot \prod_{j=1}^{2}\prod_{k\in D_{ij}} \lambda_{j0}(t_{ik}) \cdot \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_i(z_{i1}, z_{i2})d\Phi(z_{i1})d\Phi(z_{i2}),$$

where

$$g_i(z_{i1}, z_{i2}) = \exp\left(\sum_{j=1}^{2} n_{ij}u_{ij} - \sum_{j=1}^{2}\exp(\boldsymbol{x}_{ij}'\beta_j + u_{ij})\sum_{k\in D_{ijc}}\Lambda_{j0}(t_{ik-1}, t_{ik})\right),$$

and $\Lambda_{j0}(a, b) = \int_a^b \lambda_{j0}(t|\mathcal{H}(t))dt$. By a change of variable and the Gauss-Hermite rule, the integral in $L_i(\phi)$ is approximated as

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_i(z_{i1}, z_{i2})d\Phi(z_{i1})d\Phi(z_{i2}) \approx \frac{1}{\pi}\sum_{l_1=1}^{R}\sum_{l_2=1}^{R} c_{l_1}c_{l_2}g_i(\sqrt{2}z_{l_1}, \sqrt{2}z_{l_2}) \equiv \frac{1}{\pi}G_i(\phi).$$

Hence, the log-likelihood based on m subjects is approximately equal to

$$\log L(\phi) \approx \sum_{i=1}^{m} \left\{ \sum_{j=1}^{2} n_{ij} x_{ij}' \beta_j + \sum_{j=1}^{2} \sum_{k \in D_{ij}} \log(\lambda_{j0}(t_{ik})) + \log \left[\frac{1}{\pi} G_i(\phi) \right] \right\}. \tag{4.4.4}$$

This function can be maximized by a standard Newton-Raphson method. The $(k+1)$st iteration is given by

$$\hat{\phi}^{(k+1)} = \hat{\phi}^{(k)} - \left(\frac{\partial^2 \log L(\hat{\phi}^{(k)})}{\partial \phi \partial \phi'} \right)^{-1} \frac{\partial \log L(\hat{\phi}^{(k)})}{\partial \phi}. \tag{4.4.5}$$

Furthermore, starting values may be obtained from the maximum likelihood estimates of the fixed-effect model.

4.4.2 The EM Algorithm

Suppose u_i follows a discrete distribution with H mass points such that

$$\Pr(u_i = \xi_h) = \pi_h, \qquad h = 1, ..., H,$$

where the ξ_h's are bivariate mass points and the π_h's are masses. The EM algorithm is then essentially the same as its univariate counterpart (see section 2.5.3).

The number of mass points may be chosen by comparing the distance between mass points. However, this will require a large number of computations when H is even moderately large. Another approach is to use the correlation between the estimates of mass points as a measure of their distance. The idea is that if a single mass point is misspecified as two mass points, the correlation between their estimates in each component, i.e., $\text{corr}(\xi_{hj}, \xi_{kj})$, $j = 1, 2$, should be very high and positive. Therefore, we may combine two mass points if their correlations are larger than 0.95, say. However, it may not be feasible to compute the observed information matrix for each EM cycle. Limited experience showed

that the Hessian matrix of the function to be maximized in the M-step is close to singular when the iterates of some mass points are very near in distance. This suggests that we may obtain approximate correlations from the inverse of the negative Hessian matrix in the M-step, which is already available in each EM cycle.

4.5 The CHEST study

4.5.1 Model Specification

We apply the two methods of estimation described in section 4.4 to the CHEST study (section 1.4.2). The multiplicative components model (1.2.7) proposed in section 1.2.2 is employed to capture the intrinsic properties of the disease process in which the semi-Markov and the Markov components are specified as piecewise constant functions and the seasonal component is specified as a function of quarterly indicators.

More specifically, let state 1 be the AECB-free state in which patients were symptom-free, and state 2 be the AECB state in which symptoms of an exacerbation were manifested. The durations of exacerbation and inter-exacerbation periods were measured in days. We took as the origin of the basic time scale, the date of diagnosis with chronic bronchitis. The baseline transition intensities are given by

$$\lambda_{j0}(t;\psi_j|\mathcal{H}_i(t)) = S_j(c_i(t);\alpha_j)R_j(b_i(t);\gamma_j)T_j(t;\delta_j), \qquad j=1,2, \tag{4.5.1}$$

where $S_j(\cdot)$, $R_j(\cdot)$ and $T_j(\cdot)$ represent the seasonal, the semi-Markov and the Markov components respectively for the $j \to 3-j$ transition. Here $c_i(0)$ is the date of diagnosis of the disease and $c_i(t) = t + c_i(0)$ is the calendar time at t. Furthermore, $b_i(t) = t - t_{N_i.(t-)}$ is the backward recurrence time for subject i at t (i.e. the time since entry to the current state). Let $\psi_j = (\alpha_j', \gamma_j', \delta_j')'$ denote the vector of all unknown parameters in (4.5.1).

Partitioning the time axis into

$$0 = a_{j0} < a_{j1} < \cdots < a_{jq_j} = \infty, \qquad \text{and}$$
$$0 = e_{j0} < e_{j1} < \cdots < e_{jr_j} = \infty,$$

we define the semi-Markov component and the Markov component respectively as

$$R_j(s) = \sum_{k=1}^{q_j} I_{jk}(s) \exp(\gamma_{jk}), \qquad \text{and} \tag{4.5.2}$$
$$T_j(s) = \sum_{k=1}^{r_j} J_{jk}(s) \exp(\delta_{jk}), \qquad j = 1, 2, \tag{4.5.3}$$

where $I_{jk}(s) = 1$ if $a_{j,k-1} \leq s < a_{jk}$ and 0 otherwise, $J_{jk}(s) = 1$ if $e_{j,k-1} \leq s < e_{jk}$ and 0 otherwise, and $\gamma_j = (\gamma_{j1}, ..., \gamma_{jq_j})'$ and $\delta_j = (\delta_{j1}, ..., \delta_{jr_j})'$ are parameters.

By inspection of the data, the range of the duration of chronic bronchitis was 1 to 54 years, suggesting that 5-year intervals were appropriate for the piecewise constant Markov components in both transition intensities: $\{5, 10, 15, 20, 25, 30, 35, 40\} \times 365$. Since the duration of exacerbation was typically relatively short, weekly rates were adopted for the semi-Markov component for the transition out of the AECB state: $\{7, 14, 21, 28, 35, 42\}$. Monthly rates were used for the semi-Markov component of the transition out of the AECB-free state: $\{30, 60, 90, 120, 150, 180, 210\}$.

We model the seasonal effect by partitioning the calendar year into four quarters: the first quarter (winter) is from January 1 to March 31; the second quarter (spring) is from April 1 to June 30; the third quarter (summer) is from July 1 to September 30; and the fourth quarter (fall) is from October 1 to December 31. The seasonal component is defined by

$$S_j(s) = \sum_{k=1}^{4} Q_k(s) \exp(\alpha_{jk}), \qquad j = 1, 2, \tag{4.5.4}$$

where $Q_k(s) = 1$ if s is in the kth quarter and 0 otherwise, and $\alpha_j = (\alpha_{j1}, ..., \alpha_{j4})'$ are the

parameters.

For the purpose of identifiability, we constrain $\alpha_{j1} = \gamma_{j1} = \delta_{j1} = 0$ and include the constant 1 in $\boldsymbol{x}_{ij}$ for $j = 1, 2$, $i = 1, ...m$.

There are two complicating features of the design of this trial. First, selection bias was induced by requiring patients to be in the AECB state at the time of entry and randomization. The second related point is that the time of randomization did not coincide with a transition time, but the transition time just prior to the time of randomization was available here. In the presence of heterogeneity, there is no satisfactory solution to these two problems without discarding the first incomplete observed durations. Here we consider a rough adjustment by specifying the distribution of the first observed duration differently from the distribution of the subsequent durations.

Since the treatment Ciprofloxacin was not given prior to the time of randomization, the treatment variable was a time dependent covariate for $\lambda_{i2}(t|\mathcal{H}_i(t), \boldsymbol{u}_i)$. The conditional transition intensities given $\boldsymbol{u}_i$ are expressed as

$$\begin{aligned} \lambda_{i1}(t|\mathcal{H}_i(t), \boldsymbol{u}_i) &= \exp(\boldsymbol{x}_{i1}'\boldsymbol{\beta}_1 + u_{i1})\lambda_{10}(t|\mathcal{H}_i(t)), \\ \lambda_{i2}(t|\mathcal{H}_i(t), \boldsymbol{u}_i) &= \exp(\boldsymbol{x}_{i2}'(t)\boldsymbol{\beta}_2 + u_{i2})\lambda_{20}(t|\mathcal{H}_i(t)), \end{aligned} \tag{4.5.5}$$

where $\boldsymbol{x}_{i2}(t) = (\tilde{\boldsymbol{x}}_{i2}'(t), d_i(t)\tilde{\boldsymbol{x}}_{i2}'(t))'$, $\tilde{\boldsymbol{x}}_{i2}(t) = (1, trt_i(t), \boldsymbol{x}_{i2}')'$, $trt_i(t) = 0$ before randomization and $trt_i(t) = 1$ if Ciprofloxacin was given and $trt_i(t) = 0$ if standard care was given after randomization, $d_i(t) = 1$ if $t \leq t_{i1}$ and 0 otherwise, $\boldsymbol{x}_{i2}$ is a vector of other time independent covariates, and $\boldsymbol{x}_{i1}$ is a vector of time independent covariates including the treatment variable. As we model only the first observed AECB duration differently, it seems sufficient to stratify the semi-Markov component of the AECB to AECB-free baseline transition:

$$R_2(b_i(t)) = \sum_{k=1}^{q_2} I_{2k}(b_i(t)) \exp(\gamma_{2k} + \gamma_{2,q_2+k} d_i(t)), \tag{4.5.6}$$

where $\gamma_{21} = \gamma_{2,q_2+1} = 0$. We call (4.5.5) and (4.5.6) the initial model.

4.5.2 Test of Homogeneity

Assuming that the correlation between random effects is zero, we carry out the score test of homogeneity proposed in section 3.6.2 for the CHEST data. As a first step, we estimated a fixed-effect model for (4.5.5) in which a backward elimination procedure was used to select significant covariates at the 10% level of significance. As interest lies in the effect of Ciprofloxacin, the treatment variables were always included in the models. The other covariates identified as prognostic variables were the duration of AECB symptoms at randomization for the AECB to AECB-free transition, and gender and severe bronchitis for the AECB-free to AECB transition. As a second step, the tests of homogeneity were carried out with these covariates included. The test statistic for $H_0 : \sigma_1 = \sigma_2 = 0$ was equal to 5.605 ($p = 0.061$), which indicated that there was some evidence against homogeneity. The individual test statistics for $H_0 : \sigma_1 = 0$ and $H_0 : \sigma_2 = 0$ were 1.151 ($p = 0.250$) and 2.069 ($p = 0.039$) respectively. Therefore, there was evidence that mild heterogeneity existed which was mainly due to the AECB to AECB-free transition. Nevertheless, we fit the full random effect model to illustrate the procedure and inferences.

4.5.3 Parameter Estimates

We first used 12-node Gauss-Hermite integration, as described in section 4.4.1, to compute the marginal likelihood, the corresponding score functions and the information matrix. Again a backward elimination procedure was used to select important risk factors at the 10% level of significance; treatment variables were always included in the models. The maximum likelihood estimates of the regression parameters, the variances, and the correlation of the random effects with approximate 95% confidence intervals for the initial model, are given in the left panel of Table 4.8.

Second, although the QN1 algorithm (section 4.2.1) was used to speed up the EM algorithm for the non-parametric random effect model, the computational time is still quite long compared with the Gauss-Hermite integration. Since the simulation study in section 4.2 suggests that inference for covariate effects is typically quite robust to the GH and NP estimations, the non-parametric random effect model was fit with the same set of covariates chosen by the GH estimation. We started with 10 mass points and allowed mass points to combine if their distance is small. The estimation result is given in the right panel of Table 4.8.

The estimates of the covariate effects and the seasonality parameters agree quite closely for these two models, although the standard errors for the seasonality parameter estimates are larger in the non-parametric random effect model. The estimates of the variances of the random effects are smaller in the non-parametric than the log-normal random effect models. The 95% confidence intervals for σ are unacceptably wide in the non-parametric model, which is due to the large variance estimates for $\hat{\sigma}$. Moreover, the number of mass points is estimated to be only 4 (Table 4.9). The masses concentrate at a single point $(-4.025, -5.988)$, indicating that there is very mild degree of heterogeneity. This is in close agreement with the result of the test of homogeneity.

The model can be further refined. Based on the log-normal random effect model, we found little evidence of the need for the Markov component in the AECB to AECB-free transition (not shown here) and the approximate 95% confidence interval for the correlation ρ suggested that a reduced model without these parameters might be sufficient. We therefore fit a reduced model (i.e. without the Markov component for the AECB to AECB-free transition and the correlation between the random effects); the results are given in Table 4.10. The likelihood ratio test for the reduced model against the initial model gave a statistic 14.224 with 9 degrees of freedom ($p = 0.115$), which indicated that there was insufficient evidence to claim that the reduced model was significantly inferior to the initial model. This reduced model was also fit with non-parametric random effects. Again,

	Log-normal						Non-parametric					
	AECB to AECB-free Transition			AECB-free to AECB Transition			AECB to AECB-free Transition			AECB-free to AECB Transition		
Covariate[1]	Estimate	s.e.	p-value	Estimate	s.e.	p-value	Estimate	s.e.	p-value	Estimate	s.e.	p-value
	First Observed Duration						*First Observed Duration*					
intercept	-4.586	0.323	—	—[2]	—	—	—	—	—	—	—	—
treatment	0.547	0.154	—	—	—	—	0.526	0.151	< 0.001	—	—	—
symptoms	-0.106	0.016	< 0.001	—	—	—	-0.111	0.022	< 0.001	—	—	—
	Second and Subsequent Durations						*Second and Subsequent Durations*					
intercept	-4.143	0.224	-	-5.973	0.246	—	—	—	—	—	—	—
treatment	0.041	0.126	0.746	-0.038	0.131	0.770	0.100	0.109	0.357	-0.044	0.164	0.788
gender	—	—	—	0.276	0.136	0.043	—	—	—	0.257	0.149	0.083
severity	—	—	—	0.642	0.186	0.001	—	—	—	0.536	0.152	< 0.001
symptoms	-0.007	0.010	0.438	—	—	—	-0.008	0.010	0.457	—	—	—
Seasonal	Estimate	s.e.	p-value	Estimate	s.e.	p-value	Estimate	s.e.	p-value	Estimate	s.e.	p-value
α_2	0.376	0.117	0.001	-0.505	0.151	0.001	0.386	0.128	0.003	-0.525	0.246	0.033
α_3	0.159	0.144	0.269	-0.305	0.144	0.035	0.138	0.166	0.406	-0.321	0.226	0.155
α_4	0.290	0.121	0.016	-0.187	0.141	0.186	0.310	0.154	0.044	-0.178	0.202	0.378
	Estimate		95% C.I.				Estimate		95% C.I.			
σ_1	0.153		(0.069,0.338)				0.133		(0.002,8.861)			
σ_2	0.277		(0.132,0.581)				0.160		(0.022,1.163)			
ρ	0.278		(−0.245,0.676)				0.024		(−0.995,0.995)			

[1]treatment = 1 if subject was given ciprofloxacin at randomization and 0 otherwise;
gender = 1 if female and 0 if male; severity = 1 if chronic bronchitis is severe and 0 otherwise;
and symptoms = days of AECB symptoms at randomization.
[2]Not applicable.

Table 4.8: Parameter estimates for the initial model.

Mass points	Mass
(-4.889, -5.746)	0.214
(-4.374, -4.655)	0.041
(-4.025, -5.988)	0.725
(-2.866, -4.123)	0.020

Table 4.9: Estimated mass points and masses for the initial model.

estimates for covariate effects are very similar.

For the AECB to AECB-free transition, the duration of AECB symptoms at randomization remained in the model. Both the treatment variable and the duration of AECB symptoms at randomization were significant for the first observed duration of symptoms. Specifically, Ciprofloxacin increased the rate of resolution of AECB and patients with a longer duration of symptoms at randomization had lower rate of resolution. None of the covariates, however, had a significant effect in the second and subsequent AECB durations. For the AECB-free to AECB transition, we found that the treatment variable was insignificant but gender and severe bronchitis were significant. Specifically, female patients and patients with severe bronchitis tended to have higher rates of relapse of AECB.

The semi-Markov component of the AECB to AECB-free transition changed after the first observed duration from gradually increasing over time to leveling off to about exp(2) on the third week in the subsequent durations (Figure 4.2). On the other hand, the semi-Markov component of the AECB-free to AECB transition had an irregular pattern (Figure 4.3).

As mentioned earlier, the AECB to AECB-free transition did not have a Markov component. The AECB-free to AECB transition, however, showed a slightly increasing trend over time (Figure 4.3), suggesting that patients with a long history of chronic bronchitis had higher rates of relapse to AECB.

Compared to the first quarter, the AECB to AECB-free transition intensity was signif-

	Log-normal						Non-parametric					
	AECB to AECB-free Transition			AECB-free to AECB Transition			AECB to AECB-free Transition			AECB-free to AECB Transition		
Covariate[1]	Estimate	s.e.	p-value	Estimate	s.e.	p-value	Estimate	s.e.	p-value	Estimate	s.e.	p-value
	First Observed Duration											
intercept	-4.742	0.310	—	—[2]	—	—	—	—	—	—	—	—
treatment	0.498	0.155	0.001	—	—	—	0.440	0.141	0.002	—	—	—
symptoms	-0.108	0.016	< 0.001	—	—	—	-0.111	0.020	< 0.001	—	—	—
	Second and Subsequent Durations											
intercept	-4.294	0.192	—	-5.973	0.247	—	—	—	—	—	—	—
treatment	0.046	0.130	0.721	-0.039	0.132	0.771	0.038	0.105	0.720	-0.101	0.164	0.537
gender	—	—	—	0.255	0.136	0.062	—	—	—	0.222	0.140	0.113
severity	—	—	—	0.631	0.189	0.001	—	—	—	0.569	0.174	0.001
symptoms	-0.009	0.009	0.359	—	—	—	-0.007	0.010	0.485	—	—	—
Seasonal	Estimate	s.e.	p-value	Estimate	s.e.	p-value	Estimate	s.e.	p-value	Estimate	s.e.	p-value
α_2	0.386	0.118	0.001	-0.505	0.151	0.001	0.369	0.120	0.002	-0.539	0.242	0.026
α_3	0.143	0.144	0.321	-0.308	0.144	0.032	0.142	0.158	0.367	-0.333	0.228	0.144
α_4	0.301	0.120	0.012	-0.190	0.141	0.179	0.308	0.144	0.032	-0.174	0.200	0.384
	Estimate			95% Confidence Interval			Estimate			95% C.I.		
σ_1	0.189			(0.095,0.376)			0.110			(0.001,14.667)		
σ_2	0.290			(0.141,0.597)			0.171			(0.033,0.888)		

[1] See Table 4.8.
[2] Not applicable.

Table 4.10: Parameter estimates for the reduced model.

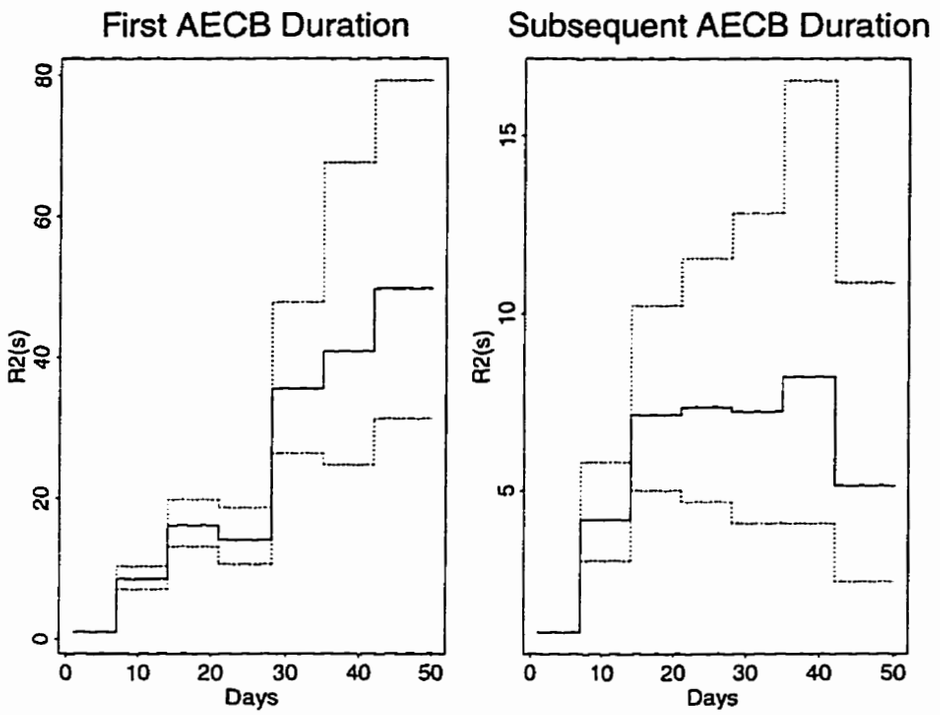

Figure 4.2: Estimates (solid line) and approximate 95% pointwise confidence intervals (dotted line) of the semi-Markov components of the AECB to AECB-free transition in the first observed AECB duration (left) and in the subsequent observed AECB durations (right) based on the reduced model (see (4.5.6)).

icantly higher in the second and the fourth quarters, but it had no significant difference in the third quarter. On the other hand, the AECB-free to AECB transition intensity was significantly lower in the second and the third quarter than the first quarter, but it was not significantly different in the fourth quarter from the first quarter. In other words, patients tended to have longer duration of AECB in winter and summer, and higher rate of relapse of AECB in winter and fall.

We also fit the model without Markov components in both transition intensities and with uncorrelated random effects. However, the likelihood ratio test indicated that the Markov component was an important feature of the AECB-free to AECB transition intensity (test statistic = 18.272 with 8 degrees of freedom ($p = 0.019$)). Furthermore, models

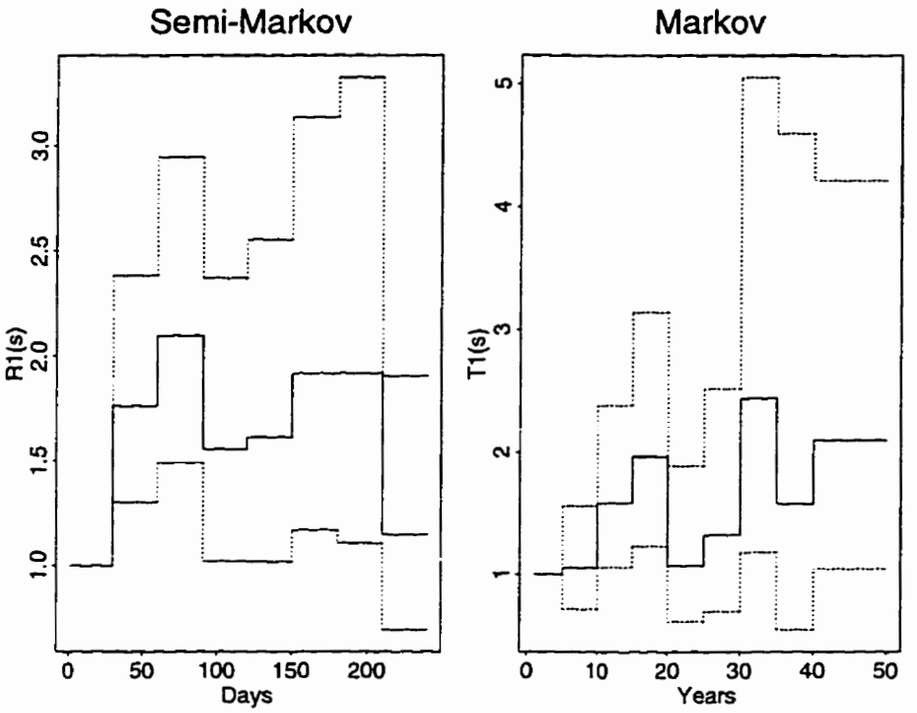

Figure 4.3: Estimates (solid line) and approximate 95% pointwise confidence intervals (dotted line) of the semi-Markov component (left) and the Markov component (right) of the AECB-free to AECB transition based on the reduced model.

with finer sub-divisions for the Markov and semi-Markov components also gave similar estimates of the covariate effects and the variances of the random effects, so the present findings appear quite robust.

4.5.4 Remark

To demonstrate the effect of "spurious" heterogeneity, we fit the data again by using a random effect model based on alternating renewal processes, which is a popular model. We adopted a further simplification by ignoring the difference between the first observed duration and the subsequent durations and assumed that the baseline intensities are of

Weibull form:

$$\lambda_{j0}(t) = \exp(\gamma_j) t^{\exp(\gamma_j)-1}, \qquad j = 1, 2,$$

where t is the backward recurrence time. Using a log-normal mixing distribution, we found that the estimates of σ_1, σ_2 and σ_{12} are 0.569, 0.568 and -0.143 respectively (Ng and Cook, 1996). The estimates of σ_1 and σ_2 are highly inflated compared with a more comprehensive model considered above.

4.6 Concluding Remarks and Discussion

4.6.1 General Remarks

In this chapter, we examined the performance of two common methods of estimation for random effect models based on Poisson and renewal processes. The regression coefficients can be estimated with negligible bias, and valid variance estimates can also be obtained. This desirable result is quite robust to the methods considered and the true mixing distribution. In contrast, the GH estimate for σ may have small positive bias if the true mixing distribution is highly discrete. Although the NP estimate for σ is unbiased, its variance estimate may be unrealistically large, probably due to the ignorance of the variability of the estimated number of mass points. In applications, using either the GH or the NP methods of estimation makes little difference with regard to covariate effects. Choosing between the GH and the NP methods should be based on practical considerations such as the ease of implementation, and prior information about the data, e.g., discrete random effects should be used if hidden subgroups are suspected.

We also propose a quite general model for analyzing data arising from common chronic conditions in which multiple time scales are incorporated to model simultaneously degenerative features of the conditions, cyclical and seasonal patterns in the disease process. When we applied this model to data from a study of chronic bronchitis, we identified sev-

eral important risk factors. We found that the transition intensities depend on more than one time scale, and found there to be mild subject-to-subject variation. Most of these clinical findings had not been found empirically before, suggesting that the two-state model provides added insight into this sort of disease activity.

Although we have not considered model diagnostics for the model proposed in the analysis of the CHEST study, the estimation result based on a simplified model specification (section 4.5.4) suggests that it is important to specify appropriate intensity functions. The analysis of the CHEST study also illustrates the practical value of a weakly parametric baseline intensity using a piecewise constant function.

Finally, we remark that a satisfactory modeling strategy should consist of the following steps: model specification; estimation; and model diagnostics. Tests of homogeneity can serve as a model diagnostic, but general methods for model diagnostics are currently lacking for random effect models, especially for point processes. We will address model diagnostics in the last chapter. The next subsection discusses an important issue in event history analysis: the problem of left-truncation.

4.6.2 Left-truncation

In many studies, the time of randomization does not coincide with a transition time, nor does it represent the beginning of the entire process. In this case, the data are said to be left-truncated because only the remaining durations after the time of randomization are observed and thus the first observed durations are incomplete. As a consequence, the selection probability of a subject may depend on the length of one or more sojourns. For example, in the CHEST study, subjects with longer exacerbation durations are more likely to be selected since one of the entry criteria stipulated that patients must be experiencing an exacerbation at randomization. This selection bias (also called length-biased sampling) must be accounted for to make valid inferences.

Consider an orderly univariate point process with intensity $\lambda(t|\mathcal{H}(t))$, where $\mathcal{H}(t)$ is the history of the process up to time t^- defined in (1.2.1) with the subscript i dropped. Let $[\tau_1, \tau_2]$ be the observation period and $t_1 < \cdots < t_n$ be the observed event times. Let $t_{-1} < \tau_1$ be the event time just prior to τ_i. If t_{-1} is known, Guo (1993) proposed a conditional argument to obtain the likelihood which amounts to adjusting the contribution to the likelihood from the first observed duration. Assuming that $T_1 = t_1$ is observed, then $\mathcal{H}(t_1) = \{T_1 > \tau_1, T_{-1} = t_{-1}, N(t), \boldsymbol{X}(t), Y(t), \tau_1 \leq t < t_1\}$. Thus,

$$\begin{aligned} \Pr[T_1|\mathcal{H}(t_1)] &= \frac{\lambda(t_1|\mathcal{H}(t_1))\exp(-\int_{t_{-1}}^{t_1}\lambda(t|\mathcal{H}(t))dt)}{\exp(-\int_{t_{-1}}^{\tau_1}\lambda(t|\mathcal{H}(t))dt)} \\ &= \lambda(t_1|\mathcal{H}(t_1))\exp(-\int_{\tau_1}^{t_1}\lambda(t|\mathcal{H}(t))dt). \end{aligned}$$

This probability does not depend on t_{-1} for Poisson processes, but in general, is a function of t_{-1}. Hence, the likelihood is given by

$$L = \prod_{j=1}^{n}\lambda(t_j|\mathcal{H}(t_j)) \cdot \exp(-\int_{\tau_1}^{\tau_2}\lambda(t|\mathcal{H}(t))dt), \tag{4.6.1}$$

where $t_0 = \tau_1$.

If t_{-1} is unknown, we have to multiply L by the density of T_{-1} and integrate this product with respect to t_{-1}. However, unless there is prior information on T_{-1} or the process is Poisson in which case L is independent of t_{-1}, the probability density of T_{-1} is difficult to specify. In the former case, discarding the first observed duration seems to be the simplest way to achieve valid inference, although part of the information will be lost. For example, 167 out of 222 patients in the CHEST study have more than one transitions during their observation periods and thereby 55 patients will be excluded from the analysis, and of the remaining 167 subjects, considerably fewer exacerbations will be contributed. Another approach is to model the first observed duration differently from the subsequent durations as we did in the analysis of the CHEST study.

In the presence of random effects, (4.6.1) is the conditional likelihood given the random effect. The marginal likelihood can be maximized by the methods considered in this chapter.

In the context of multi-state processes with random effects, selection bias can be due to extreme subject-specific random effects. The conditional argument considered by Guo (1993) may not be able to correct this selection bias. The approach we used in the analysis of the CHEST study may serve as an approximate adjustment for the selection bias. As a rough check on the appropriateness of this approach, we fit a Cox proportional hazard model for the first observed durations and found that the treatment variable had significant effect. This result is consistent with the finding based on the stratified model (4.5.6). Nevertheless, further research in this direction is necessary.

In general, the construction of likelihood function under left-truncation relies on the information available before the time of randomization and the sampling scheme. Lawless and Fong (1997) recently described some difficulties and modeling approaches pertaining to this problem.

Chapter 5

Robust Inference for Bivariate Point Processes

5.1 Overview

Inference for random effect models considered in chapter 4 is likelihood-based, which requires a full specification of the process conditional on the random effects. Although a certain degree of robustness against misspecification can be achieved through the use of a weakly parametric baseline intensity and a non-parametric mixing distribution, appropriate time scales must still be specified. For situations in which one is not sure about the appropriate time scale and when interest lies primarily on studying covariate effects on the number of events, models based directly on the mean number of events over time may be entertained. This approach belongs to another popular class of models for longitudinal data analysis and is termed "marginal models" (Diggle et al., 1994). Unlike random effect models, marginal models only require minimal assumptions on the probabilistic structure of the process. Lawless (1995) reviewed a number of methods of analysis for recurrent events using both conditional (intensity function) models and marginal models. A variety

of conditional models were considered in chapter 4.

In this chapter, we propose marginal models for the mean and covariance functions of bivariate point processes for which inference is based on the theory of estimating functions. Specifically, we construct unbiased linear estimating functions for the marginal cumulative mean functions (CMFs) of bivariate point processes, in which the correlation structure for the jumps can be taken into account. A criterion for obtaining the optimal estimating function for the CMF is suggested in section 5.3. Tests of hypotheses using Wald-type and score-type tests are discussed in section 5.4. Some estimating functions for the covariance functions are also considered in section 5.5. Since the correlation structure is usually unknown and any assumed structures are difficult to verify, we suggest using the estimating function arising from mixed bivariate Poisson processes as a working estimating function. These mixed bivariate Poisson estimating functions are easy to implement and yet provide a "working" correlation to account for the underlying correlation structure, analogous to the generalized estimating equation (GEE) used in discrete time longitudinal data (Liang and Zeger, 1986). The performance of the mixed bivariate Poisson estimating function is examined through simulations in section 5.7. In section 5.8, the proposed model is applied to data from the asthma trial described in chapter 1. Finally, some concluding remarks and discussion are given in section 5.9.

The rest of this section is devoted to a brief review of the marginal approach for univariate point processes using estimating function. The idea originated from quasi-likelihood (QL) estimation, which is most commonly applied in the context of generalized linear models (GLM) (McCullagh and Nelder, 1989 Chapter 9). By specifying the mean and covariance of the response over time, and introducing covariate effects, one can construct a quasi-likelihood which behaves in many ways like an ordinary likelihood function.

Let $\boldsymbol{Y}_i$ be an $n \times 1$ vector of responses and $\boldsymbol{x}_i$ be a $p \times 1$ vector of covariates for the ith subject. The mean response is assumed to be a function of the linear predictor, $\boldsymbol{x}_i'\boldsymbol{\beta}$,

through a link function,

$$\mu_i = \mathrm{E}(\boldsymbol{Y}_i) = g(\boldsymbol{x}_i'\boldsymbol{\beta}),$$

and the variance is a function of μ_i,

$$\mathrm{var}(\boldsymbol{Y}_i) = \phi V_i(\mu_i),$$

where $\phi > 0$ is a dispersion parameter. The QL estimator for $\boldsymbol{\beta}$ based on a sample $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_m\}$, is obtained as the solution to the quasi-score function

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{m} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} V_i^{-1}(\mu_i)(\boldsymbol{Y}_i - \mu_i) = 0.$$

The dispersion parameter is often estimated by the method of moments.

Under some mild regularity conditions, the QL estimator has asymptotic properties similar to the ordinary maximum likelihood estimator, namely consistency and asymptotic normality. In fact, the quasi-score functions are the optimal estimating functions among the class of unbiased linear estimating functions provided the mean and covariance of the response are correctly specified (McCullagh and Nelder, 1989 Chapter 9).

In many longitudinal studies, the responses from the same subject constitute a time series whose the autocorrelations may be difficult to specified, especially for discrete responses. In a series of articles (Liang and Zeger, 1986; Zeger and Liang, 1986; Zeger et al., 1988; Liang et al., 1992), generalized estimating equations (GEE) are introduced in which only the mean function has to be correctly specified and the covariance matrix may take a convenient form,

$$V(\mu) = A(\mu)R(\rho)A(\mu),$$

where $A(\mu) = diag(\mathrm{var}(\boldsymbol{Y}))^{1/2}$ and $R(\rho)$ is a correlation matrix parameterized by ρ. $R(\rho)$ is not necessarily the true correlation matrix, and thus it is referred to as a "working"

correlation matrix. These authors showed that, using the results of White (1982), the GEE estimators are consistent and asymptotically normally distributed even if the covariance structure is misspecified. As the generalized estimating equations belong to the class of unbiased linear estimating functions, the efficiency is higher if the working correlation matrix is closer to the true correlation matrix.

Inspired by this approach, Lawless and Nadeau (1995) proposed a robust model for univariate point processes in which estimation is based on a Poisson likelihood. The estimation is robust with respect to misspecification of the distribution. Let $N(t)$ be the number of events occurring over $(0, t]$ and $dN(t) = N(t) - N(t-) = \lim_{h \to 0^+}(N(t) - N(t-h))$ be the number of events (or jumps) at the instant t. The cumulative mean function (CMF) for $N(t)$ is defined as $\mathrm{E}[N(t)] = \Lambda(t)$ which is assumed to be continuous, non-decreasing and differentiable with respect to t if the time scale is continuous. The CMF can be modeled parametrically or non-parametrically. For the ease of exposition, we consider a parametric CMF which is completely specified by a $p \times 1$ vector $\boldsymbol{\theta}$. Regression models may also be entertained with multiplicative covariate effects of the form $\Lambda(t) = \exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t)$, where $\boldsymbol{x}$ is a vector of covariates and $\Lambda_0(t)$ is the baseline mean function.

We note that the probabilistic structure of the process is not fully specified unless it is a Poisson process. Lawless and Nadeau (1995) argued that parameter estimates based on Poisson models are valid quite generally, and derive robust variance estimates using the theory of estimating functions.

We define a function $\nu(t)$ such that $\nu(t) = t$ for continuous time processes and $\nu(t) = [t]$ for discrete time processes. Let $\lambda(t) = d\Lambda(t)/d\nu(t)$. Based on a sample $\{N_1(t), ..., N_m(t)\}$, the Poisson estimating function (PEF) can be shown to be

$$U(\boldsymbol{\theta}) = \sum_{i=1}^{m} \int_0^{\infty} Y_i(t) \frac{\partial \log(\lambda_i(t))}{\partial \boldsymbol{\theta}} [dN_i(t) - \lambda_i(t) d\nu(t)], \tag{5.1.1}$$

where $Y_i(t)$ is independent of $N_i(t)$. The process $Y_i(t)$ usually indicates whether subject i

is under observation at time t. The robust asymptotic variance of the estimate is given by

$$\text{var}(\hat{\boldsymbol{\theta}}) = \text{E}\left(-\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right)^{-1} \text{E}(U(\boldsymbol{\theta})U(\boldsymbol{\theta})')\text{E}\left(-\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{-1}.$$

When the jump sizes are uncorrelated, i.e., $\text{cov}(dN(s), dN(t)) = 0$ for $s \neq t$, Nadeau and Lawless (1996) showed that the PEF is optimal among all unbiased linear estimating functions. In general, (5.1.1) may produce inefficient estimates for correlated jump processes, although consistency is still preserved. Nadeau and Lawless (1996) extended the PEF to incorporate the covariance structure of the process. If the CMF and the covariance between jumps ($\text{cov}(dN(s), dN(t))$) are correctly specified, inference can be based on a quadratic estimating function of the form

$$\begin{aligned} U(\boldsymbol{\phi}) \;=\; & \sum_{i=1}^{m} \left\{ \int_0^\infty Y_i(t)\boldsymbol{a}_i(t)dN_i^c(t) + \right. \\ & \left. \int_0^\infty \int_0^\infty Y_i(s)Y_i(t)\boldsymbol{b}_i(s,t)\left[dN_i^c(s)dN_i^c(t) - \text{cov}(dN_i^c(s), dN_i^c(t))\right] \right\} \quad (5.1.2) \end{aligned}$$

where

$$\begin{aligned} dN_i^c(t) &= dN_i(t) - \lambda_i(t)d\nu(t), \\ \text{cov}(dN_i^c(s), dN_i^c(t)) &= I(s=t)v_i(t)d\nu(t) \\ &\quad + I(s \neq t)c_i(s,t)d\nu(s)d\nu(t), \end{aligned}$$

$v_i(t)$ and $c_i(s,t)$ are variance and covariance functions with parameters $\boldsymbol{\phi} = (\boldsymbol{\theta}', \boldsymbol{\sigma}')'$ respectively, $\boldsymbol{a}_i(t)$ and $\boldsymbol{b}_i(s,t)$ are known differentiable functions of $\boldsymbol{\phi}$, and $I(\cdot)$ is the indicator function.

The optimal quadratic estimating function requires knowledge of the third and the fourth moments of the process which are usually unavailable (Godambe and Thompson, 1989). Instead of working with (5.1.2) directly, Nadeau and Lawless consider breaking

(5.1.2) into two components:

$$\begin{aligned} U_1(\boldsymbol{\theta};\boldsymbol{\sigma}) &= \sum_{i=1}^{m}\int_0^\infty Y_i(t)\boldsymbol{a}_i(t)dN_i^c(t), \\ U_2(\boldsymbol{\sigma};\boldsymbol{\theta}) &= \sum_{i=1}^{m}\int_0^\infty\int_0^\infty Y_i(s)Y_i(t)\boldsymbol{b}_i(s,t)\left[dN_i^c(s)dN_i^c(t) - \mathrm{cov}(dN_i^c(s), dN_i^c(t))\right], \end{aligned}$$

where the dimensions of $\boldsymbol{a}_i(t)$ and $\boldsymbol{b}_i(s,t)$ are the same as the dimensions of $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ respectively. Let $U(\boldsymbol{\phi}) = (U_1(\boldsymbol{\theta};\boldsymbol{\sigma})', U_2(\boldsymbol{\sigma};\boldsymbol{\theta})')'$.

This leads to a 2-step estimation procedure in which we solve $U_1(\boldsymbol{\theta};\boldsymbol{\sigma}) = 0$ for $\boldsymbol{\theta}$ given $\boldsymbol{\sigma}$ and substitute this solution into $U_2(\boldsymbol{\sigma};\boldsymbol{\theta}) = 0$ to find a root of $\boldsymbol{\sigma}$. This iterative procedure continues until convergence is met. Let $\hat{\boldsymbol{\phi}}$ be the solution to $U(\boldsymbol{\phi}) = 0$. It can be shown that the asymptotic variance of $\hat{\boldsymbol{\phi}}$ is block diagonal with the partition conformable to $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$, and the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is equal to

$$\mathrm{asvar}(\hat{\boldsymbol{\theta}}) = \mathrm{E}\left(-\frac{\partial U(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}'}\right)^{-1}\mathrm{E}(U(\boldsymbol{\phi})U(\boldsymbol{\phi})')\mathrm{E}\left(-\frac{\partial U(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}}\right)^{-1}.$$

This implies that even when the covariance is not correctly specified, $\hat{\boldsymbol{\theta}}$ is still consistent, whereas the estimator for $\boldsymbol{\theta}$ obtained from (5.1.2) may not be. Furthermore, given $\boldsymbol{\sigma}$, an optimal linear estimating function for $\boldsymbol{\theta}$ exists in which $\boldsymbol{a}_i(t)$ is a function of $\lambda_i(t)$, $v_i(t)$ and $c_i(s,t)$; see Theorem 1 of Nadeau and Lawless (1996) for the expression of $\boldsymbol{a}_i(t)$ in the optimal estimating function. Inference for $\boldsymbol{\theta}$ can be based on the asymptotic normality of $\hat{\boldsymbol{\theta}}$.

5.2 Model Formulation

The marginal approach for univariate point processes may be extended to multivariate point processes. In addition to the marginal means and covariances of jumps, we have to specify the cross-covariances of jumps between components of a process as well. For the

ease of exposition, we focus on bivariate point processes in this chapter. We propose a robust method via the estimating function in which optimal linear estimating functions for the parameters in the mean functions can be obtained along the line of Nadeau and Lawless (1996).

Suppose there are m independent bivariate point processes $\{\boldsymbol{N}_1(t), ..., \boldsymbol{N}_m(t)\}$, where $\boldsymbol{N}_i(t) = (N_{i1}(t), N_{i2}(t))'$ such that $N_{i1}(t)$ and $N_{i2}(t)$ are two orderly point processes characterizing two types of events. Suppose $\boldsymbol{N}_i(t)$ is observed over $(0, \tau_i]$ or at time points $T_i = \{t_{i1}, ..., t_{im_i}\}$ for continuous time processes or discrete time processes, respectively. Let $Y_i(t)$ be the indicator function that the ith process is under observation at time t, i.e. $Y_i(t) = I(t \leq \tau_i)$ for continuous time or $Y_i(t) = I(t \in T_i)$ for discrete time. Therefore, for any function $g(t)$,

$$\int_0^\infty Y_i(t)g(t)d\nu(t) = \begin{cases} \int_0^{\tau_i} g(t)dt & \text{for continuous time,} \\ \sum_{t \in T_i} g(t) & \text{for discrete time.} \end{cases}$$

We assume that the censoring (or observation) process $\{Y_i(t)\}$ is independent of the bivariate point process $\{N_i(t)\}$. Consider parametric models and define the following marginal moments for the jumps:

$$\begin{aligned} \mathrm{E}(dN_{ij}(t)) &= \lambda_{ij}(t;\boldsymbol{\theta}_j)d\nu(t), \\ \mathrm{var}(dN_{ij}(t)) &= v_{ij}(t;\boldsymbol{\theta}_j)d\nu(t), \\ \mathrm{cov}(dN_{ij}(s), dN_{ij}(t)) &= I(s=t)v_{ij}(t;\boldsymbol{\theta}_j)d\nu(t) \\ &\quad + I(s \neq t)c_{ij}(s,t;\boldsymbol{\sigma}_j,\boldsymbol{\theta}_j)d\nu(s)d\nu(t), \\ \mathrm{cov}(dN_{i1}(s), dN_{i2}(t)) &= I(s=t)v_{i,12}(t;\boldsymbol{\sigma}_{12},\boldsymbol{\theta})d\nu(t) \\ &\quad + I(s \neq t)c_{i,12}(s,t;\boldsymbol{\sigma}_{12},\boldsymbol{\theta})d\nu(s)d\nu(t), \end{aligned}$$

for $j = 1, 2$, where $\boldsymbol{\theta}_j$ is a $p_j \times 1$ vector of parameters for the mean functions, $j = 1, 2$, and $\boldsymbol{\sigma}_j$ is a $q_j \times 1$ vector of parameters for the covariance functions, $j = 1, 2$ and 12. We assume that there is no common element in $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ and $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1', \boldsymbol{\sigma}_2', \boldsymbol{\sigma}_{12}')'$. The mean rate functions $\lambda_{ij}(\cdot)$, $j = 1, 2$, the variance rate functions $v_{ij}(\cdot)$, $j = 1, 2$ and the simultaneous covariance rate function $v_{i,12}(\cdot)$ are assumed to be non-negative and cadlag. The covariance rate functions $c_{ij}(\cdot, \cdot)$, $j = 1, 2, 12$ are cadlag such that $c_{ij}(s, t) = c_{ij}(t, s)$, $j = 1, 2$. Covariates may be incorporated into the mean functions and covariance functions. In what follows, inference is based conditionally on the censoring process and the covariate process which is assumed to be external.

The CMFs are parameterized by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively. We note that for continuous time processes, $dN_{ij}(t)$ is a 0–1 random variable and thus $v_{ij}(t) = \lambda_{ij}(t)$ and $v_{i,12}(t)dt$ is the probability that both types of events occur simultaneously.

Let

$$\Sigma_i(t) = \begin{pmatrix} v_{i1}(t) & v_{i,12}(t) \\ v_{i,12}(t) & v_{i2}(t) \end{pmatrix} \quad \text{and} \quad \Omega_i(s,t) = \begin{pmatrix} c_{i1}(s,t) & c_{i,12}(s,t) \\ c_{i,12}(t,s) & c_{i2}(s,t) \end{pmatrix},$$

for $s \neq t$. For convenience, we set $\Omega_i(t, t) = 0$ for $t \geq 0$. Assuming that $\Sigma_i(t)$ is positive definite for $t \geq 0$. Since our interest usually lies in estimating $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we assume that the other parameters ($\boldsymbol{\sigma}$) are known for the moment.

5.3 Optimal Estimating Functions For CMF

A linear unbiased estimating function (LUEF) for $\boldsymbol{\theta}$ is defined by

$$U_{1,2}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \int_0^\infty Y_i(t)\boldsymbol{a}_i(t)[dN_i(t) - \boldsymbol{\lambda}_i(t)d\nu(t)], \tag{5.3.1}$$

where $\boldsymbol{a}_i(t)$ is a $(p_1 + p_2) \times 2$ matrix function of $\boldsymbol{\phi} = (\boldsymbol{\theta}', \boldsymbol{\sigma}')'$ and $\boldsymbol{\lambda}_i(t) = (\lambda_{i1}(t), \lambda_{i2}(t))'$. Let $\mathcal{U}_{1,2}$ be the set of all LUEFs of the form (5.3.1).

An optimal estimating function for $\boldsymbol{\theta}$ may be derived in a fashion similar to Theorem 1 of Nadeau and Lawless (1996). We state the result as follows:

Proposition 5.1 *Let* $R_i = \{t \in [0, \infty) | Y_i(t) = 1\}$ *for* $i = 1, ..., m$. *Let* $F_i(s, t; \boldsymbol{\phi}) : R_i \times R_i \longrightarrow (-\infty, \infty) \times (-\infty, \infty)$ *be a* 2×2 *matrix of functions of* $\boldsymbol{\phi}$ *such that* $F_i(s, t)' = F_i(t, s)$ *for* $s, t \in R_i$, *and satisfying*

$$\int_0^\infty Y_i(u) F_i(s, u) \Omega_i(u, t) d\nu(u) + F_i(s, t) \Sigma_i(t) + \Sigma_i^{-1}(s) \Omega_i(s, t) = 0, \tag{5.3.2}$$

for $s, t \in R_i$ *and* $i = 1, ..., m$. *Then the optimal LUEF for* $\boldsymbol{\theta}$ *is given by* $U_{1,2}^*(\boldsymbol{\theta})$ *for which*

$$\boldsymbol{a}_i^*(t) = \frac{\partial \boldsymbol{\lambda}_i(t)}{\partial \boldsymbol{\theta}} \Sigma_i^{-1}(t) + G_i(t), \tag{5.3.3}$$

for $t \in R_i$ *and* $i = 1, ..., m$, *where*

$$G_i(t) = \int_0^\infty Y_i(s) \frac{\partial \boldsymbol{\lambda}_i(s)}{\partial \boldsymbol{\theta}} F_i(s, t) d\nu(s), \tag{5.3.4}$$

which satisfies

$$\int_0^\infty Y_i(s) \boldsymbol{a}_i^*(s) \Omega_i(s, t) d\nu(s) + G_i(t) \Sigma_i(t) = 0, \tag{5.3.5}$$

for $t \in R_i$.

Proof: Let

$$U_{1,2,i}(\boldsymbol{\theta}) = \int_0^\infty Y_i(t) \boldsymbol{a}_i(t) [d\boldsymbol{N}_i(t) - \boldsymbol{\lambda}_i(t) d\nu(t)].$$

A well-known result of the theory of estimating functions states that

$$U_{1,2}^*(\boldsymbol{\theta}) = \sum_{i=1}^m U_{1,2,i}^*(\boldsymbol{\theta}) \in \mathcal{U}_{1,2}$$

is optimal within $\mathcal{U}_{1,2}$ if for any $U_{1,2}(\boldsymbol{\theta}) \in \mathcal{U}_{1,2}$, we have

$$\mathrm{E}(U_{1,2,i}(\boldsymbol{\theta})U^{*}_{1,2,i}(\boldsymbol{\theta})') = \mathrm{E}\left(-\frac{\partial U_{1,2,i}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right), \tag{5.3.6}$$

for $i = 1, ..., m$ (Godambe and Heyde, 1987). This criterion implies that the optimal estimating function is the projection of the score function onto $\mathcal{U}_{1,2}$. To see $U^{*}_{1,2}(\boldsymbol{\theta})$ satisfies (5.3.6), pre-multiply the left-hand side of (5.3.2) with $Y_i(s)\partial\boldsymbol{\lambda}_i(s)/\partial\boldsymbol{\theta}$, and then integrate with respect to s to get

$$\int_0^\infty Y_i(s)\boldsymbol{a}^{*}_i(s)\Omega_i(s,t)d\nu(s) = -G_i(t)\Sigma_i(t), \tag{5.3.7}$$

for $t \in R_i$. The optimal criterion (5.3.6) can be proved by noting that,

$$\begin{aligned}
\mathrm{E}(U_{1,2,i}(\boldsymbol{\theta})U^{*}_{1,2,i}(\boldsymbol{\theta})') &= \int_0^\infty Y_i(t)\boldsymbol{a}_i(t)\Sigma_i(t)Y_i(t)\boldsymbol{a}^{*}_i(t)'d\nu(t) \\
&\quad + \int_0^\infty \int_0^\infty Y_i(s)\boldsymbol{a}_i(s)\Omega_i(s,t)Y_i(t)\boldsymbol{a}^{*}_i(t)'d\nu(s)d\nu(t) \\
&= \int_0^\infty Y_i(t)\boldsymbol{a}_i(t)\Sigma_i(t)(\boldsymbol{a}^{*}_i(t)' - G_i(t)')d\nu(t), \\
&\quad \text{(by (5.3.7) and } \Omega_i(s,t)' = \Omega_i(t,s)), \\
&= \mathrm{E}\left(-\frac{\partial U_{1,2,i}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right), \\
&\quad \text{(as } \boldsymbol{a}^{*}_i(t) = \frac{\partial\boldsymbol{\lambda}_i(t)}{\partial\boldsymbol{\theta}}\Sigma_i^{-1}(t) + G_i(t)).
\end{aligned}$$

Hence, $\boldsymbol{a}^{*}_i(t)$ gives the optimal LUEF for $\boldsymbol{\theta}$. Finally, (5.3.7) proves the result (5.3.5). □

The optimal weight function $\boldsymbol{a}^{*}_i(t)$ consists of two parts. The first term in (5.3.3) leads to the optimal LUEF for uncorrelated jump processes, while the second term in (5.3.3) can be viewed as a correction factor that incorporates the covariance structure of jumps.

Consider a special case for which $dN_{i1}(s)$ and $dN_{i2}(t)$ are uncorrelated. This gives $v_{i,12}(t) = c_{i,12}(s,t) = 0$ and $F_i(s,t)$ becomes a diagonal matrix. It is readily seen that

Proposition 5.1 leads to two optimal LUEFs for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively. These optimal LUEFs are exactly the results of Nadeau and Lawless (1996) obtained by treating the components of the bivariate processes separately.

It is in general difficult to solve (5.3.2) or (5.3.5). The following corollary provides a solution for a particular type of covariance function.

Corollary 5.1 *If the covariance function is of the form*

$$\Omega_i(s,t) = A_i(s)B_i(t), \qquad s,t \in R_i, i = 1,...,m, \tag{5.3.8}$$

where $A_i(s)$ and $B_i(t)$ are 2×2 matrices such that $B_i(t)$ is invertible for all $t \in R_i$, then the optimal weight function is given by

$$\begin{aligned} \boldsymbol{a}_i^*(t) = \frac{\partial \boldsymbol{\lambda}_i(t)}{\partial \boldsymbol{\theta}} \Sigma_i^{-1}(t) - \left(\int_0^\infty Y_i(s) \frac{\partial \boldsymbol{\lambda}_i(s)}{\partial \boldsymbol{\theta}} \Sigma_i^{-1}(s) A_i(s) d\nu(s) \right) \times \\ \left(I_2 + \int_0^\infty Y_i(s) B_i(s) \Sigma_i^{-1}(s) A_i(s) d\nu(s) \right)^{-1} B_i(t) \Sigma_i^{-1}(t), \end{aligned} \tag{5.3.9}$$

for $t \in R_i$ and $i = 1,...,m$, where I_2 is the 2×2 identity matrix.

Proof: Since $B_i(t)$ is invertible, equation (5.3.5) can be expressed as

$$\int_0^\infty Y_i(s) G_i(s) A_i(s) d\nu(s) + \int_0^\infty Y_i(s) \frac{\partial \boldsymbol{\lambda}_i(s)}{\partial \boldsymbol{\theta}} \Sigma_i^{-1}(s) A_i(s) d\nu(s) + G_i(t) \Sigma_i(t) B_i^{-1}(t) = 0.$$

We observe that the two integrals in the above equation are time independent and so is $H \equiv G_i(t)\Sigma_i(t)B_i^{-1}(t)$. Thus, we easily find H as

$$H = - \left(\int_0^\infty Y_i(s) \frac{\partial \boldsymbol{\lambda}_i(s)}{\partial \boldsymbol{\theta}} \Sigma_i^{-1}(s) A_i(s) d\nu(s) \right) \left(I_2 + \int_0^\infty Y_i(s) B_i(s) \Sigma_i^{-1}(s) A_i(s) d\nu(s) \right)^{-1}.$$

Hence, the solution for $G_i(t)$ is $G_i(t) = HB_i(t)\Sigma_i^{-1}(t)$, $t \in R_i$ and $i = 1,...,m$. □

We are now going to derive the asymptotic variance for the solution to an arbitrary LUEF (5.3.1). Let $\hat{\boldsymbol{\theta}}$ be the solution to (5.3.1) for $\boldsymbol{\theta}$. Given $\boldsymbol{\sigma}$, the asymptotic variance for $\hat{\boldsymbol{\theta}}$ is given by the so-called sandwich estimator:

$$asvar(\hat{\boldsymbol{\theta}}) = (A'_m(\boldsymbol{\theta}))^{-1} B_m(\boldsymbol{\theta}) A_m^{-1}(\boldsymbol{\theta}),$$

where

$$\begin{aligned} A_m(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial U_{1,2}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] &= \sum_{i=1}^{m} \int_0^\infty Y_i(t) \frac{\partial \boldsymbol{\lambda}_i(t)}{\partial \boldsymbol{\theta}} \boldsymbol{a}_i'(t) d\nu(t), \\ B_m(\boldsymbol{\theta}) = \mathrm{E}\left[U_{1,2}(\boldsymbol{\theta}) U'_{1,2}(\boldsymbol{\theta})\right] &= \sum_{i=1}^{m} \int_0^\infty Y_i(t) \boldsymbol{a}_i(t) \Sigma_i(t) \boldsymbol{a}_i'(t) d\nu(t) \\ &\quad + \sum_{i=1}^{m} \int_0^\infty \int_0^\infty Y_i(s) Y_i(t) \boldsymbol{a}_i(s) \Omega_i(s,t) \boldsymbol{a}_i'(t) d\nu(s) d\nu(t). \end{aligned}$$

If the specification of the covariance is in doubt, an empirical covariance estimate in which B_m is replaced by its empirical form,

$$B_m^e(\boldsymbol{\theta}) = \sum_{i=1}^{m} U_{1,2,i}(\boldsymbol{\theta}) U'_{1,2,i}(\boldsymbol{\theta}),$$

is often used (White, 1982; Liang and Zeger, 1986).

In particular, $A_m(\boldsymbol{\theta}) = B_m(\boldsymbol{\theta})$ for optimal LUEF with correct covariance specification and thereby, the asymptotic variance becomes

$$asvar(\hat{\boldsymbol{\theta}}) = A_m^{-1}(\boldsymbol{\theta}),$$

which is the variance formula in the quasi-likelihood estimation; see Godambe and Heyde (1987) for the discussion of estimating functions and quasi-likelihood.

Under some mild regularity conditions (White, 1982), the estimate $\hat{\boldsymbol{\theta}}$ is approximately normally distributed with mean $\boldsymbol{\theta}$ and variance $asvar(\hat{\boldsymbol{\theta}})$ which can be estimated by sub-

stituting $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$.

5.4 Tests of Hypotheses

Tests of hypotheses concerning $\boldsymbol{\theta}$ can be carried out using the asymptotic normality of $\hat{\boldsymbol{\theta}}$ or the estimating function (5.3.1). The former is a Wald-type test and the latter is a score-type test. Specifically, consider the null hypothesis

$$H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0,$$

where $\boldsymbol{\psi}$ is a r-subvector of $\boldsymbol{\theta}$ and for simplicity, we let $\boldsymbol{\theta} = (\boldsymbol{\psi}', \boldsymbol{\pi}')'$.

The Wald test is based on the statistic

$$W_{\boldsymbol{\psi}} = (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)'(V_{\boldsymbol{\psi}}^W(\hat{\boldsymbol{\theta}}))^{-1}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0), \tag{5.4.1}$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\psi}}', \hat{\boldsymbol{\pi}}')'$ is the estimating function estimate for $\boldsymbol{\theta}$ without restrictions, and $V_{\boldsymbol{\psi}}^W(\hat{\boldsymbol{\theta}})$ is the asymptotic variance of $\hat{\boldsymbol{\psi}}$ evaluated at $\hat{\boldsymbol{\theta}}$. Under H_0, $W_{\boldsymbol{\psi}}$ is asymptotically χ^2 distributed with degrees of freedom r.

The score test uses the estimating function as the test statistic. Partition $U_{1,2}(\boldsymbol{\theta})$ into two components $(U_{\boldsymbol{\psi}}(\boldsymbol{\theta})', U_{\boldsymbol{\pi}}(\boldsymbol{\theta})')$ conformably to $\boldsymbol{\psi}$ and $\boldsymbol{\pi}$. Given σ, an estimate for $\boldsymbol{\pi}$ under H_0, $\tilde{\boldsymbol{\pi}}$, is obtained by solving $U_{\boldsymbol{\pi}}(\boldsymbol{\pi}; \boldsymbol{\psi}_0) = 0$ for $\boldsymbol{\pi}$. Denote $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\psi}_0', \tilde{\boldsymbol{\pi}}')'$. Standard Taylor series expansion for $U_{\boldsymbol{\psi}}(\tilde{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0 = (\boldsymbol{\psi}_0', \boldsymbol{\pi}')'$ under H_0 gives

$$U_{\boldsymbol{\psi}}(\tilde{\boldsymbol{\theta}}) = U_{\boldsymbol{\psi}}(\boldsymbol{\theta}_0) - \frac{\partial U_{\boldsymbol{\psi}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\pi}'} \left(\frac{\partial U_{\boldsymbol{\pi}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\pi}} \right)^{-1} U_{\boldsymbol{\pi}}(\boldsymbol{\theta}_0) + O_p(n^{-1}).$$

We also partition the matrices A_m and B_m conformably to ψ and π as

$$A_m = \begin{pmatrix} A_{11,m} & A_{12,m} \\ A_{21,m} & A_{22,m} \end{pmatrix} \quad \text{and} \quad B_m = \begin{pmatrix} B_{11,m} & B_{12,m} \\ B_{21,m} & B_{22,m} \end{pmatrix},$$

where the dependence of θ is suppressed. The asymptotic variance of $U_\psi(\tilde{\theta})$ is equal to

$$\begin{aligned} V_\psi^S(\tilde{\theta}) &= B_{11,m} + A'_{21,m}A_{22,m}^{-1}B_{22,m}(A'_{22,m})^{-1}A_{21,m} \\ &\quad - B_{12,m}(A'_{22,m})^{-1}A_{21,m} - A'_{21,m}A_{22,m}^{-1}B_{21,m}. \end{aligned}$$

The optimal estimating function provides a simpler form because $A_m = B_m$, so that

$$V_\psi^S(\tilde{\theta}) = A_{11,m} - A_{12,m}A_{22,m}^{-1}A_{21,m}.$$

The score statistic is then given by

$$S_\psi = U'_\psi(\tilde{\theta})(V_\psi^S(\tilde{\theta}))^{-1}U_\psi(\tilde{\theta}), \tag{5.4.2}$$

which is asymptotically χ^2 distributed with degrees of freedom r under H_0.

In addition, the empirical covariance matrix B_m^e can be used in place of B_m for W_ψ and S_ψ if the covariance structure is unclear. We remark that the test statistics require an estimate for σ which may be obtained from a saturated model or a full regression model which includes a large number of covariates (Breslow, 1990).

Breslow (1990) studied the performance of Wald and score tests for overdispersed Poisson regression. He demonstrated, based on simulations, that the Wald and score tests derived from quasi-likelihood generally have reasonable performance in terms of size and power. He recommended the use of a score test with the empirical covariance matrix in practice because this empirical score test not only is robust to the specification of covari-

ance, but also performs satisfactorily even for small sample sizes and large overdispersion. Similar results should hold for bivariate point processes considered here, especially when the optimal estimating function is used.

5.5 Joint Estimation of CMF and Covariances

The discussion in previous sections relies on the knowledge of the covariance parameter σ. However, the parameters σ are usually unknown. Quadratic unbiased estimating functions (QUEFs) for σ of the following forms may be constructed, given θ, for $j = 1, 2$,

$$\begin{aligned} U_{jj}(\sigma_j) &= \sum_{i=1}^{m} \int_0^\infty \int_0^\infty Y_i(s)Y_i(t)\boldsymbol{b}_{ij}(s,t;\phi)[dN_{ij}^c(s)dN_{ij}^c(t) - \mathrm{cov}(dN_{ij}(s), dN_{ij}(t))], \\ U_{12}(\sigma_{12}) &= \sum_{i=1}^{m} \int_0^\infty \int_0^\infty Y_i(s)Y_i(t)\boldsymbol{b}_{i,12}(s,t;\phi)[dN_{i1}^c(s)dN_{i2}^c(t) - \mathrm{cov}(dN_{i1}(s), dN_{i2}(t))], \end{aligned} \tag{5.5.1}$$

where $dN_{ij}^c(t) = dN_{ij}(t) - \lambda_{ij}(t)d\nu(t)$ and $\boldsymbol{b}_{ij}(s,t)$ is a known function of dimension q_j for $j = 1, 2, 12$. It should be noted that U_{jj} and U_{12} are in fact functions of ϕ and they aim to estimate σ_j and σ_{12} respectively. In the estimation procedure, we fix the values of the parameters other than σ_j in U_{jj} to find a solution for σ_j. The other estimating functions are treated similarly. The procedure proceeds by solving $U_{1,2}(\theta) = 0$, $U_{11}(\sigma_1) = 0$, $U_{22}(\sigma_2) = 0$ and $U_{12}(\sigma_{12}) = 0$ iteratively. Due to the unbiasedness of these estimating functions, the resulting estimates are consistent.

The next question is how to choose the weight functions $\boldsymbol{b}_{ij}(s,t)$. A convenient candidate is $b_{ij}(s,t) = 1$ for $j = 1, 2, 12$ which gives the moment estimators for σ if $q_j = 1$. An optimal QUEF requires knowledge of the third and the fourth moments of $dN_{ij}(t)$ (Crowder, 1987; Godambe and Thompson, 1989) which are often unavailable. The orderly continuous time process is an exception. Recall that $v_{ij}(t) = \lambda_{ij}(t)$ for $j = 1, 2$. It is easily

seen that $(dN_{ij}^c(t))^2 = (1-2\lambda_{ij}(t)dt)dN_{ij}^c(t) + \lambda_{ij}(t)dt - (\lambda_{ij}(t)dt)^2$ and thus,

$$\begin{aligned}\mathrm{var}(dN_{ij}^c(s)dN_{ij}^c(t)) &= (1-2\lambda_{ij}(s)ds)(1-2\lambda_{ij}(t)dt)\mathrm{E}(dN_{ij}^c(s)dN_{ij}^c(t)) \\ &= I(s=t)\lambda_{ij}(t)dt + I(s\neq t)c_{ij}(s,t)dsdt + o(ds,dt),\end{aligned}$$

and $\mathrm{var}(dN_{i1}(s)dN_{i2}(t)) = I(s=t)v_{i,12}(t)dt + I(s\neq t)c_{i,12}(s,t)dsdt + o(ds,dt)$ similarly, where $o(ds,dt) = o(ds) + o(dt) + o(dsdt)$. Estimating functions similar to the PEF may be constructed as follows, for $j = 1, 2$,

$$\begin{aligned}U_{jj}(\boldsymbol{\sigma}_j) &= \sum_{i=1}^{m}\left\{\int_0^\infty Y_i(t)\frac{\partial\log(\lambda_{ij}(t))}{\partial\boldsymbol{\sigma}_j}[dN_{ij}^c(t)-\lambda_{ij}(t)dt]\right. \\ &\quad \left. + \int_0^\infty\int_0^\infty Y_i(s)Y_i(t)I(s\neq t)\frac{\partial\log(c_{ij}(s,t))}{\partial\boldsymbol{\sigma}_j}[dN_{ij}^c(s)dN_{ij}^c(t)-c_{ij}(s,t)dsdt]\right\}, \\ &= \sum_{i=1}^{m}\left\{\int_0^\infty\int_0^\infty Y_i(s)Y_i(t)I(s\neq t)\frac{\partial\log(c_{ij}(s,t))}{\partial\boldsymbol{\sigma}_j}[dN_{ij}^c(s)dN_{ij}^c(t)-c_{ij}(s,t)dsdt]\right\},\end{aligned} \tag{5.5.2}$$

$$\begin{aligned}U_{12}(\boldsymbol{\sigma}_{12}) &= \sum_{i=1}^{m}\left\{\int_0^\infty Y_i(t)Y_i(t)\frac{\partial\log(v_{i,12}(t))}{\partial\boldsymbol{\sigma}_{12}}[dN_{i1}^c(t)dN_{i2}^c(t)-v_{i,12}(t)dt]\right\} + \\ &\quad \sum_{i=1}^{m}\left\{\int_0^\infty\int_0^\infty Y_i(s)Y_i(t)I(s\neq t)\frac{\partial\log(c_{i,12}(s,t))}{\partial\boldsymbol{\sigma}_{12}}[dN_{i1}^c(s)dN_{i2}^c(t)-c_{i,12}(s,t)dsdt]\right\}.\end{aligned} \tag{5.5.3}$$

For discrete time processes, we may use the moment estimator or define some working covariance structures as suggested by Prentice and Zhao (1991).

The asymptotic variance for the joint estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ can be derived similarly as in section 5.3. We stack the estimating functions (5.3.1) and (5.5.1) together and write

$$U(\boldsymbol{\phi}) = (U_{1,2}(\boldsymbol{\theta})', U_{11}(\boldsymbol{\sigma}_1)', U_{22}(\boldsymbol{\sigma}_2)', U_{12}(\boldsymbol{\sigma}_{12})')'.$$

The asymptotic variance of $\hat{\boldsymbol{\phi}}$ is given by

$$\text{asvar}(\hat{\boldsymbol{\phi}}) = \mathrm{E}\left(-\frac{\partial U(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right)^{-1} \mathrm{E}\left(U(\boldsymbol{\phi})U(\boldsymbol{\phi})'\right) \mathrm{E}\left(-\frac{\partial U(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right)^{-1}. \tag{5.5.4}$$

The expectation $\mathrm{E}\left(U(\boldsymbol{\phi})U(\boldsymbol{\phi})'\right)$ involves third and fourth moments of the jumps which are not available in most cases. An empirical variance formula similar to B_m^e considered in section 5.3 can be used.

Since $\lambda_{ij}(t)$ does not depend on $\boldsymbol{\sigma}$, $E(-\partial U_{1,2}(\boldsymbol{\theta})/\partial\boldsymbol{\sigma}) = 0$ and thus $E(-\partial U(\boldsymbol{\phi})/\partial\boldsymbol{\phi})$ is an upper triangular matrix in which the lower $q_\cdot \times p_\cdot$ submatrix is equal to zero, where $p_\cdot = p_1 + p_2$ and $q_\cdot = q_1 + q_2 + q_{12}$. Therefore, the asymptotic variance of $\hat{\boldsymbol{\theta}}$ remains unchanged whether $\boldsymbol{\sigma}$ is known or consistently estimated. Furthermore, $\hat{\boldsymbol{\phi}}$ is asymptotically distributed as $N(\boldsymbol{\phi}, \text{asvar}(\hat{\boldsymbol{\phi}}))$.

The estimating functions are easily computed for discrete time processes. Here we express the computational forms for continuous time processes. Let $0 \le t_{ij1} < \cdots < t_{ijm_j} \le \tau_i$ be the observed event times in $\{N_{ij}(t)\}$. The optimal LUEF for $\boldsymbol{\theta}$ and the estimating functions for $\boldsymbol{\sigma}$ using (5.5.2) and (5.5.3) can be expressed as follows:

$$\begin{aligned}
U^*_{1,2}(\boldsymbol{\theta}) &= \sum_{i=1}^{m}\left\{\sum_{j=1}^{2}\left(\sum_{t\in\mathcal{D}_{ij}} a^*_{ij}(t) - \int_0^{\tau_i} a^*_{ij}(t)\lambda_{ij}(t)dt\right)\right\}, \\
U_{jj}(\boldsymbol{\sigma}_j) &= \sum_{i=1}^{m}\left\{\sum_{\{s,t\in\mathcal{D}_{ij}|s\neq t\}} \frac{\partial \log(c_{ij}(s,t))}{\partial \boldsymbol{\sigma}_j} - 2\sum_{s\in\mathcal{D}_{ij}}\int_0^{\tau_i}\frac{\partial \log(c_{ij}(s,t))}{\partial \boldsymbol{\sigma}_j}\lambda_{ij}(t)dt \right. \\
&\qquad \left. + \int_0^{\tau_i}\int_0^{\tau_i}\frac{\partial \log(c_{ij}(s,t))}{\partial \boldsymbol{\sigma}_j}(\lambda_{ij}(s)\lambda_{ij}(t) - c_{ij}(s,t))dsdt\right\}, \\
U_{12}(\boldsymbol{\sigma}_{12}) &= \sum_{i=1}^{m}\left\{\sum_{t\in\mathcal{D}_i}\frac{\partial \log(v_{i,12}(t))}{\partial \boldsymbol{\sigma}_{12}} - \int_0^{\tau_i}\frac{\partial v_{i,12}(t)}{\partial \boldsymbol{\sigma}_{12}}dt + \right. \\
&\qquad \sum_{\{s\in\mathcal{D}_{i1},t\in\mathcal{D}_{i2}|s\neq t\}}\frac{\partial \log(c_{i,12}(s,t))}{\partial \boldsymbol{\sigma}_{12}} - \sum_{t\in\mathcal{D}_{i2}}\int_0^{\tau_i}\frac{\partial \log(c_{i,12}(s,t))}{\partial \boldsymbol{\sigma}_{12}}\lambda_{i1}(s)ds \\
&\qquad \left. - \sum_{s\in\mathcal{D}_{i1}}\int_0^{\tau_i}\frac{\partial \log(c_{i,12}(s,t))}{\partial \boldsymbol{\sigma}_{12}}\lambda_{i2}(t)dt\right.
\end{aligned}$$

$$+\int_0^{\tau_i}\int_0^{\tau_i}\frac{\partial \log(c_{i,12}(s,t))}{\partial \sigma_{12}}(\lambda_{i1}(s)\lambda_{i2}(t)-c_{i,12}(s,t))dsdt\Big\}, \qquad (5.5.5)$$

where $\boldsymbol{a}_{ij}^{*}(t)$ is the jth column of $\boldsymbol{a}_i^{*}(t)$, $\mathcal{D}_{ij}=\{t_{ij1},...,t_{ijm_j}\}$, for $j=1,2$ and $i=1,...,m$, and $\mathcal{D}_i=\mathcal{D}_{i1}\bigcap\mathcal{D}_{i2}$ is the collection of the simultaneous event times for the ith process.

The above integrations may have to be computed numerically. In such situations, simple numerical integration methods such as the Simpson's rule may be used.

5.6 Applications

5.6.1 Mixed Bivariate Poisson Process

We consider a continuous time mixed bivariate Poisson process in which conditional on a bivariate random effect $\boldsymbol{Z}=(Z_1,Z_2)'$, the bivariate process comprises two independent Poisson processes. Suppose $\mathrm{E}(\boldsymbol{Z})=\mathbf{1}$ and

$$\mathrm{var}(\boldsymbol{Z})=\Sigma_Z=\begin{pmatrix}\sigma_1 & \sigma_{12}\\ \sigma_{12} & \sigma_2\end{pmatrix}$$

is positive definite. The conditional moments of jumps are given by

$$\begin{aligned}\mathrm{E}(dN_{ij}(t)|\boldsymbol{Z}_i) &= z_{ij}\lambda_{ij}(t)dt,\\ \mathrm{cov}(dN_{ij}(s),dN_{ij}(t)|\boldsymbol{Z}_i) &= I(s=t)z_{ij}\lambda_{ij}(t)dt,\\ \mathrm{cov}(dN_{i1}(s),dN_{i2}(t)|\boldsymbol{Z}_i) &= 0.\end{aligned}$$

The marginal moments can be shown to be

$$\begin{aligned}\mathrm{E}(dN_{ij}(t)) &= \lambda_{ij}(t)dt,\\ \mathrm{cov}(dN_{ij}(s),dN_{ij}(t)) &= I(s=t)\lambda_{ij}(t)dt+I(s\neq t)\sigma_j\lambda_{ij}(s)\lambda_{ij}(t)dsdt,\end{aligned}$$

$$\text{cov}(dN_{i1}(s), dN_{i2}(t)) = I(s \neq t)\sigma_{12}\lambda_{i1}(s)\lambda_{i2}(t)dsdt.$$

Therefore, $v_{i,12}(t) = 0$, $c_{ij}(s,t) = \sigma_j\lambda_{ij}(s)\lambda_{ij}(t)$ and $c_{i,12}(s,t) = \sigma_{12}\lambda_{i1}(s)\lambda_{i2}(t)$. By noting that $\Omega_i(s,t) = \Sigma_i(s)\Sigma_Z\Sigma_i(t)$ and using Corollary 5.1, the optimal LUEF for $\boldsymbol{\theta}$ is equal to

$$U_{1,2}(\boldsymbol{\theta}) = \sum_{i=1}^{m}\int_0^{\tau_i}\frac{\partial\log(\boldsymbol{\lambda}_i(t))}{\partial\boldsymbol{\theta}}[d\boldsymbol{N}_i(t) - \boldsymbol{\lambda}_i(t)dt] - \sum_{i=1}^{m}\frac{\partial\boldsymbol{\Lambda}_i(\tau_i)}{\partial\boldsymbol{\theta}}M_i[\boldsymbol{N}_i(\tau_i) - \boldsymbol{\Lambda}_i(\tau_i)], \quad (5.6.1)$$

where

$$\begin{aligned} M_i &= \Sigma_Z\left(I_2 + \int_0^{\infty} Y_i(u)\Sigma_i(u)du\Sigma_Z\right)^{-1} \\ &= \frac{1}{K_i}\begin{pmatrix} \sigma_1(1+\sigma_2\Lambda_2(\tau_i)) - \sigma_{12}^2\Lambda_2(\tau_i) & \sigma_{12} \\ \sigma_{12} & \sigma_2(1+\sigma_1\Lambda_1(\tau_i)) - \sigma_{12}^2\Lambda_1(\tau_i) \end{pmatrix}, \\ K_i &= (1+\sigma_1\Lambda_1(\tau_i))(1+\sigma_2\Lambda_2(\tau_i)) - \sigma_{12}^2\Lambda_1(\tau_i)\Lambda_2(\tau_i). \end{aligned}$$

The estimators for $\boldsymbol{\sigma}$ obtained from (5.5.5) turn out to be:

$$\begin{aligned} \hat{\sigma}_j &= \frac{\sum_{i=1}^m[(n_{ij} - \hat{\Lambda}_{ij}(\tau_i))^2 - n_{ij}]}{\sum_{i=1}^m \hat{\Lambda}_{ij}^2(\tau_i)}, \qquad j = 1,2, \\ \hat{\sigma}_{12} &= \frac{\sum_{i=1}^m(n_{i1} - \hat{\Lambda}_{i1}(\tau_i))(n_{i2} - \hat{\Lambda}_{i2}(\tau_i))}{\sum_{i=1}^m \hat{\Lambda}_{i1}(\tau_i)\hat{\Lambda}_{i2}(\tau_i)}, \end{aligned} \quad (5.6.2)$$

where $n_{ij} = N_{ij}(\tau_i)$. The estimator for σ_j is approximately equal to its moment estimator which is

$$\frac{\sum_{i=1}^m[(n_{ij} - \hat{\Lambda}_{ij}(\tau_i))^2 - \hat{\Lambda}_{ij}(\tau_i)]}{\sum_{i=1}^m \hat{\Lambda}_{ij}^2(\tau_i)},$$

for $j = 1,2$, while $\hat{\sigma}_{12}$ is exactly the moment estimator for σ_{12}. In cases when $\hat{\sigma}_j$ is negative, it is usually set to 0.

Furthermore, when regression models are entertained, covariate effects are often mod-

eled multiplicatively to the mean functions:

$$\lambda_{ij}(t) = \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j)\lambda_{0j}(t;\boldsymbol{\gamma}_j),$$

where $\lambda_{0j}(t)$ is the baseline rate completely specified by a vector of parameters $\boldsymbol{\gamma}_j$, $j = 1, 2$ and $i = 1, ..., m$. It remains to specify the forms of the baseline rates. A weakly parametric specification using piecewise constant functions may be used if there is no prior knowledge of appropriate forms.

The optimal estimating function (5.6.1) arising from mixed bivariate Poisson processes may have great potential for practical applications. In most situations, any particular covariance structure is unlikely to be correctly specified in a model. Thus, the covariance structure implied by mixed bivariate Poisson processes can be viewed as a working covariance similar in spirit to the GEE working correlation (Liang and Zeger, 1986). We call (5.6.1) the *mixed Poisson estimating function (MPEF)*. The merit of the MPEF is that it provides consistent estimates and valid inference for the parameters in the CMFs through the use of a robust variance estimate with the efficiency being higher the "closer" the working covariance is to the true covariance. The parameters in $\boldsymbol{\sigma}$ arising from the "working" random effects should be viewed as dispersion parameters in general. In practice, the empirical formula of the sandwich estimator should be used. Some simulation studies will be considered in the next section to investigate the performance of the MPEF.

In the case of equal duration of follow-up for each subject, i.e., $\tau_i = \tau$ for all i, the estimates for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\sigma}$ are invariant to the forms of the baseline rates. This can be seen from the following arguments. Let $\alpha_{0j} = \Lambda_{0j}(\tau)$, $j = 1, 2$ where $\Lambda_{0j}(t) = \int_0^t \lambda_{0j}(s)ds$. Given $\boldsymbol{\sigma}$, the estimating function for $\boldsymbol{\beta}_j$ is given by, from (5.6.1),

$$U_{\beta_j}(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{m} [r_{ij} - (r_{ij}M_{i,jj} + r_{i,3-j}M_{i,12})\Lambda_{ij}]\, \boldsymbol{x}_{ij},$$

$$= \sum_{i=1}^{m} \frac{1}{K_i} [(1 + \sigma_{3-j}\Lambda_{i,3-j})r_{ij} - \sigma_{12}\Lambda_{ij}r_{i,3-j}]\, \boldsymbol{x}_{ij}, \qquad (5.6.3)$$

where $\Lambda_{ij} = \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j + \alpha_{0j})$, $r_{ij} = n_{ij} - \Lambda_{ij}$, $j = 1, 2$, and $M_{i,kl}$ is the (k,l)th element of M_i, $k, l = 1, 2$. Hence, $U_{\beta_j}(\boldsymbol{\theta})$ depends on the baseline rates only through α_{0j}. Similarly, $\hat{\sigma}$ obtained from (5.6.2) is also invariant to the forms of the baseline rates. Therefore, estimation of the regression coefficients and overdispersion parameters is essentially based on the total numbers of events occurring in $(0, \tau)$. In fact, (5.6.3) is equivalent to the quasi-score equation for the overdispersed Poisson regression.

5.6.2 Longitudinal Analysis of Bivariate Count Data

We consider a longitudinal study in which each subject experiences two types of events and is observed at a set of fixed discrete time points, for example, dates of the follow-up visit in a clinical trial. Suppose without loss of generality that the ith subject is observed at times $R_i = \{1, 2, ..., \tau_i\}$. Let $D_{ij}(t)$ be the number of j type events for subject i occurring in the interval $(t-1, t]$ for $j = 1, 2$ and $i = 1, ..., m$. In other words, $D_{ij}(t) = dN_{ij}(t)$. The specifications for the means and covariances are, using the above notation,

$$\begin{aligned}
\mathrm{E}[D_{ij}(t)] &= \lambda_{ij}(t; \boldsymbol{\theta}_j), \\
\mathrm{cov}(D_{ij}(s), D_{ij}(t)) &= I(s = t)v_{ij}(t; \boldsymbol{\theta}_j) + I(s \neq t)c_{ij}(s, t; \sigma_j, \boldsymbol{\theta}_j), \\
\mathrm{cov}(D_{i1}(s), D_{i2}(t)) &= I(s = t)v_{i,12}(t; \sigma_{12}; \boldsymbol{\theta}) + I(s \neq t)c_{i,12}(s, t; \sigma_{12}; \boldsymbol{\theta}),
\end{aligned}$$

for $s, t \in R_i$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$.

Here we introduce some more notation and let

$$\begin{aligned}
\boldsymbol{D}_i(t) &= (D_{i1}(t), D_{i2}(t))', & \boldsymbol{\lambda}_i(t) &= (\lambda_{i1}(t), \lambda_{i2}(t))', \\
\boldsymbol{D}_i &= (\boldsymbol{D}_i'(1), ..., \boldsymbol{D}_i'(\tau_i))', & \boldsymbol{\lambda}_i &= (\boldsymbol{\lambda}_i'(1), ..., \boldsymbol{\lambda}_i'(\tau_i))',
\end{aligned}$$

$$\Sigma_i = diag(\Sigma_i(1), ..., \Sigma_i(\tau_i)), \qquad \Omega_i = \Big(\Omega_i(s,t)\Big)_{s,t=1,....,\tau_i},$$
$$\text{and} \qquad V_i = \text{var}(\boldsymbol{D}_i) = \Sigma_i + \Omega_i.$$

An estimating function for $\boldsymbol{\theta}$ is given by

$$U_{1,2}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \boldsymbol{a}_i(\boldsymbol{d}_i - \boldsymbol{\lambda}_i),$$

where $\boldsymbol{a}_i = (\boldsymbol{a}_i(1), ..., \boldsymbol{a}_i(\tau_i))$ is a matrix of weight functions. The optimal weight is obtained by finding $F_i(s,t)$ in Proposition 5.1. Let $F_i = \Big(F_i(s,t)\Big)_{s,t=1,....,\tau_i}$; then (5.3.2) becomes

$$F_i\Omega_i + F_i\Sigma_i + \Sigma_i^{-1}\Omega_i = \mathbf{0}.$$

Simplifying, we find $F_i = -\Sigma_i^{-1} + V_i^{-1}$. Therefore, from (5.3.3), the optimal weight function is given by,

$$\boldsymbol{a}_i^* = \frac{\partial \boldsymbol{\lambda}_i}{\partial \boldsymbol{\theta}}\Sigma_i^{-1} + \frac{\partial \boldsymbol{\lambda}_i}{\partial \boldsymbol{\theta}}F_i = \frac{\partial \boldsymbol{\lambda}_i}{\partial \boldsymbol{\theta}}V_i^{-1}.$$

Hence, the optimal LUEF for $\boldsymbol{\theta}$ is

$$U_{1,2}^*(\boldsymbol{\theta}) = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\lambda}_i}{\partial \boldsymbol{\theta}}V_i^{-1}(\boldsymbol{d}_i - \boldsymbol{\lambda}_i), \tag{5.6.4}$$

which is the well-known quasi-score equation (McCullagh and Nelder, 1989).

Given $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1', \boldsymbol{\sigma}_2', \boldsymbol{\sigma}_{12}')'$, a solution to (5.6.4) can be found by the following scoring algorithm (Liang and Zeger, 1986):

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \left(\sum_{i=1}^{m} C_i V_i^{-1} C_i'\right)^{-1} \left(\sum_{i=1}^{m} C_i V_i^{-1} \boldsymbol{r}_i\right), \tag{5.6.5}$$

where $C_i = \partial \boldsymbol{\lambda}_i / \partial \boldsymbol{\theta}$ and $\boldsymbol{r}_i = \boldsymbol{d}_i - \boldsymbol{\lambda}_i$.

The estimates for $\boldsymbol{\sigma}$ are often found by the method of moments in which $b_{ij}(s,t) = 1$ in

(5.5.1) for $j = 1, 2, 12$. However, the method of moments usually leads to slow convergence in the iteration procedure. Crowder (1985) argued that a Gaussian estimation procedure for correlated binary data may provide faster convergence. Crowder's procedure uses a Gaussian pseudo-likelihood for $\boldsymbol{r}_i$:

$$l(\boldsymbol{\theta}, \boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i=1}^{m}\left(\log(|V_i|) + \boldsymbol{r}_i' V_i^{-1}\boldsymbol{r}_i\right).$$

The estimates for $\boldsymbol{\sigma}$ are obtained by solving $\boldsymbol{g}(\boldsymbol{\sigma}) = \partial l/\partial\boldsymbol{\sigma} = 0$. Labeling the elements of $\boldsymbol{\sigma}$ as $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_{q.})'$, the jth element of $\boldsymbol{g}(\boldsymbol{\sigma})$ is equal to

$$g_j(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i=1}^{m}\left(tr\left[V_i^{-1}\frac{\partial V_i}{\partial\sigma_j}\right] - \boldsymbol{r}_i' V_i^{-1}\frac{\partial V_i}{\partial\sigma_j}V_i^{-1}\boldsymbol{r}_i\right).$$

It is noted that the entries in $\boldsymbol{g}(\boldsymbol{\sigma})$ are quadratic estimating functions for $\boldsymbol{\sigma}$. A scoring algorithm for $\boldsymbol{\sigma}$ is given by (Rochon, 1996)

$$\boldsymbol{\sigma}^{(k+1)} = \boldsymbol{\sigma}^{(k)} + H(\boldsymbol{\sigma}^{(k)})^{-1}\boldsymbol{g}(\boldsymbol{\sigma}^{(k)}), \tag{5.6.6}$$

where the (a, b)th element of $H(\boldsymbol{\sigma})$ is

$$H_{ab}(\boldsymbol{\sigma}) = \frac{1}{2}\sum_{i=1}^{m} tr\left[V_i^{-1}\frac{\partial V_i}{\partial\sigma_a}V_i^{-1}\frac{\partial V_i}{\partial\sigma_b}\right],$$

for $a, b = 1, ..., q..$ Hence, the estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ can be found by iterating through (5.6.5) and (5.6.6).

Furthermore, a "working" correlation structure can be used as in the usual GEE approach. For example, the equicorrelation model and the autoregression model are common choices. As seen in section 5.5, a consistent estimator for $\boldsymbol{\theta}$ with a robust variance estimate is still available.

5.7 Simulation Study for MPEF

5.7.1 Design of Simulation Study

In this section, we study the performance of the estimators obtained from the MPEF by examining their biases and coverage probabilities through simulations. We consider m independent mixed bivariate Poisson processes with multiplicative covariate and random effects. Specifically, given random effect $\boldsymbol{z}_i = (z_{i1}, z_{i2})'$, the conditional intensity of the jth Poisson process for the ith subject is given by

$$\lambda_{ij}(t|\boldsymbol{z}_i) = z_{ij}\exp(\beta_{j0} + \beta_{j1}x_i)\alpha_j t^{\alpha_j - 1}, \qquad t \in (0, \tau_i),$$

where $x_i \sim Bin(1, 0.5)$ to mimic a random treatment assignment and $\alpha_j = \exp(\gamma_j)$, $j = 1, 2$, $i = 1, ..., m$.

The bivariate random effects $\boldsymbol{z}_i$ are independent and identically distributed with mean $(1, 1)$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1 & \sigma_{12} \\ \sigma_{12} & \sigma_2 \end{pmatrix},$$

according to one of the following distributions:

1. bivariate log-normal, i.e., $\log(\boldsymbol{Z}_i)$ is bivariate normal;

2. bivariate binary:

$$\Pr(Z_{i1} = 1-\sqrt{\sigma_1}, Z_{i2} = 1-\sqrt{\sigma_2}) = \frac{1+\rho}{4}; \quad \Pr(Z_{i1} = 1-\sqrt{\sigma_1}, Z_{i2} = 1+\sqrt{\sigma_2}) = \frac{1-\rho}{4};$$

$$\Pr(Z_{i1} = 1+\sqrt{\sigma_1}, Z_{i2} = 1-\sqrt{\sigma_2}) = \frac{1-\rho}{4}; \quad \Pr(Z_{i1} = 1+\sqrt{\sigma_1}, Z_{i2} = 1+\sqrt{\sigma_2}) = \frac{1+\rho}{4},$$

where $\rho = \text{corr}(Z_{i1}, Z_{i2}) = \sigma_{12}/\sqrt{\sigma_1\sigma_2}$;

3. mixture of two bivariate log-normals: $\boldsymbol{Z}_i = \boldsymbol{Z}_i^{(1)}$ with probability 1/2 and $\boldsymbol{Z}_i = \boldsymbol{Z}_i^{(2)}$ with probability 1/2, where $\boldsymbol{Z}_i^{(1)}$ and $\boldsymbol{Z}_i^{(2)}$ are independent bivariate log-normal random variables such that

$$\mathrm{E}(\boldsymbol{Z}_i^{(1)}) = \frac{4}{5}\mathbf{1}; \qquad \mathrm{var}(\boldsymbol{Z}_i^{(1)}) = \frac{8}{13}(\Sigma - \frac{1}{25}\mathbf{1}\mathbf{1}'),$$

$$\mathrm{E}(\boldsymbol{Z}_i^{(2)}) = \frac{6}{5}\mathbf{1}; \qquad \mathrm{var}(\boldsymbol{Z}_i^{(2)}) = \frac{18}{13}(\Sigma - \frac{1}{25}\mathbf{1}\mathbf{1}'),$$

where $\mathbf{1} = (1, 1)'$.

The purpose of considering these mixing distributions is to investigate the sensitivity of the estimators, especially for the dispersion parameters σ, to different mixing distributions.

The lengths of follow-up (τ_i) are generated independently of the mixed Poisson processes from an exponential distribution with mean log(0.5) such that the follow-up periods for about 50% of the subjects are less than the target follow-up of 1 (year). The true values of the parameters are taken to be: $\beta_{10} = \log(2)$, $\beta_{11} = 1$, $\gamma_1 = \log(2)$, $\beta_{20} = \log(4)$, $\beta_{21} = -1$, $\gamma_2 = \log(1)$, $\sigma_1 = \sigma_2 = 0.5$ and $\sigma_{12} = -0.25$ and 0.25. The average numbers of events generated from each component of the process under this specification are small.

We generate $m = 100$ mixed bivariate Poisson processes according to the above scheme, and apply the MPEF (5.6.1) to estimate the parameters in the CMFs and (5.6.2) to estimate the dispersion parameters. The baseline rates are chosen either as Weibull, which is the correct model specification, or piecewise constant functions with 5 cut-points. The cut-points for each component are determined by the empirical percentiles of the observed event times in that component. The variance matrix is computed by the empirical formula of the sandwich estimator. A total of 1,000 simulations are performed.

The standardized biases, 95% empirical coverage probabilities, average of the model-based standard errors and the standard deviations of the empirical distributions of the estimates (simulation-based standard errors) are reported in Tables 5.1 to 5.4. The stan-

dardized bias is the average bias of an estimate divided by its simulation standard error and is approximately distributed as standard normal. The 95% coverage probability is the proportion of 95% confidence intervals, which are computed as $\hat{\theta} \pm 1.96 s.e.(\hat{\theta})$, that cover the true parameter, where $s.e.$ is the square root of the robust variance estimate for that parameter. Since σ_1 and σ_2 are non-negative, we also consider inference based on their logarithmic transforms in order to obtain less skewed asymptotic and empirical distributions for the estimates of $\log(\sigma_1)$ and $\log(\sigma_2)$.

5.7.2 Results of Simulation

We first look at the results reported in Tables 5.1 and 5.3 which correspond to the correctly specified conditional intensities. The standardized biases in all parameters are very small regardless of the assumed mixing distributions, particularly for β_{11}, β_{21}, γ_1 and γ_2, indicating that the bias produced by using the MPEF is negligible and the estimates are robust to the mixing distribution. For the estimation of the parameters in the CMFs, the 95% coverage probabilities are slightly smaller than the nominal level 0.95 and the robust standard errors are also only slightly less than the simulation standard errors. The same pattern is observed across the assumed mixing distributions for different σ_{12}, and thus suggests that the estimators perform quite satisfactorily.

On the other hand, although the dispersion parameters are estimated consistently, there are some discrepancies between their coverage probabilities and the nominal level, as well as between the robust standard errors and the simulation standard errors. The discrepancies seem to depend on the mixing distribution in which the processes generated from binary random effects lead to smaller discrepancies. Use of the logarithmic transformations improves the performance slightly in some cases. There is also some loss of efficiency for the estimate of the covariance σ_{12}, although to a lesser extent than the variances σ_1 and σ_2.

	Log-normal				Binary				Log-normal Mixture			
	bias[1]	cover[2]	se-m[3]	se-s[4]	bias	cover	se-m	se-s	bias	cover	se-m	se-s
β_{10}	-0.059	0.938	0.166	0.177	-0.100	0.941	0.171	0.178	-0.111	0.931	0.167	0.173
β_{11}	-0.014	0.933	0.213	0.224	0.056	0.951	0.219	0.221	0.052	0.937	0.214	0.221
γ_1	-0.016	0.937	0.060	0.062	0.012	0.939	0.060	0.062	0.014	0.943	0.059	0.059
β_{20}	-0.104	0.937	0.131	0.137	-0.123	0.931	0.137	0.144	-0.125	0.938	0.134	0.134
β_{21}	0.009	0.936	0.216	0.233	0.036	0.936	0.222	0.230	0.027	0.948	0.219	0.221
γ_2	0.006	0.939	0.066	0.066	0.046	0.927	0.066	0.069	0.027	0.945	0.067	0.067
σ_1	-0.188	0.737	0.164	0.231	-0.088	0.915	0.135	0.143	-0.151	0.748	0.167	0.251
σ_2	-0.217	0.766	0.169	0.241	-0.086	0.926	0.141	0.144	-0.145	0.796	0.177	0.239
σ_{12}	0.116	0.929	0.083	0.088	-0.025	0.914	0.097	0.105	0.106	0.928	0.083	0.087
$\log(\sigma_1)$	-0.424	0.832	0.367	0.502	-0.230	0.951	0.284	0.298	-0.406	0.845	0.364	0.494
$\log(\sigma_2)$	-0.452	0.900	0.406	0.542	-0.229	0.961	0.298	0.304	-0.385	0.905	0.394	0.497

[1]Standardized bias: bias divided by simulation s.e. [2]95% coverage probability.
[3]Robust model-based s.e. [4]Simulation-based s.e.

Table 5.1: Weibull baseline rates, $\sigma_{12} = -0.25$.

	Log-normal				Binary				Log-normal Mixture			
	bias[1]	cover[2]	se-m[3]	se-s[4]	bias	cover	se-m	se-s	bias	cover	se-m	se-s
β_{11}	-0.051	0.932	0.213	0.222	0.014	0.954	0.219	0.219	0.014	0.938	0.214	0.220
β_{21}	0.028	0.937	0.215	0.233	0.054	0.937	0.220	0.229	0.044	0.945	0.217	0.220
σ_1	-0.075	0.756	0.174	0.253	-0.076	0.923	0.138	0.153	-0.045	0.767	0.177	0.277
σ_2	-0.222	0.762	0.169	0.240	-0.094	0.920	0.141	0.144	-0.151	0.792	0.177	0.238
σ_{12}	0.073	0.930	0.083	0.090	-0.061	0.917	0.098	0.106	0.059	0.935	0.084	0.088
$\log(\sigma_1)$	-0.325	0.823	0.367	0.515	-0.071	0.938	0.276	0.302	-0.305	0.845	0.364	0.507
$\log(\sigma_2)$	-0.457	0.898	0.406	0.542	-0.237	0.963	0.297	0.305	-0.390	0.901	0.395	0.499

[1]Standardized bias: bias divided by simulation s.e. [2]95% coverage probability.
[3]Robust model-based s.e. [4]Simulation-based s.e.

Table 5.2: Piecewise constant baseline rates, $\sigma_{12} = -0.25$.

	Log-normal				Binary				Log-normal Mixture			
	bias[1]	cover[2]	se-m[3]	se-s[4]	bias	cover	se-m	se-s	bias	cover	se-m	se-s
β_{10}	-0.081	0.939	0.167	0.176	-0.103	0.941	0.170	0.177	-0.037	0.942	0.167	0.170
β_{11}	0.005	0.944	0.213	0.214	0.055	0.940	0.219	0.227	-0.018	0.929	0.213	0.219
γ_1	0.031	0.932	0.059	0.061	0.000	0.942	0.060	0.061	0.009	0.938	0.059	0.060
β_{20}	-0.091	0.944	0.132	0.130	-0.069	0.927	0.137	0.144	-0.066	0.928	0.133	0.139
β_{21}	-0.033	0.944	0.218	0.218	0.001	0.945	0.222	0.227	-0.008	0.947	0.218	0.226
γ_2	0.057	0.945	0.066	0.068	-0.014	0.926	0.067	0.070	-0.029	0.947	0.067	0.064
σ_1	-0.216	0.749	0.163	0.228	-0.082	0.927	0.136	0.138	-0.183	0.763	0.167	0.229
σ_2	-0.151	0.779	0.175	0.246	-0.093	0.920	0.141	0.145	-0.135	0.780	0.175	0.249
σ_{12}	-0.073	0.893	0.122	0.147	-0.022	0.931	0.107	0.113	-0.016	0.882	0.120	0.146
$\log(\sigma_1)$	-0.457	0.848	0.364	0.473	-0.219	0.961	0.284	0.284	-0.415	0.856	0.370	0.490
$\log(\sigma_2)$	-0.390	0.889	0.402	0.544	-0.237	0.958	0.297	0.300	-0.380	0.888	0.400	0.525

[1]Standard bias: bias divided by simulation s.e. [2]95% coverage probability.
[3]Robust model-based s.e. [4]Simulation-based s.e.

Table 5.3: Weibull baseline rates, $\sigma_{12} = 0.25$.

	Log-normal				Binary				Log-normal Mixture			
	bias[1]	cover[2]	se-m[3]	se-s[4]	bias	cover	se-m	se-s	bias	cover	se-m	se-s
β_{11}	-0.033	0.943	0.214	0.212	0.018	0.947	0.219	0.226	-0.057	0.934	0.214	0.217
β_{21}	-0.047	0.944	0.219	0.218	-0.015	0.946	0.223	0.227	-0.022	0.946	0.219	0.226
σ_1	-0.096	0.764	0.173	0.251	0.095	0.937	0.139	0.147	-0.067	0.787	0.177	0.251
σ_2	-0.135	0.782	0.177	0.249	-0.065	0.920	0.142	0.146	-0.120	0.784	0.177	0.251
σ_{12}	-0.028	0.897	0.125	0.152	0.035	0.933	0.110	0.116	0.028	0.887	0.123	0.151
$\log(\sigma_1)$	-0.348	0.857	0.364	0.486	-0.046	0.945	0.276	0.289	-0.311	0.857	0.371	0.506
$\log(\sigma_2)$	-0.376	0.889	0.403	0.547	-0.211	0.960	0.296	0.301	-0.366	0.886	0.397	0.524

[1]Standard bias: bias divided by simulation s.e. [2]95% coverage probability.
[3]Robust model-based s.e. [4]Simulation-based s.e.

Table 5.4: Piecewise constant baseline rates, $\sigma_{12} = 0.25$.

Loss of efficiency for the dispersion parameters is the result of relaxing the model assumptions. The merits of this method of estimation are that the CMF parameters can still be estimated quite efficiently and the dispersion parameters can be estimated consistently. Since within this framework, focus is on the CMFs, inference on the dispersion parameters is often a secondary concern and hence we do not view this deficiency as problematic.

The simulation results for the models with piecewise constant baseline rates agree closely with those of the correct model (Tables 5.2 and 5.4). This indicates that the piecewise constant baseline rates are an attractive robust alternative for the specification of the baseline rates, demonstrating the practical value of a weakly parametric specification for the baseline rate function.

5.8 Bronchial Asthma Study

We consider the first stage of the asthma study described in chapter 1 to illustrate the use of the MPEF. Two treatments for the control of asthma were given to 64 subjects: placebo and fenoterol. In the first stage of this study, subjects were randomized to either of the two treatments for a period of 167 days. The purpose is to estimate the mean rates of two daytime symptoms: wheezing and coughing in relation to the treatment and other covariates.

Let $N_{i1}(t)$ and $N_{i2}(t)$ be the cumulative numbers of days with wheezing and coughing respectively for subject i in $(0,t]$. We model the mean rate function multiplicatively with respect to covariate effects:

$$\mathrm{E}[dN_{ij}(t)] = \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}_j)\lambda_{0j}(t;\boldsymbol{\gamma}_j)dt, \qquad j = 1,2, \quad i = 1,...,m,$$

where $\boldsymbol{x}_{ij}$ is a vector of time-independent covariates for process j of subject i, $\boldsymbol{\beta}_j$ is the

corresponding vector of regression coefficients and $\lambda_{0j}(t)$ is the baseline rate specified by a vector of parameters γ_j. Although inference on the regression coefficients is independent of the form of the baseline rates due to common follow-up time (section 5.6.1), estimating the trends is also of interest. We therefore model the baseline rates as piecewise constant functions with 5 cut-points determined by the empirical percentiles of the observed event times in each component.

We examine the assumption of multiplicative covariate effects by constructing the Nelson-Aalen estimators for the cumulative mean functions of the numbers of days with wheezing and coughing stratified by gender and treatment which are believed to be two major effects. The Nelson-Aalen estimator at time t is given by

$$\Lambda_j^{(h)}(t) = \int_0^t \frac{dN_{\cdot j}^{(h)}(s)}{Y_{\cdot j}^{(h)}(s)}, \qquad t > 0,$$

where $N_{\cdot j}^{(h)}(t) = \sum_{i \in G_h} N_{ij}(t)$, $Y_{\cdot j}^{(h)}(t) = \sum_{i \in G_h} Y_{ij}(t)$, and G_h is the set of subjects who belong to group h, $h = 1, 2, 3, 4$. Here group 1 corresponds to the placebo-female group, group 2 the placebo-male group, group 3 the fenoterol-female group and group 4 the fenoterol-male group. In this example, $Y_{\cdot j}^{(h)}(t)$ is equal to the number of subjects in group h for $s \in (0, 167]$. These estimates are displayed in Figures 1 and 2. There is no clear evidence that the multiplicative assumption is violated.

Furthermore, the sample correlation between the numbers of wheezes and coughs observed in $(0, 167]$ is equal to 0.47 which clearly indicates that the wheezing and coughing processes are correlated. Certainly, the correlation structure is unknown and the MPEF is a convenient choice of a working model.

We applied the MPEF with the dispersion parameters estimated from (5.6.2). Robust variance estimates were computed using the empirical formula. Parameter estimates are presented in Table 5.5. We label this model as the full model.

	Wheeze			Cough		
Covariate*	Estimate	s.e.	p-value	Estimate	s.e.	p-value
Age	-0.023	0.049	0.631	-0.042	0.052	0.425
Age^2	<0.001	0.001	0.871	0.001	0.001	0.404
Gender	-0.738	0.290	0.011	-0.846	0.323	0.009
Drug	-0.225	0.280	0.422	0.400	0.300	0.182
Gender×Drug	0.403	0.402	0.316	0.315	0.444	0.478
Smoke	0.157	0.213	0.460	-0.052	0.243	0.832
Daytime Symptoms:						
< weekly	0.218	0.578	0.707	0.586	0.666	0.379
> weekly	0.338	0.552	0.541	0.944	0.666	0.156
daily	1.243	0.584	0.033	1.247	0.756	0.099
> 1 daily	0.961	0.608	0.114	1.005	0.799	0.209
Nocturnal:						
< weekly	0.002	0.271	0.994	-0.831	0.278	0.003
> weekly	0.004	0.244	0.988	0.062	0.276	0.823
most nights	-0.070	0.391	0.859	0.409	0.462	0.376
Tightness on waking:						
< weekly	0.504	0.290	0.083	0.025	0.382	0.948
> weekly	0.795	0.284	0.005	0.169	0.396	0.669
most mornings	0.501	0.306	0.101	-0.215	0.440	0.625

*Age in year at entry; Gender: F=0, M=1; Drug: Placebo=0, Active=1; Smoke: Non-smoker=0, Smoker=1; Daytime Symptoms, Nocturnal and Tightness on waking are symptoms over the past 4 weeks at entry with no symptoms at baseline, and the categories are frequencies of symptoms.

	Dispersion	
Parameter	Estimate	s.e.
σ_1	0.302	0.073
σ_2	0.608	0.178
σ_{12}	0.252	0.070

Table 5.5: Estimation of the full model for the asthma study.

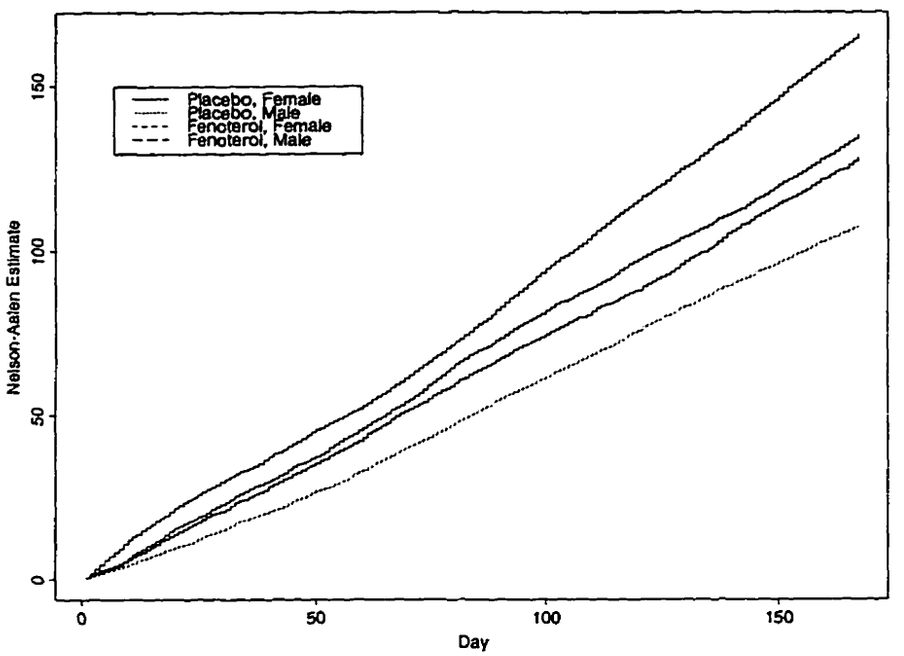

Figure 5.1: Nelson-Aalen estimates for the cumulative mean functions of the wheeze processes with respect to the placebo-female group, the fenoterol-female group, the fenoterol-male group and the placebo-male group, in descending order of the curves.

The p-values for testing covariate effects based on the Wald test described in section 5.4 indicate that some covariates have no significant effects on the CMFs. We therefore consider a reduced model with results given in Table 5.6. The score test proposed in section 5.4 was used to test if the full and reduced models are significantly different. Using the estimates of σ from the full model, we found the score statistic to be 20.606 with degrees of freedom 26 (p-value = 0.762). Therefore, there is no evidence that the reduced model is significantly different from the full model.

The effects of fenoterol and placebo on the number of days with wheezing are not significantly different. In contrast, use of fenoterol led to higher number of days with coughing than did placebo. Males appear to have a smaller number of days with symptoms than females on average. The severity of daytime symptoms had an adverse effect

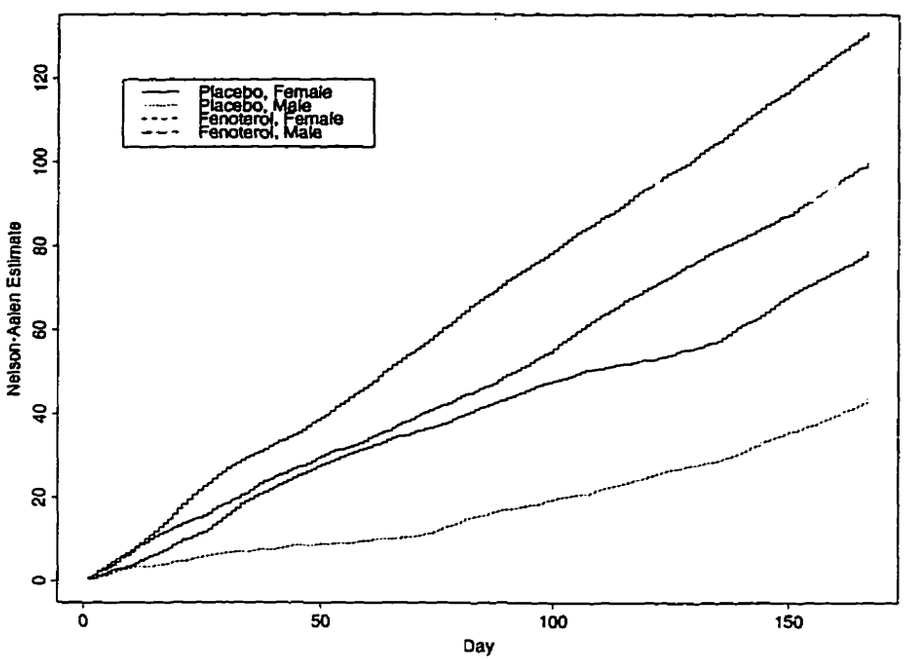

Figure 5.2: Nelson-Aalen estimates for the cumulative mean functions of the cough processes with respect to the fenoterol-female group, the placebo-female group, the fenoterol-male group and the placebo-male group, in descending order of the curves.

on wheezing but insignificant effect on coughing. However, it is puzzling that for the nocturnal activities variable, the number of days with coughing is smaller for subjects whose frequency of symptoms at night is less than weekly than for subjects without symptoms at night. Possible explanation may be due to the fact that only symptomatic activities at randomization were considered in the analysis, but subsequent symptomatic activities are ignored. Such covariates are time-dependent and certainly complicate the estimation procedure considered here. The analysis of the bronchial asthma study presented here is a preliminary investigation and an illustration of the method proposed in this chapter. We do not intend to give a thorough analysis of this study in this thesis.

Finally, the estimates for σ indicate substantial overdispersion. The estimated baseline rates are also shown in Figures 5.3 and 5.4. The wheezing process has a slowly increasing

	Wheeze			Cough		
Covariate	Estimate	s.e.	p-value	Estimate	s.e.	p-value
Gender	—	—	—	-0.441	0.204	0.030
Drug	—	—	—	0.405	0.210	0.053
Daytime Symptoms:						
daily	0.823	0.167	< 0.001	—	—	—
> 1 daily	0.707	0.200	< 0.001	—	—	—
Nocturnal:						
$<$ weekly	—	—	—	-0.809	0.233	< 0.001

Dispersion		
Parameter	Estimate	s.e.
σ_1	0.434	0.102
σ_2	0.816	0.185
σ_{12}	0.330	0.087

Table 5.6: Estimation of the reduced model for the asthma study.

trend while the coughing process has a slowly changing bath-shape trend, although there seems to be no significant change in the trends as suggested by the 95% confidence intervals.

5.9 Concluding Remarks and Discussion

In this chapter, we proposed some marginal models for the CMFs of bivariate point processes using estimating functions. A criterion for obtaining the optimal weight function for the CMF is provided. A procedure for joint estimation of the parameters in the CMF and the covariance functions is also suggested. In situations for which the covariance structure is unknown, estimating functions arising from mixed bivariate Poisson processes, the MPEF, may be useful. We examined the properties of the estimators for the MPEF. The estimators for the CMF parameters generally perform satisfactorily. The robust variance estimates are quite efficient compared with the sample variance estimates. The dispersion parameters can be estimated consistently, although some loss of efficiency is unavoidable.

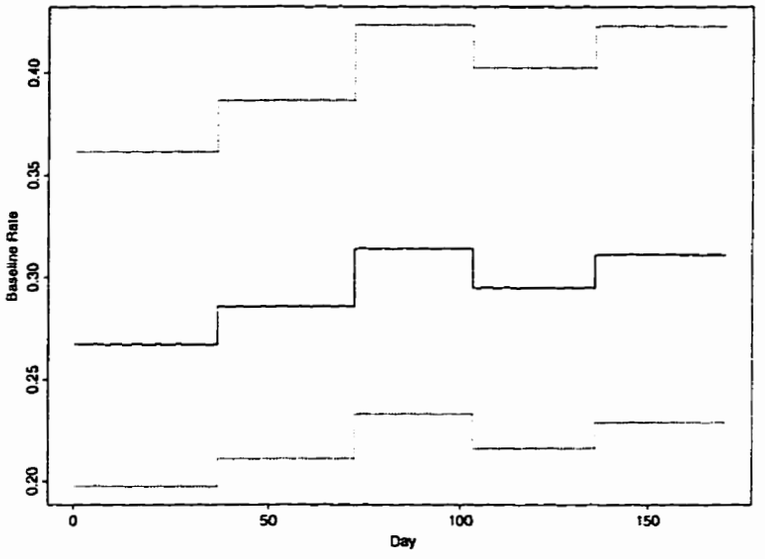

Figure 5.3: Estimate of the baseline mean rate function for the wheeze process (solid line) with 95% pointwise confidence intervals (dashed line).

We discuss some aspects for further research of the marginal approach here.

5.9.1 Semi-parametric Models

It is interesting to extend the above method to semi-parametric models in which the mean function is expressed as

$$\mathrm{E}(dN_{ij}(t)) = \lambda_{ij}(t)d\nu(t) = g_{ij}(t;\boldsymbol{\beta}_j)\lambda_{j0}(t)d\nu(t),$$

for $j = 1, 2$ and $i = 1, ..., m$, where $g_i(t)$ is a known positive function of covariates $\boldsymbol{x}_{ij}(t)$ parameterized by $\boldsymbol{\theta}_j$, for example the widely used log-link function $\exp(\boldsymbol{x}'_{ij}(t)\boldsymbol{\beta}_j)$, and $\lambda_{j0}(t)$ is an unknown positive cadlag function independent of $\boldsymbol{x}_{ij}(t)$ and is common to all subjects.

Nadeau and Lawless (1996) suggested estimating functions for $\boldsymbol{\beta}$ in univariate processes

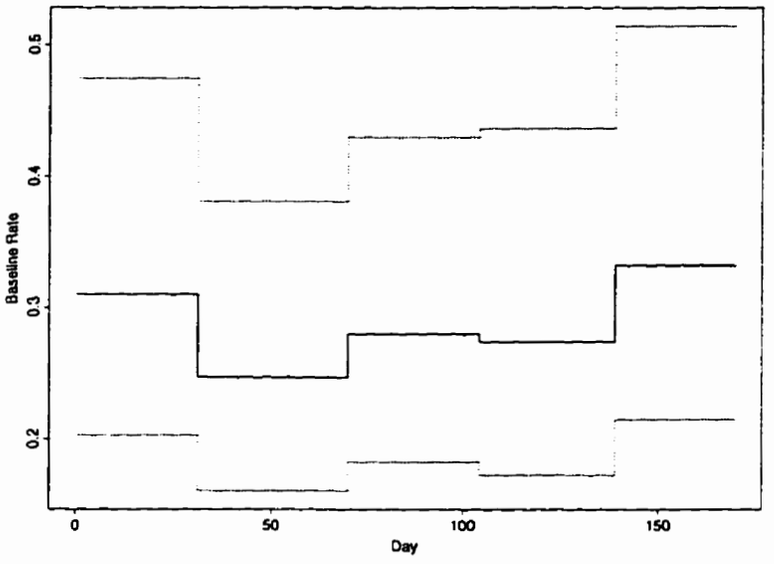

Figure 5.4: Estimate of the baseline mean rate function for the cough process (solid line) with 95% pointwise confidence intervals (dashed line).

where the well-known Cox partial likelihood score is included as a special case. Optimal linear estimating functions exist for processes with uncorrelated jumps.

For more general correlated-jump processes, piecewise constant specifications for the mean and covariance functions should give a reasonable approximation to the semiparametric model, as empirical evidence has been given for mixed bivariate Poisson processes in section 5.7.

5.9.2 Extension to Multivariate Point Processes

Since the proof of Proposition 5.1 does not rely on the dimension of the point processes as long as it is finite, Proposition 5.1 still holds for higher dimensional point processes.

Consider a K-dimensional point process. We have to define K CMFs $(\lambda_{ij}(t))$, K variance functions $(v_{ij}(t))$, K autocovariance functions $(c_{ij}(s,t))$ of jumps within components and $2\binom{K}{2}$ cross-covariance functions of jumps between components $(v_{i,jk}(t), c_{i,jk}(s,t))$. This

requires totally $K(K+2)$ specifications of the mean and covariance functions. Frequently some covariance functions have unknown forms. In this case, a convenient working model can be obtained from mixed multivariate Poisson processes as in section 5.6.1.

5.9.3 Two-state Processes

The marginal approach for bivariate processes does not apply to two-state processes. This is because the numbers of $1 \to 2$ and $2 \to 1$ transitions over the same period of time differ by at most 1. This creates difficulties in constructing the marginal expected numbers of transitions which are functionally related in an unknown manner if the probabilistic structure is not fully specified. Even if the transition intensities are specified, the expected numbers of transitions are not always available, for example in the alternating renewal process. Nevertheless, the main interest for such processes is usually focussed on the analysis of transition probabilities in which case an intensity model seems to be more appealing.

Chapter 6

Further Research

6.1 Overview

In previous chapters of this thesis, we have presented extensions of established methodology. Here we describe two major additional areas for further research. The first topic pertains to the estimation of mixed multi-state processes and the second topic involves methods for model diagnostics for random effect and marginal point process models.

The difficulties in the estimation of random effect models arising from multi-state processes lie in the specification of a genuine multivariate mixing distribution and the accuracy of an approximate marginal likelihood computed from an estimation procedure. We briefly discuss some potential methods of estimation using a mixed illness-death model as an illustration in section 6.2.

Model diagnostics are important in assessing the performance of models. They are usually carried out by visual inspection of residual plots and more formally tests for goodness of fit. In fixed-effect failure time models, there has been quite a lot of work directed at residual analyses (Kay, 1977; Barlow and Prentice, 1988; Therneau et al., 1990; Lin et al., 1993), goodness-of-fit tests (Lin and Wei, 1989; 1991; Crouchley and Pickles, 1993), and

tests of proportionality in Cox regression models (Arjas, 1988; Grambsch and Therneau, 1994). However, there is a lack of literature on diagnostics for random effect models, due in part to the complex distributional structure of residuals arising from the mixed models. We will suggest some ways of constructing residuals, a deletion diagnostic for detecting influential observations, and a goodness-of-fit test using the IM test for mixed point processes in sections 6.3 to 6.5. The investigation here is quite preliminary. As a starting point, we focus on the univariate point process. Finally, some concluding remarks are given in section 6.6.

6.2 Mixed Multi-state Processes

Accommodating heterogeneity in multi-state processes requires the specification of a probably high-dimensional mixing distribution. For instance, the three-state illness-death process illustrated in Figure 6.1 is frequently modeled in applications. Subjects make transitions between the healthy state and the diseased state until a transition to death, an absorbing state. The arrows indicate possible transitions between states. The CHEST study described in section 1.4.2 can be regarded as illness-death processes with heterogeneity if the observed deaths were modeled. Another application is the Studies of Left Ventricular Dysfunction (SOLVD) (The SOLVD Investigators, 1991) which aims to investigate the rates of mortality and hospitalization due to congestive heart failure in asymptomatic and symptomatic patients with reduced left ventricular ejection fractions.

Let $\delta(t)$ be the state occupied by a subject at time $t \geq 0$. Conditional on the history of the process and the covariates up to time t, and the random effects $\boldsymbol{u} = (u_{12}, u_{21}, u_{13}, u_{23})'$, the $j \to k$ transition intensity assuming the multiplicative model, is given by

$$\lambda_{jk}(t|\mathcal{H}(t), \boldsymbol{u}) = \exp(\boldsymbol{x}'_{jk}(t)\boldsymbol{\beta}_{jk} + u_{jk})\lambda_{jk0}(t|\mathcal{H}(t)),$$

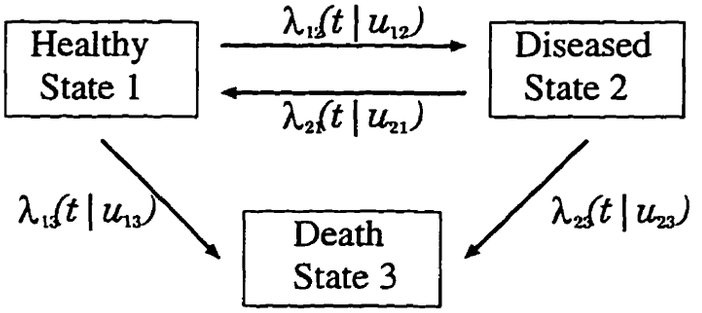

Figure 6.1: An illness-death model with multivariate random effects.

where $\boldsymbol{x}_{jk}(t)$ is the vector of covariates for the $j \to k$ transition,

$$\mathcal{H}(t) = \{\delta(s), \boldsymbol{x}_{12}(s), \boldsymbol{x}_{21}(s), \boldsymbol{x}_{13}(s), \boldsymbol{x}_{23}(s) | 0 \leq s < t\},$$

and $\lambda_{jk0}(t|\mathcal{H}(t))$ is the baseline intensity. Aalen (1987) suggested a construction of a multivariate mixing distribution based on a transformation of multivariate normal random variables for a relatively simple time-homogeneous illness-death process.

For more general time-inhomogeneous illness-death processes, the multivariate normal mixing distribution seems to be the most feasible genuine multivariate mixing distribution. The marginal likelihood may be evaluated in principle by numerical integration. Evans and Swartz (1995) recommended that Gaussian quadrature rules be used for integration problems of dimension less than about 6. The Gauss-Hermite rule is thus appropriate for the mixed illness-death model which involves integrations of dimension 4.

However, the numerical integration method may not be suitable for mixed multi-state processes involving high-dimensional integrations. Evans and Swartz (1995) suggested using Monte Carlo integrations for high-dimensional integration problems. Alternatively, if the number of transitions per subject is large, the penalized likelihood may be tried. Nevertheless, there is still no satisfactory estimation method for high-dimensional integration problems.

Non-parametric mixing distributions may also be considered. However, we have to deal with a large number of parameters, which is likely to cause numerical problems even for moderately large samples.

As advocated in chapter 4, specifying proper intensity functions is more important in the sense that the degree of observed heterogeneity may be minimized and thus with an adequate model, even a fixed-effect model may be satisfactory.

6.3 Residual Analysis

6.3.1 Fixed-effect Models

In fixed-effect models, there are three major model departures: specification of the functional forms of covariates, specification of the baseline intensity and the assumption of multiplicity. Useful residuals should have two basic properties of being sensitive to model departures and having known distributions under the assumed model.

We consider the univariate intensity model $\lambda(t) = \exp(\boldsymbol{x}'\boldsymbol{\beta})\lambda_0(t)$. Residuals are usually constructed via the cumulative intensity between two consecutive events:

$$r_j = \exp(\boldsymbol{x}'\boldsymbol{\beta}) \int_{t_{j-1}}^{t_j} \lambda_0(t)dt,$$

where t_j is the occurrence time for the jth event such that $t_0 = 0$. If the model is adequate, the r's evaluated under the true parameter values are iid exponentially distribution with mean 1 (Lawless and Thiagarajah, 1996). A probability plot for the unit exponential distribution and an index plot may be constructed to examine the goodness of fit. In practice, estimates are used in place of the true values. The consequences are that the residuals are not independent and may be distributed differently from $Exp(1)$.

The martingale residual is also popular in the counting process approach. It is defined

as

$$M(t) = N(t) - \exp(\boldsymbol{x}'\boldsymbol{\beta}) \int_0^t Y(s)\lambda_0(s)ds,$$

for $t > 0$, where $Y(t)$ is the indicator that the process is under observation at time t. This is the difference between the number of observed events and the number of expected events under the model in $[0, t]$. It should be noted that the intensity generally depends on the history of the process and thus the cumulative intensity itself is a model-based estimator for the expected cumulative counts, whereas the cumulative intensity of a Poisson process is equal to the expected cumulative counts. Strictly speaking, $M(t)$ is the difference between the observed and the estimated cumulative counts.

The residual is usually taken as $M(\infty)$. An index plot is often constructed to check the model adequacy. Also, plots of the martingale residuals against covariates provides useful clues on the appropriateness of the functional forms of covariates.

Nevertheless, all residuals have to be evaluated at the estimates of the unknown parameters. Baltazar-Aban and Pena (1995) showed that even for the ordinary Cox regression model, the properties of these estimated residuals are not well understood.

The residuals may be adjusted for the bias induced by the substitution of parameter estimates. Using a similar argument in the adjusted score test proposed in chapter 3 and treating the residuals as functions of the parameters, we may construct bias-adjusted residuals. Further investigations are necessary to study their properties.

6.3.2 Random Effect Models

In addition to the three main model departures mentioned in section 6.3.1, the specification of a mixing distribution is another possible departure. For the widely-used gamma frailty models in survival analysis, Shih and Louis (1995) proposed a graphical method for the gamma mixing assumption. They pointed out that covariate effects are usually multiplicative in most situations and the unknown baseline hazard may well be specified by either a

piecewise exponential model or a non-parametric model as in Clayton and Cuzick (1985), Klein (1992) and Nielsen et al. (1992). The remaining major departure is then the gamma frailty assumption. By treating the posterior mean of the gamma frailty, given the data at time t, as a stochastic process, Shih and Louis considered a time-plot of the centered posterior means with confidence bands, which resembles a usual residual plot. For frailties other than gamma, or for multivariate frailties, similar approaches may be computationally intensive as numerical integrations are necessary. Nevertheless, simulation studies (section 4.2) demonstrate that log-normal random effect models using the Gauss-Hermite integration and non-parametric random effect models using the EM algorithm provide similar and valid inference for parameters in the intensity function, even for misspecified mixing distribution. Therefore, checking the assumption of a mixing distribution may be of only secondary relevance for practical problems.

In general, the adequacy of the model may be examined by constructing residuals similar to those in fixed-effect multiplicative intensity models. Here we propose some methods for the construction of such residuals.

Conditional Residuals

Suppose the random effects and parameters were known. We may consider the conditional residual, given the random effect, defined as

$$r_j = v \exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t_{j-1}, t_j),$$

where $\Lambda_0(a, b) = \int_a^b \lambda_0(t)dt$. The r_j's are independently and identically distributed as $Exp(1)$. In the case of censoring, the conditional expectation given that the event time is greater than its censoring time, may be used (Lawless, 1982). It is expressed as

$$r_j^c = v \exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t_{j-1}, t_j) + 1.$$

Residual plots can then be constructed. For example, a probability plot for an exponential distribution with unit mean can be used to assess the distribution of the residuals. Plots of the log-transformed residuals against covariates provide useful clues about the functional form of covariates. In applications, the residuals may be estimated by replacing the parameters by their estimates and the random effects by their posterior means, $\mathrm{E}[V|\mathcal{H}(\tau)]$, where $\mathcal{H}(\tau)$ is the history of the entire observation period.

Marginal Residuals

For small numbers of recurrences, the posterior mean of V may not be a good estimate of V. In such situations, it might be more appropriate to consider the unconditional residual. Since the marginal intensity conditional on the history is given by

$$\lambda(t|\mathcal{H}(t)) = \mathrm{E}[V|\mathcal{H}(t)]\exp(\boldsymbol{x}'\boldsymbol{\beta})\lambda_0(t|\mathcal{H}(t)),$$

the unconditional residual is thus defined as

$$r_j = \int_{t_{j-1}}^{t_j} \mathrm{E}[V|\mathcal{H}(t)]\exp(\boldsymbol{x}'\boldsymbol{\beta})\lambda_0(t)dt,$$

where $\mathcal{H}(t) = \{N(s)|0 \leq s < t\}$. For simplicity, we drop the dependence of $\mathcal{H}(t)$ on $\lambda(t|\mathcal{H}(t))$ and $\lambda_0(t|\mathcal{H}(t))$. Let $\mathcal{L}_v(\cdot)$ be the Laplace transform of V. Using the conditional likelihood for $\mathcal{H}(t)$ given v,

$$v^{N(t-)}\exp(N(t-)\boldsymbol{x}'\boldsymbol{\beta})\prod_{j=1}^{N(t-)}\lambda_0(t_j)\exp(-v\exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t)),$$

it is straightforward to show that

$$\mathrm{E}[V|\mathcal{H}(t)] = \frac{-\mathcal{L}_v^{(N(t-)+1)}(\exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t))}{\mathcal{L}_v^{(N(t-))}(\exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t))}.$$

Suppose V is gamma distributed with mean 1 and variance σ; then

$$\mathrm{E}[V|\mathcal{H}(t)] = \frac{1+\sigma N(t-)}{1+\sigma\exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t)},$$

and thus the residual becomes

$$\begin{aligned} r_j &= \int_{t_{j-1}}^{t_j} \frac{1+\sigma N(t-)}{1+\sigma\exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t)} \exp(\boldsymbol{x}'\boldsymbol{\beta})\lambda_0(t)dt \\ &= (\frac{1}{\sigma} + N(t_j-)) \log\left(\frac{1+\sigma\exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t_j;\mathcal{H}(t_j))}{1+\sigma\exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t_{j-1};\mathcal{H}(t_j))}\right), \end{aligned}$$

because $N(t)$ is constant in the interval $[t_{j-1}, t_j)$. Recall that the baseline intensity is in general a function of $\mathcal{H}(t)$. Hence, residual plots for these r's can be constructed. Nevertheless, other mixing distributions may not result in such a nice expression. Further studies are necessary.

Martingale Residuals

Alternatively, a martingale residual is given by

$$M(t) = N(t) - \int_0^t \mathrm{E}[V|\mathcal{H}(s)]\exp(\boldsymbol{x}'\boldsymbol{\beta})Y(s)\lambda_0(s)ds.$$

A gamma mixing distribution would lead to a simple expression:

$$\begin{aligned} M(\infty) = N(\tau) \; - \; \Big[&(\frac{1}{\sigma} + N(\tau))\log(1+\sigma\exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(\tau;\mathcal{H}(\tau))) \\ &- \sum_{j=1}^{N(\tau)} \log[1+\sigma\exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(t_j)]\Big]. \end{aligned}$$

Now the martingale residual and the covariates are related in a complex way. Although it is difficult to reveal the functional forms by plotting residuals against covariates, an index plot is still able to provide a valid overall check.

Pearson Residuals

In GLM, the Pearson residual is widely used (McCullagh and Nelder, 1989). This approach may be useful for mixed point processes provided we can find $\mathrm{E}(N(\tau))$ and $\mathrm{var}(N(\tau))$. Consider a mixed Poisson process. The marginal mean and variance are equal to

$$\begin{aligned} \mathrm{E}(N(\tau)) &= \exp(\boldsymbol{x}'\boldsymbol{\beta})\Lambda_0(\tau) \\ \mathrm{var}(N(\tau)) &= \mathrm{E}(N(\tau)) + \sigma(\mathrm{E}(N(\tau)))^2. \end{aligned}$$

The Pearson residual is defined as

$$r = \frac{N(\tau) - \mathrm{E}(N(\tau))}{\sqrt{\mathrm{var}(N(\tau))}}. \tag{6.3.1}$$

However, other point processes such as the renewal process may not have closed-form expressions for the marginal mean and variance of $N(\tau)$.

On the whole, there are serious limitations for the residuals defined in this section. Difficulties are mainly due to the random effect, which often causes computational difficulties. Despite this, studying the properties of residuals under certain restricted assumptions, such as gamma mixing distributions, is useful in its own right and may provide insight for more general situations.

6.3.3 Marginal Models

Pearson residuals can be constructed for the marginal model discussed in chapter 5. Recall that for a univariate point process, $\mathrm{E}(dN(t)) = \lambda(t)d\nu(t)$, $\mathrm{var}(dN(t)) = v(t)d\nu(t)$ and $\mathrm{cov}(dN(s), dN(t)) = c(s,t)d\nu(s)d\nu(t)$ for $s \neq t$. We find

$$\mathrm{E}(N(\tau)) = \int_0^\tau \lambda(t)d\nu(t)$$

$$\text{var}(N(\tau)) = \int_0^\tau v(t)d\nu(t) + \int_0^\tau \int_0^\tau c(s,t)d\nu(s)d\nu(t).$$

Hence, Pearson residuals with robust variance can be defined as in (6.3.1).

6.4 Detecting Influential Observations

It is possible that the observed heterogeneity may be due to a few influential subjects. One should first re-examine the observations from these subjects. If no human errors are found, removal of these subjects may lead to a simpler model that provides a better fit to the data. Analogous to the ordinary linear regression model (Cook and Weisberg, 1982), we suggest a deletion diagnostic for the random effect model.

Recall that the parameters ϕ come from three components: the regression, the baseline intensity function and the mixing distribution. Let ϕ_k be the kth element of ϕ. Let $\hat{\phi}_k$ and $\hat{\phi}_{k(i)}$ be the maximum likelihood estimates of ϕ using all subjects with and without the ith subject in the sample respectively. The ith subject is considered to be influential in the estimation of ϕ_k if

$$\frac{\hat{\phi}_{k(i)} - \hat{\phi}_k}{\sqrt{se(\hat{\phi}_{k(i)})}} \tag{6.4.1}$$

is large. A summary influence statistic similar to the Cook's distance may be defined as

$$(\hat{\phi}_{(i)} - \hat{\phi})'\hat{V}_{(i)}^{-1}(\hat{\phi}_{(i)} - \hat{\phi}), \tag{6.4.2}$$

where $\hat{V}_{(i)}$ is the estimated asymptotic covariance matrix of $\hat{\phi}_{(i)}$. Again, the ith subject is considered influential if this statistic is large.

Further study is certainly warranted to explore the value and role of deletion diagnostics in the context of mixed point processes.

6.5 Goodness-of-Fit Tests

In addition to the unclear distributional behavior of the estimated residuals, visual inspection of residual plots may be too subjective. It is therefore desirable to develop global goodness-of-fit tests. In particular, the IM test has been widely used for this purpose.

Using the IM test, Lin and Wei (1991) constructed goodness-of-fit tests for Cox regression models. Crouchley and Pickles (1993) illustrated the use of the IM test in univariate and multivariate parametric proportional hazards models. Using a real example, Crouchley and Pickles demonstrated that the conventional residual plot failed to indicate a departure due to the omission of important covariates, while the IM test of homogeneity did detect such a departure.

The IM test may be considered as an omnibus goodness-of-fit test for the random effect model. The model consists of three components, the covariate effect, the baseline intensity, and the random effect, parameterized by β, γ and σ respectively. Let $\phi = (\beta', \gamma', \sigma')'$. An IM test can be constructed from the distinct elements of the information difference:

$$\frac{\partial^2 \ell(\phi)}{\partial \phi \partial \phi'} + \frac{\partial \ell(\phi)}{\partial \phi} \frac{\partial \ell(\phi)}{\partial \phi'}.$$

The variance of the statistic can be found using the formulae in section 3.3.1. This is a χ^2 test with degree of freedom $p^*(1+p^*)/2$, where $p^* = dim(\phi)$. Large value of the test statistic indicates model inadequacy. The total number of parameters is often quite large, and in such a case, one may use the diagonal elements of the information difference in the test.

Specific types of model departure may be investigated by considering subsets of the elements of the information difference. Since the effects of unmeasured covariates have mostly been taken into account by the random effect, it seems appropriate to consider the following three tests:

1. Testing for constant regression coefficients, except for the intercept. Rejection of the test may indicate non-multiplicative covariates or incorrect functional forms of the covariate effect.

2. Testing for constant parameters of the baseline intensity. Non-constancy may indicate functional forms other than the assumed form, or perhaps that more sub-divisions of the time axis are required for a piecewise exponential model.

3. Testing for constant parameters of the mixing distribution. This may indicate that the mixing distribution is misspecified.

Although the test for a particular condition may be sensitive to the other conditions, these tests nevertheless may provide a crude indication to the type of model departure. Since the mixing distribution may have little influence on the estimators for covariate effects based on previous studies, and a piecewise constant baseline intensity should approximate well the true baseline intensity, we recommend that one should focus on tests related to the regression coefficients. Further investigation of these goodness of fit tests are required.

6.6 Concluding Remarks

In this thesis, we investigated three important aspects in the analysis of event history data with an emphasis on developing an appropriate modeling strategy. This strategy should consist of cycling through stages of model specification, model estimation and model diagnostics. In the first step of the strategy, we consider two types of models for analyzing two different properties of the processes: the intensity functions and the mean rate functions. The choice between these models depends mainly on the problem at hand and the assumptions one is willing to make. For example, an intensity model with random effects is used for the CHEST study because the rates of transitions to exacerbation and symptom-free states are the main concern. The transitional intensities are specified quite generally by

incorporating different time scales, and the piecewise constant functions are employed to achieve robustness to misspecification. In contrast, a marginal model is used for the analysis of the bronchial asthma study because the mean number of days with symptoms is the study objective.

In the second step, we studied the performance of two popular methods of estimation for random effect models, and developed an estimating function approach based on mixed Poisson processes for marginal models. Gauss-Hermite integration and the EM algorithm perform quite satisfactorily for parametric and non-parametric mixing distributions respectively for univariate and bivariate processes. The mixed Poisson estimating function serves effectively as a working structure for robust inference based on the marginal model if the covariance structure is unknown. Consistent estimates for the parameters in the mean function are still available with robust variance estimates.

The last step of the cycle has not yet been well developed in the literature. The tests of homogeneity we proposed may serve as goodness-of-fit tests for fixed-effect models. As discussed in the above sections, satisfactory model diagnostics for random effect and marginal models are still unavailable.

Methodology for the analysis of event history data continues to be the focus of much ongoing research in statistics. This thesis contributes some potentially useful models and methods to this field, and provides useful insight for further research.

Appendix A

Tests of Homogeneity

A.1 Parametric Models

A.1.1 Expression for $I_{\theta\sigma}(\boldsymbol{\theta})$

From (3.5.2), one can easily show that

$$\frac{\partial T_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\frac{1}{2}\sum_{i=1}^{m}\begin{pmatrix} (2(n_i - \Lambda_i(\boldsymbol{\theta})) + 1)\Lambda_i(\boldsymbol{\theta})\boldsymbol{x}_i \\ (2(n_i - \Lambda_i(\boldsymbol{\theta})) + 1)\partial\Lambda_i(\boldsymbol{\theta})/\partial\boldsymbol{\gamma} \end{pmatrix},$$

where $\Lambda_i(\boldsymbol{\theta}) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\Lambda_0(\tau_i)$. Under H_0, the n_i are independent and distributed as Poisson with mean $\Lambda_i(\boldsymbol{\theta})$, therefore, we have

$$I_{\theta\sigma}(\boldsymbol{\theta}) = \mathrm{E}[-\frac{\partial T_p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}] = \frac{1}{2}\sum_{i=1}^{m}\begin{pmatrix} \Lambda_i(\boldsymbol{\theta})\boldsymbol{x}_i \\ \partial\Lambda_i(\boldsymbol{\theta})/\partial\boldsymbol{\gamma} \end{pmatrix}.$$

A.1.2 Expression for $I_\theta(\boldsymbol{\theta})$

The log-likelihood under H_0 is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left[n_i \boldsymbol{x}_i' \boldsymbol{\beta} + \sum_{j=1}^{n_i} \log \lambda_0(t_{ij}) - \exp(\boldsymbol{x}_i' \boldsymbol{\beta}) \Lambda_0(\tau_i) \right].$$

Thus, the score functions are easily seen to be

$$U_\beta(\boldsymbol{\theta}) = \sum_{i=1}^{m} [n_i - \Lambda_i(\boldsymbol{\theta})] \boldsymbol{x}_i \quad \text{and} \quad U_\gamma(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left[\sum_{j=1}^{n_i} \frac{\partial \log \lambda_0(t_{ij})}{\partial \gamma} - \frac{\partial \Lambda_i(\boldsymbol{\theta})}{\partial \gamma} \right].$$

By partitioning the expected Fisher information matrix conformably to $(\boldsymbol{\beta}', \gamma')$, one can show that

$$I_\theta(\boldsymbol{\theta}) = \begin{pmatrix} \sum_{i=1}^{m} \Lambda_i(\boldsymbol{\theta}) \boldsymbol{x}_i \boldsymbol{x}_i' & \sum_{i=1}^{m} \boldsymbol{x}_i \, \partial \Lambda_i(\boldsymbol{\theta}) / \partial \gamma' \\ \sum_{i=1}^{m} \partial \Lambda_i(\boldsymbol{\theta}) / \partial \gamma \, \boldsymbol{x}_i' & I_\gamma(\boldsymbol{\theta}) \end{pmatrix},$$

where

$$I_\gamma(\boldsymbol{\theta}) = \sum_{i=1}^{m} \left[-\mathrm{E} \left(\sum_{j=1}^{n_i} \frac{\partial^2 \log \lambda_0(t_{ij})}{\partial \gamma \partial \gamma'} \right) + \frac{\partial^2 \Lambda_i(\boldsymbol{\theta})}{\partial \gamma \partial \gamma'} \right].$$

The expectation in $I_\gamma(\boldsymbol{\theta})$ can be obtained by noting that given $N_i(\tau_i) = n_i$, the event times $t_{i1}, ..., t_{in_i}$ are i.i.d. with probability density function $f(t|n_i) = \lambda_0(t)/\Lambda_0(\tau_i)$, for $0 \leq t \leq \tau_i$. Therefore, we have

$$\begin{aligned} \mathrm{E} \left(\sum_{j=1}^{n_i} \frac{\partial^2 \log \lambda_0(t_{ij})}{\partial \gamma \partial \gamma'} \right) &= \mathrm{E} \left[\mathrm{E} \left(\sum_{j=1}^{n_i} \frac{\partial^2 \log \lambda_0(t_{ij})}{\partial \gamma \partial \gamma'} \middle| N_i(\tau_i) = n_i \right) \right] \\ &= \mathrm{E} \left(\frac{n_i}{\Lambda_0(\tau_i)} \int_0^{\tau_i} \frac{\partial^2 \log \lambda_0(t_{ij})}{\partial \gamma \partial \gamma'} d\Lambda_0(t) \right) \\ &= \exp(\boldsymbol{x}_i' \boldsymbol{\beta}) \int_0^{\tau_i} \frac{\partial^2 \log \lambda_0(t_{ij})}{\partial \gamma \partial \gamma'} d\Lambda_0(t). \end{aligned}$$

Depending on the functional form of $\lambda_0(t)$, one may need to evaluate the above integral numerically. Nevertheless, closed form expressions are available for Weibull and piecewise exponential specifications.

A.1.3 Proof of $\mathbf{cov}(n_i, U(\boldsymbol{\theta})) = \partial\Lambda_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$

Since the $N_i(t)$ are independent, $\mathrm{cov}(n_i, U(\boldsymbol{\theta})) = \mathrm{cov}(n_i, U_i(\boldsymbol{\theta}))$, where $U_i(\boldsymbol{\theta})$ is the score function for subject i. First we consider the covariance between n_i and the score function of $\boldsymbol{\beta}$:

$$\mathrm{cov}(n_i, (n_i - \Lambda_i(\boldsymbol{\theta}))\boldsymbol{x}_i) = \Lambda_i(\boldsymbol{\theta})\boldsymbol{x}_i = \partial\Lambda_i(\boldsymbol{\theta})/\partial\boldsymbol{\beta},$$

because n_i is distributed as Poisson with mean $\Lambda_i(\boldsymbol{\theta})$. Next, using the argument given in section A.2, the covariance between n_i and the score function of $\boldsymbol{\gamma}$ is equal to

$$\begin{aligned}
\mathrm{cov}\left(n_i, \sum_{j=1}^{n_i} \frac{\partial \log \lambda_0(t_{ij})}{\partial\boldsymbol{\gamma}} - \frac{\partial\Lambda_i(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}}\right) &= \mathrm{cov}\left[n_i, \mathrm{E}\left(\sum_{j=1}^{n_i} \frac{\partial \log \lambda_0(t_{ij})}{\partial\boldsymbol{\gamma}} \middle| N_i(\tau_i) = n_i\right)\right] \\
&= \mathrm{cov}\left(n_i, \frac{n_i}{\Lambda_0(\tau_i)} \int_0^{\tau_i} \frac{\partial \log \lambda_0(t)}{\partial\boldsymbol{\gamma}} d\Lambda_0(t)\right) \\
&= \frac{\mathrm{var}(n_i)}{\Lambda_0(\tau_i)} \frac{\partial\Lambda_0(\tau_i)}{\partial\boldsymbol{\gamma}} \\
&= \frac{\partial\Lambda_i(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}}.
\end{aligned}$$

Hence,

$$\mathrm{cov}(n_i, U(\boldsymbol{\theta})) = \frac{\partial\Lambda_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}.$$

A.2 Semi-parametric Models

A.2.1 Derivation of the Score Statistic

The score statistic is given by

$$T_{sp}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{m}\left[\left(\frac{\partial \log(L_{pi}(\boldsymbol{\beta},\sigma|u_i))}{\partial \eta_i}\right)^2 + \frac{\partial^2 \log(L_{pi}(\boldsymbol{\beta},\sigma|u_i))}{\partial \eta_i^2}\right]_{\sigma=0},$$

where $L_{pi}(\boldsymbol{\beta},\sigma|u_i)$ is the conditional partial likelihood due to the i subject in (3.5.9) and $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma u_i$. It is straightforward to show that

$$\left.\frac{\partial \log(L_{pi}(\boldsymbol{\beta},\sigma|u_i))}{\partial \eta_i}\right|_{\sigma=0} = \int_0^\infty (1 - w_i(s))\, dN_i(s) = \hat{M}_i(\boldsymbol{\beta})$$

$$\left.\frac{\partial^2 \log(L_{pi}(\boldsymbol{\beta},\sigma|u_i))}{\partial \eta_i^2}\right|_{\sigma=0} = \int_0^\infty \left(w_i^2(s) - w_i(s)\right) dN_\cdot(s).$$

A.2.2 Martingale Representation of Score Statistic

By the integration by part formula (Fleming and Harrington, 1991 Theorem A.1.2), we can express the squared martingale residual (3.5.11) as

$$\hat{M}_i^2(t,\boldsymbol{\beta}) = 2\int_0^t \hat{M}_i(s-,\boldsymbol{\beta}) d\hat{M}_i(s,\boldsymbol{\beta}) + \sum_{0\le s\le t} \triangle \hat{M}_i^2(s,\boldsymbol{\beta}). \qquad \text{(A.2.1)}$$

Since no two processes can jump at the same time and the $N_i(\cdot)$ are orderly processes, the last term in (A.2.1) can be written as

$$\begin{aligned}\sum_{0\le s\le t} \triangle \hat{M}_i^2(s,\boldsymbol{\beta}) &= \int_0^t (dN_i(s) - w_i(s) dN_\cdot(s))^2 \\ &= N_i(t) + \int_0^t w_i^2(s) dN_\cdot(s) - 2\int_0^t w_i(s) dN_i(s)\end{aligned}$$

Furthermore, by (3.5.8) with $\sigma = 0$, we can rewrite

$$d\hat{M}_i(t,\beta) = dM_i(t) - w_i(t)dM_\cdot(t).$$

Therefore, the score statistic at time t is given by

$$\begin{aligned} T_{sp}(t,\beta) &= \frac{1}{2}\sum_{i=1}^{m}\Big(2\int_0^t \hat{M}_i(s-,\beta)[dM_i(s) - w_i(s)dM_\cdot(s)] + N_i(t) + \int_0^t w_i^2(s)dN_\cdot(s) \\ &\qquad -2\int_0^t w_i(s)dN_i(s) - \int_0^t (w_i(s) - w_i^2(s))dN_\cdot(s)\Big) \\ &= \sum_{i=1}^{m}\Big(\int_0^t \hat{M}_i(s-,\beta)[dM_i(s) - w_i(s)dM_\cdot(s)] + \int_0^t w_i^2(s)[dM_\cdot(s) + S^{(0)}d\Lambda_0(s)] \\ &\qquad - \int_0^t w_i(s)[dM_i(s) + Y_i(s)\exp(\boldsymbol{x}_i'\beta)d\Lambda_0(s)]\Big) \\ &= \sum_{i=1}^{m}\int_0^t H_i(s,\beta)dM_i(s). \end{aligned}$$

A.2.3 Proof of $\mathbf{cov}(\hat{M}_i(\beta), U(\beta)) = -\mathbf{E}(J_i(\beta))$

The score function under H_0 at time t is given by (Fleming and Harrington, 1991 p150)

$$U(t,\beta) = \sum_{i=1}^{m}\int_0^t (\boldsymbol{x}_i - E(s))dM_i(s).$$

The predictable covariance process of $\hat{M}_i(t,\beta)$ and $U(t,\beta)$ is then equal to

$$< \hat{M}_i(\cdot,\beta), U(\cdot,\beta) > (t)$$

$$=< \int_0^{\cdot} dM_i(s) - \int_0^{\cdot} w_i(s) dM_{\cdot}(s), \sum_{j=1}^{m} \int_0^{\cdot} (\boldsymbol{x}_j - E(s)) dM_j(s) > (t)$$

$$= \int_0^t (\boldsymbol{x}_i - E(s)) d < M_i > (s) - \sum_{j=1}^{m} \int_0^t (\boldsymbol{x}_j - E(s)) w_i(s) d < M_j > (s)$$

$$= \int_0^t (\boldsymbol{x}_i - E(s)) Y_i(s) \exp(\boldsymbol{x}_i' \boldsymbol{\beta}) d\Lambda_0(s),$$

because $\sum_{j=1}^{m} (\boldsymbol{x}_j - E(s)) Y_j(s) \exp(\boldsymbol{x}_j' \boldsymbol{\beta}) = 0$ for $s \geq 0$. Hence,

$$\text{cov}(\hat{M}_i(\boldsymbol{\beta}), U(\boldsymbol{\beta})) = \text{E}[< \hat{M}_i(\cdot, \boldsymbol{\beta}), U(\cdot, \boldsymbol{\beta}) > (\infty)] = -\text{E}(J_i(\boldsymbol{\beta})).$$

References

Aalen, O.O. (1987a). Two examples of modelling heterogeneity in survival analysis. *Scand. J. Statist.* **6**, 701–726.

Aalen, O.O. (1987b). Mixing distributions on a Markov chain. *Scand. J. Statist.* **14**, 281-289.

Aalen, O.O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine* **7**, 1121–1137.

Aalen, O.O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *The Annals of Applied Probability* **2**, 951–972.

Aalen, O.O. (1994). Effects of frailty in survival analysis. *Statistical Methods in Medical Research* **3**, 227–243.

Aalen, O.O. and Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine* **10**, 1227–1240.

Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions*. Dover.

Andersen, P.K. (1986). Time-dependent covariates and Markov processes. In Moolgavkar, S.H. and Prentice, R.L., editors, *Modern Statistical Methods in Chronic Disease Epidemiology*. John-Wiley: New York, 82–103

Andersen, P.K. (1988). Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Statistics in Medicine* **7**, 661–670.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag: New York.

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100–1120.

Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J. Am. Statist. Assoc.* **83**, 204–212.

Baltazar-Aban, I. and Pena, E.A. (1995). Properties of hazard-based residuals and implications in model diagnostics. *J. Am. Statist. Assoc.* **90**, 185–197.

Barlow, W.E. and Prentice, R.L. (1988). Residuals for relative risk regression. *Biometrika* **75**, 65–74.

Breslow, N.E. (1989). Score tests in overdispersed GLM's. In *Proceedings of GLIM 89 and the Fourth International Workshop on Statistical Modelling*, eds. A. Decarli, B.J. Francis, R. Gilchrist, and G.U.H. Seeber, Springer-Verlag: New York, 64–74.

Breslow, N.E. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *J. Am. Statist. Assoc.* **85**, 565–571.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

Bretagnolle, J. and Huber-Carol, C. (1988). Effects of omitting covariates in Cox's model for survival data. *Scand. J. Statist.* **15**, 125–138.

Butler, J.S., Anderson, K.H. and Burkhauser, R.V. (1989). Work and health after retirement: a competing risk model with semiparametric unobserved heterogeneity. *Review of Economics and Statistics* **71**, 47-53.

Butler, S.M. and Louis, T.A. (1992). Random effects models with non-parametric priors. *Statistics in Medicine* **11**, 1981–2000.

Byar, D.P. (1980). The Veterans Administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine, and topical thiotepa. In *Bladder Tumors and Other Topics in Urological Oncology* edited by M. Pavone-Macaluso, P.H. Smith and F. Edsmyr. New York: Plenum, 363–370.

Chesher, A.D. (1983). The information matrix test: simplified calculation via a score test interpretation. *Economics Letters* **13**, 45–48.

Chesher, A.D. (1984). Testing for neglected heterogeneity. *Econometrica* **52**, 865–872.

Chesher, A.D. and Spady, R. (1991). Asymptotic expansions of the information matrix test statistic. *Econometrica* **59**, 787–815.

Chunge, R.N. (1989). Intestinal parasites in a rural community in Kiambu district, Kenya with special reference to giardia lamblia. *Ph.D. Thesis, University College Galway.*

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency toward chronic disease. *Biometrika* **65**, 141–151.

Clayton, D.G. (1988). The analysis of event history data: a review of progress and outstanding problems. *Statistics in Medicine* **7**, 819–841.

Clayton, D. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467–486.

Clayton, D. (1994). Some approaches to the analysis of recurrent event data. *Statistical Methods in Medical Research* **3**, 244–262.

Clayton, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *J. R. Statist. Soc. A* **148**, 82–117.

Commenges, D. and Andersen, P.K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis* **1**, 145–156.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall.

Cook, R.J. and Ng, E.T.M. (1997). A logistic-bivariate normal model for overdispersed two-state Markov processes. *Biometrics* **53**, 358–364.

Cox, D.R. (1962). *Renewal Theory.* John Wiley & Sons: New York.

Cox, D.R. (1972a). The statistical analysis of dependencies in point processes. In *Stochastic Point Processes*, ed. P.A.W. Lewis, John Wiley & Sons: New York.

Cox, D.R. (1972b). Regression models and life-tables (with discussion). *J. R. Statist. Soc. B* **34**, 187–220.

Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D.R. and Isham, V. (1980). *Point Processes.* Chapman & Hall.

Crouchley, R. and Pickles, A. (1993). A specification test for univariate and multivariate proportional hazards models. *Biometrics* **49**, 1067–1076.

Crowder, M. (1985). Gaussian estimation for correlated binomial data. *J. R. Statist. Soc. B* **47**, 229–237.

Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika* **74**, 591–597.

Dean, C. (1988). Mixed Poisson models and regression methods for count data. *PhD Thesis, University of Waterloo.*

Dean, C. (1992). Testing for overdispersion in Poisson and binomial regression models. *J. Am. Statist. Assoc.* **87**, 451–457.

Dempster, A.P., Laird, N. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B* **39**, 1–38.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall.

Elbers, C. and Ridder, G. (1982). True and spurious duration dependence: The identifiability of the proportional hazard model. *Review of Economic Studies* **49**, 403–409.

Evans, M. and Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science* **10**, 254–272.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer-Verlag.

Farrington, C.P. (1996). On assessing goodness of fit of generalized linear models to sparse data. *J. R. Statist. Soc. B* **58**, 349–360.

Fleming and Harrington (1991). *Counting Processes and Survival Analysis.* John Wiley & Sons: New York.

Fietta, A.M., Merlini, C., Uccelli, M., Grassi, G. and Grassi, C. (1992). Immunological and clinical effect of long-term oral treatment with RU 41740 in patients with chronic

bronchitis: double-blind trial long-term versus standard dose regimen. *Respiration* **59**, 253–258.

Frank, E., Kupfer, D.F., Perel, J.M., Cornes, C., Jarrett, D.B., Maillinger, A.G., Thase, M.E., McEachran, A.B. and Grochocinski, V.J. (1990). Three-year outcomes for maintenance therapies in recurrent depression. *Arch. Gen. Psychiatry* **47**, 1093–1099.

Gail, M.H., Santner, T.J. and Brown, C.C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics* **36**, 255–266.

Gail, M.H., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effects in randomized experiments with non-linear regressions and omitted covariates. *Biometrika* **71**, 431–444.

Gamerman, D. (1991). Dynamic Bayesian models for survival data. *Applied Statistics* **40**, 63–79.

Genest, C. and MacKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *Am. Statistician* **40**, 280–283.

Godambe, V.P. and Heyde, C.C. (1987). Quasi-likelihood and optimal estimation. *International Statistical Review* **55**, 231–244.

Grambsch, P.M. and Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526.

Godambe, V.P. and Thompson, M.E. (1989). An extension of quasi-likelihood estimation. *J. Statistical Planning and Inference* **22**, 137–152.

Gray, R.J. (1995). Tests for variation over groups in survival data. *J. Am. Statist. Assoc.* **90**, 198–203.

Guo, G. (1993). Event-history analysis for left-truncated data. *Sociological Methodology* **23**, 217–243.

Heckman, J.J. and Singer, B. (1982). Population heterogeneity in demographic models. In *Multidimensional Mathematical Demography*, 567–595.

Heckman, J.J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models of duration data. *Econometrica* **52**, 271–320.

Heckman, J.J. and Taber, C.R. (1994). Econometric mixture models and more general models for unobservables in duration analysis. *Statistical Methods in Medical Research* **3**, 279–299.

Honore, B.E. (1993). Identification results for duration models with multiple spells. *Review of Economic Studies* **60**, 241–246.

Hougaard, P. (1986a). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73**, 387–396.

Hougaard, P. (1986b). A class of multivariate failure time distributions. *Biometrika* **73**, 671–678.

Hougaard, P. (1987). Modelling multivariate survival. *Scand. J. Statist.* **14**, 291–304.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis* **1**, 255-273.

Hougaard, P., Harvald, B. and Holm, N.V. (1992). Measuring the similarities between the lifetimes of adult Danish twins born between 1881–1930. *J. Am. Statist. Assoc.* **87**, 17–24.

International Headache Society Committee on Clinical Trials in Migraine (1991). Guidelines for controlled trials of drugs in migraine. Cephalalgia, II, 1–12.

Jacqmin-Gadda, H. and Commenges, D. (1995). Tests of homogeneity for generalized linear models. *J. Am. Statist. Assoc.* **90**, 1237–1246.

Jamshidian, M. and Jennrich, R.I. (1997). Acceleration of the EM Algorithm by using quasi-Newton methods. *J. R. Statist. Soc. B* **59**, 569–587.

Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *J. R. Statist. Soc. B* **40**, 214–221.

Kalbfleisch, J.D. and Lawless, J.F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine* **7**, 149–160.

Kalbfleisch, J.D., Lawless, J.F., and Robinson, J.A. (1991). Methods for the analysis and prediction of warranty claims. *Technometrics* **33**, 273–285.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley & Sons: New York.

Karatzas, I. and Shreve, S.E. (1991). *Brownian Motion and Stochastic Calculus.* 2nd ed. Springer-Verlag: New York.

Karlin, S. and Taylor, H.M. (1975). *A First Course in Stochastic Processes.* 2nd ed. Academic Press.

Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics* **26**, 227–237.

Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 759–806.

Kuk, A.Y.C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *J. R. Statist. Soc. B* **57**, 395–407.

Lancaster, T. (1984). The covariance matrix of the information matrix test. *Econometrica* **52**, 1051–1053.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data.* Cambridge University Press: New York.

Lancaster, T. and Nickell, S. (1980). The analysis of re-employment probabilities for the unemployed. *J. R. Statist. Soc. A* **143**, 141–165.

Laird, N. (1978). Non-parametric maximum likelihood estimation of a mixing distribution. *J. Am. Statist. Assoc.* **73**, 805–811.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* John Wiley & Sons: New York.

Lawless, J.F. (1987). Regression methods for Poisson process data. *J. Am. Statist. Assoc.* **82**, 808–815.

Lawless, J.F. (1995). The analysis of recurrent events for multiple subjects. *Applied Statistics* **44**, 487–498.

Lawless, J.F. and Fong, D.Y.T. (1997). State duration models in clinical and observational studies. *Manuscript.*

Lawless, J.F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**, 158–168.

Lawless, J.F. and Thiagarajah, K. (1996). A point-process model incorporating renewals and time trends, with application to reparable systems. *Technometrics* **38**, 131–138.

Liang, K.Y. (1987). A locally most powerful test for homogeneity with many strata. *Biometrika* **74**, 295–264.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *J. Royal Statist. Soc. B* **54**, 3–40.

Liang, K.Y., Self, S.G., Bandeen-Roche, K.J. and Zeger, S.L. (1995). Some recent developments for regression analysis of multivariate failure time data. *Lifetime Data Analysis* **1**, 403–415.

Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine* **13**, 2233–2247.

Lin, D.Y. and Wei, L.J. (1989). The robust inference for the Cox proportional hazards model. *J. Am. Statist. Assoc.* **84**, 1074–1078.

Lin, D.Y. and Wei, L.J. (1991). Goodness-of-fit tests for the general Cox regression model. *Statistica Sinica* **1**, 1–17.

Lin, D.Y., Wei, L.J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

Lin, X. (1997). Variance components testing in generalized linear models with random effects. To appear in *Biometrika*.

Lindeboom, M. and Van Den Berg, G.J. (1994). Heterogeneity in models for bivariate survival: the importance of the mixing distribution. *J. R. Statist. Soc. B* **56**, 49–60.

Lindley, D.V. and Singpurwalla, N.A. (1986). Multivariate distributions for the reliability of a system of components sharing a common environment. *J. Applied Probability* **23**, 418–431.

Lindsey, J.C. and Ryan, L.M. (1993). A three-state multiplicative model for rodent tumorigenicity experiments. *Applied Statistics* **42**, 283–300.

Louis, T.A. (1982). Finding observed information using the EM algorithm. *J. R. Statist. Soc. B* **44**, 98–130.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman & Hall.

McGilchrist, C.A. (1993). REML estimation for survival models with frailty. *Biometrics* **49**, 221–225.

McGilchrist, C.A. and Aisbett, C.W. (1991). Regression with frailty in survival analysis. *Biometrics* **47**, 461–466.

McLachlan, G.J. and Basford, K.E. (1987). *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker: New York and Basel.

Meng, X.L. and van Dyk, D. (1997). The EM algorithm — an old folk-song sung to a fast new tune (with discussion). *J. R. Statist. Soc. B* **59**, 511–567.

Meng, X.L. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices. *J. Am. Statist. Assoc.* **86**, 899–909.

Moran, P.A. (1971). Maximum likelihood estimation in non-standard conditions. *Proceedings of the Cambridge Philosophical Society* **70**, 441–450.

Murphy, S.A. (1994). Consistency in proportional hazards model incorporating a random effect. *Annals of Statistics* **22**, 712–731.

Murphy, S.A. (1995). Asymptotic theory for the frailty model. *Annals of Statistics* **23**, 182–198.

Nadeau, C. and Lawless, J.F. (1996). Inference for means and covariances of point processes through estimating functions. Technical Report STAT-96-02, Dept. of Statist. and Act. Sci., University of Waterloo.

Nagelkerke, N.J.D., Chunge, R.N. and Kinoti, S.N. (1990). Estimation of parasitic infection dynamics when detectability is imperfect. *Statistics in Medicine* **9**, 1211–1219.

Neuhaus, J.M., Hauck, W.W. and Kalbfleisch, J.D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755–762.

Ng, E.T.M. and Cook, R.J. (1996). Alternating renewal processes with bivariate random effects. Technical Report STAT-96-06, Dept. of Statist. and Act. Sci., University of Waterloo.

Nielsen, G.G., Andersen, P.K., Gill, R.D. and Sorensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.* **19**, 25–43.

Oakes, D. (1989). Bivariate survival models induced by frailties. *J. Am. Statist. Assoc.* **84**, 487-493.

Oakes, D. and Cui, L. (1994). On semi-parametric inference for modulated renewal processes. *Biometrika* **81**, 83–90.

O'Hara Hines, R.J. (1995). A comparison of scores tests for overdispersion in generalized linear models. *Tech. Report STAT 95-02, Dept. of Stats. & Act. Sci., University of Waterloo.*

O'Quigley, J. and Pessione, F. (1989). Score tests for homogeneity of regression effect in the proportional hazard model. *Biometrics* **45**, 135–144.

Orme, C. (1990). The small-sample performance of the information matrix test. *J. of Econometrics* **46**, 309–331.

Paik, M.C., Tsai, W.Y. and Ottman, R. (1994). Multivariate survival analysis using piecewise gamma frailty. *Biometrics* **50**, 975–988.

Petersen, J.H., Andersen, P.K. and Gill, R.D. (1995). Variance components models for survival data. Research report, Department of Biostatistics, University of Copenhagen, Denmark.

Pickles, A. (1994). Generalizations and applications of frailty models for survival and event data. *Statistical Methods in Medical Research* **3**, 263–278.

Pickles, A. and Crouchley, R. (1995). A comparison of frailty models for multivariate survival data. *Statistics in Medicine* **14**, 1447–1461.

Pierce, D.A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Annals of Statistics* **10**, 475–478.

Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous response. *Biometrics* **47**, 825–839.

Prentice, R.L., Williams, B.J. and Peterson, A.V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68**, 373–379.

Rai, S.N. and Matthews, D.E. (1993). Improving the EM algorithm. *Biometrics* **49**, 587–591.

Rochon, J. (1996). Accounting for covariates observed post randomization for discrete and continuous repeated measures data. *J. R. Statist. Soc. B* **58** 205–219.

Rokkas, T., Karameris and A., Maurogeorgis, A. (1995). Eradication of helicobacter pylori reduces the possibility of rebleeding in peptic ulcer disease. *Gastrointestinal Endoscopy* **41**, 1–4.

Sears, M.R., Taylor, D.R., Print, C.G., Lake, D.C., Li, Q., Flannery, E.M., Yates, D.M., Lucas, M.K. and Herbison, G.P. (1990). Regular inhaled beta-agonist treatment in bronchial asthma. *Lancet* **336**, 1391–1396.

Shih, J.H. and Louis, T.A. (1995). Assessing gamma frailty models for clustered failure time data. *Lifetime Data Analysis* **1**, 205–220.

Smith, P.J. and Heitjan, D.F. (1993). Testing and adjusting for departures from nominal dispersion in generalized linear models. *Applied Statistics* **42**, 31–41.

The SOLVD Investigators (1991). Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *New England J. Medicine* **325**, 293–302.

Struthers, C.A. and Kalbfleisch, J.D. (1986). Misspecified proportional hazard models. *Biometrika* **73**, 363–369.

Sutradhar, B.C. and Das, K. (1995). Analysing exponential family based messy data. *Commun. Statist. — Theory and Methods* **24**, 2683–2699.

Therneau, T.M., Grambsch, P.M. and Fleming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147–160.

Tibshirani, R.J. and Ciampi, A. (1983). A family of proportional and additive hazards models for survival data. *Biometrics* **39**, 141–147.

Vaupel, J.W., Harvald, B., Holm, N.V., Yashin, A. and Xiu, L. (1992). Survival analysis in genetics: Danish twin data applied to a gerontological question. In *Survival Analysis:*

State of the Art. Eds. J.P. Klein and P.K. Goel, 121–138.

Vaupel, J.W., Manton, K.G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.

Vaupel, J.W., Yashin, A.I. (1985). Heterogeneity's ruses: Some surprising effects of selection on population dynamics. *The American Statistician* **39**, 176–185.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Am. Statist. Assoc.* **91**, 217–221.

Wei, L.J. (1984). Testing goodness-of-fit for proportional hazards model with censored observations. *J. Am. Statist. Assoc.* **79**, 649–652.

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

Whitemore, G.A. and Lee, M.L.T. (1991). A multivariate survival distribution generated by an inverse Gaussian mixture of exponentials. *Technometrics* **33**, 39–50.

Wolfe, J.H. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. *Technical Bulletin STB 72-2.* San Diego: U.S. Naval Personnel and Training Research Laboratory.

Xue, X. and Brookmeyer, R. (1996). Bivariate frailty model for the analysis of multivariate survival time. *Lifetime Data Analysis* **2**, 277–289.

Yashin, A., Vaupel, J.W. and Iachine, I. (1995). Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies* **5**, 145–159.

Yau, K.K.W. and McGilchrist, C.A. (1996). ML and REML estimation in survival analysis with time dependent correlated frailty. *To appear in Statistics in Medicine.*

Yue, H. and Chan, K.S. (1997). A dynamic frailty model for multivariate survival data. *Biometrics* **53**, 785–793.

Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–1060.