SPLINE MODELS FOR THE ANALYSIS OF

RECURRENT EVENT PANEL DATA

by

Jason D. Nielsen

B.Sc., University of Calgary, 1999

M.Sc., Simon Fraser University, 2001

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in the Department

of

Statistics and Actuarial Science

© Jason D. Nielsen 2007

SIMON FRASER UNIVERSITY

Summer 2007

APPROVAL

Name: Jason D. Nielsen

Degree: Doctor of Philosophy

Title of thesis: Spline Models for the Analysis of Recurrent Event Panel Data

Examining Committee: Dr. Richard Lockhart
Chair

Dr. Charmaine Dean,
Simon Fraser University
Senior Supervisor

Dr. Robert Balshaw,
Supervisory Committee

Dr. Carl Schwarz,
Supervisory Committee

Dr. Brad McNeney,
Internal External Examiner

Dr. Nancy Heckman,
External Examiner,
University of Britich Columbia

Date Approved: May 28, 2007

ii

Abstract

There has been a substantial interest in longitudinal studies, particularly for monitoring changes in profiles of specific sub-sectors of society, for example, Statistics Canada's Canadian National Longitudinal Study of Children and Youth (Statistics Canada, 1996), the Women's Health Australia national longitudinal study (Women's Health Australia, 2005) and the U.S. National Longitudinal Surveys on Labor Statistics (National Longitudinal Surveys Handbook, 2005). For these and many longitudinal studies, so-called panel data are collected, with information gathered between specific follow-up times. When interest focuses on multiple or recurrent episodes of an event of interest, recurrent event panel data arise, where only information on the number of recurrences between follow-up times is recorded. Such data collection designs are typical in clinical studies where it is not possible to record exact event times, for example, if examinations are invasive or occur too frequently as in the study of chronic diseases such as epilepsy (Thall and Vail, 1990) or certain incidences of tumors in cancer patients (Abu-Libdeh et al., 1990).

This thesis discusses semiparametric methods for the analysis of recurrent event panel data and offers a comprehensive framework for such analysis requiring only minimal distributional assumptions. The basic model assumes that the counts for each subject are generated by a mixed nonhomogeneous Poisson process (NHPP) where frailties account for heterogeneity common to this type of data. The generating intensity of the counting process is assumed to be a smooth function modeled with splines. Covariate effects are also represented as splines; this permits covariate effects to change over time. The development offers several special limiting cases which are common, for example, a constant intensity, or fixed covariate effects. The thesis also considers discrete mixtures of these mixed NHPP models accommodating clusters of hidden sub-populations which generate counts with differing

intensity functions. Several recent applications investigated suggested a need for accommodating such unobservable sub-populations. For example, in the motivating application that is used throughout this thesis, moth matings in the summer seem to be generated by emergence of at least two types of moths in the spring: those which overwinter in the pupal stage and emerge earlier in the spring, and those which overwinter in the egg stage; the finite mixture approach accommodates this type of behavior. The thesis concludes with a discussion of several areas for further investigation in this important field of study.

Acknowledgments

I would like to begin by thanking my senior supervisor Dr. Charmaine Dean. The support, guidance and the many opportunities she has provided have had a profound influence on my life. I will be forever grateful.

Many thanks to the faculty and staff of the Department of Statistics and Actuarial Science of Simon Fraser University for all their assistance and for providing such an excellent environment in which to study. A special thanks to Dr. Michael Stephens for all the good advice and the wonderful trip to Uruguay. To Giovani Silva for his friendship and showing me around Brazil. A high five to the members of the department basketball team for all the spirited games. I hope you eventually win the coveted championship t-shirt! I would also like to thank Laurie Ainsworth, Chunfang Lin, Crystal Linkletter, Jason Loeppky, Farouk Nathoo, Pritam Ranjan, Darby Thompson and all the other great graduate students at SFU I have had the pleasure of meeting.

I am also very grateful to my parents for a lifetime of love. I cannot express how much you mean to me. To my brothers Devin and Steven for the good times, particularly those amazing ski trips.

Last but not least I would like to thank Maria Lorenzi, my marvelous little lady. Your love and support have been a constant source of strength.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The development of methods for longitudinal data analysis has seen tremendous growth over the last few decades. In particular, methods which incorporate smoothers and time-varying coefficients have provided more flexible and realistic methods for data analysis. Much of this work has been in the context of multivariate normal data, or data which may be approximated by normality assumptions. Studies of a longitudinal nature abound in many applications, from health to economics and in many instances such longitudinal data arise as counts, with the processes being monitored giving rise to recurrent events. This field may also be considered as an extension of survival analysis, where, instead of monitoring the time to a (typically crucial or terminal) event, times to recurrences are available and it is of interest to determine intensity functions generating such recurrences.

This thesis develops methods for the analysis of longitudinal count data arising from such recurrent event processes. There are special features which will be considered and one major feature is that of missing information as typical when processes are not continuously monitored but observed at specific time periods.

We begin with a review of some preliminary ideas that are the building blocks of the models developed in later chapters followed by an outline of the material presented in these chapters.

1.1 Recurrent Event Panel Data

In recurrent event studies objects or subjects are each observed for a period of time and the number of recurrences of a phenomenon of interest is recorded. For the ith subject, let

$[T_{io}, T_{ie_i})$ denote the total observation period and let $T_{i1} < T_{i2} < T_{i3} < \cdots < T_{ie_i}$ where $T_{ij} \in (T_{io}, T_{ie_i}]$, $j = 1, 2, \ldots, e_i$, denote followup times, the times at which the subjects are examined for recurrences. Let I denote the number of individuals under study so i ranges from 1 to I. In general the start of the observation period T_{io} may not be the same for all individuals, but here, they are all set to zero as the time scale of interest is time from T_{io} and marks a particular intervention such as diagnosis, surgery or entry into the study. In some cases, it is important to work with calendar time, rather than or in addition to time from T_{io}. Note that the followup times T_{ij} may be different for each individual to accommodate different followup schedules as well as drop-out.

Ideally the type of data one would like to collect are the exact times of the events of interest for each individual i that occur in $[T_{io}, T_{ie})$. To collect such data each individual needs to be monitored continuously. Such monitoring can be costly, considered as overly intrusive, particularly in human health studies that do not involve chronic care, or simply not possible as may be the case with animal populations for instance. Instead what is commonly done in such situations is to simply collect the number of events, N_{ij}, that occur in each observation window or panel $[T_{i,j-1}, T_{ij})$. For each individual a vector of counts $\mathbf{N}_i = [N_{i1}, \ldots, N_{ie_i}]^T$ is recorded. Panel count data of this form is the type that will be considered in this thesis as it is quite generally the type of data that is collected. In fact, one could argue that in practice all data are collected in this fashion as we do not measure time continuously but record events by the second or microsecond which would then define the lengths of the panels for the study. Of course, if these observation windows are fine enough so that the counts are simply the presence or absence of an event then the use of models based upon exact event time responses is reasonable.

1.2 The Nonhomogeneous Poisson Process Model

The Poisson process is a stochastic process that counts the number of randomly occurring events over time. It is treated in detail in most books on stochastic processes such as Karlin and Taylor (1975) and Ross (1996). A nonhomogeneous Poisson process (NHPP) is a modification of a Poisson process which allows the intensity of the process to be time dependent. Let $N[t, t + h)$ be the number of events occurring in the time interval $[t, t + h)$, where $t \geq 0$ and $h \geq 0$ and let $N(t) = N[0, t)$. Then a process $\{N(t), t \geq 0\}$ is said to be a

nonhomogeneous Poisson process with intensity $\lambda(t)$ if

$$1) \quad N(0) = 0 \tag{1.1}$$

$$2) \quad \{N(t), t \geq 0\} \text{ has independent increments} \tag{1.2}$$

$$3) \quad P\{N(t, t+h) = 0\} = 1 - \lambda(t)h + o(h) \tag{1.3}$$

$$4) \quad P\{N(t, t+h) = 1\} = \lambda(t)h + o(h) \tag{1.4}$$

where $\lambda(t) \geq 0$ is called the intensity function of the process and $o(h)$ denotes a remainder quantity $g(h)$ which approaches zero faster than h, i.e. $\lim_{h \to 0} \frac{g(h)}{h} = 0$. The intensity function $\lambda(t)$ is continuous and

$$\Lambda(t) = \int_0^t \lambda(u) \mathrm{d}u$$

is called the cumulative intensity function. Conditions (1.3) and (1.4) imply that $P\{N[t, t + h) \geq 2\} = o(h)$ and that the occurrence of events prior to time t does not affect those in $[t, t + h)$. This gives rise to the property that counts in non-overlapping intervals are independent of one another or more succinctly that the process is memoryless. It can also be shown (Ross, 1996) that 1.1-1.4 imply that $N[t, t + h)$ follows a Poisson distribution:

$$P\{N[t, t+h) = n\} = \frac{\left(\int_t^{t+h} \lambda(u)\mathrm{d}u\right)^n \exp\left\{-\int_t^{t+h} \lambda(u)\mathrm{d}u\right\}}{n!}, \ n = 0, 1, 2, \ldots .$$

Thus $N(t)$ is distributed as a Poisson random variable with mean $\Lambda(t)$. For this reason the cumulative intensity $\Lambda(t)$ is often referred to as the cumulative mean function.

1.3 Spline Smoothing

Suppose there exists data (Y_i, t_i), $i = 1, \ldots, n$ where $l < t_1 < \cdots < t_n < u$ and the Y_i's are independently distributed $N(f(t_i), \sigma^2)$ random variables. The relationship between the Y_i's and the t_i's is then simply expressed as follows

$$\mathbf{Y} = \mathbf{f} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}^T = \{Y_1, \ldots, Y_n\}$, $\mathbf{f}^T = \{f(t_1), \ldots, f(t_n)\}$ and $\boldsymbol{\epsilon}^T = \{\epsilon_1, \ldots, \epsilon_n\} \sim N_n(\mathbf{0}, \mathbf{I}_n\sigma^2)$; $\mathbf{I}_n$ is the $n \times n$ identity matrix. Further assume that $f(\cdot)$ is some unknown smooth function of which an estimate is desired. One simple approach would be to assume that $f \in \mathbf{P}_d$ where

$\mathbf{P}_d$ is the space of polynomials of degree d. The function $f(\cdot)$ can then be expressed as a linear combination of the basis functions $\{1, x, x^2, \ldots, x^d\}$ which span $\mathbf{P}_d$:

$$f(x) = \psi_1 + \psi_2 x + \psi_3 x^3 + \cdots + \psi^{d+1} x^d$$

Estimates of the function $f(\cdot)$ are then easily obtained by finding the values of $\psi^T = \{\psi_1, \ldots, \psi_{m+1}\}$ that minimize $||\mathbf{Y} - \mathbf{f}||^2$ by least squares. Unfortunately when $d << n < \infty$ polynomials are not flexible enough to fit slowly varying functions or those that fail to be extremely smooth which results in poor approximating power for general smooth functions $f(\cdot)$. Polynomials however are known to have very good local approximating properties (Schumaker, 1981) and so it is logical to partition the data into segments and fit different polynomials to each of these subintervals. These piece-wise polynomials can fit a much larger class of functions. Piece-wise polynomials however can be discontinuous at the break points between segments, which will be referred to as knots. As estimation of a smooth function is the goal, discontinuities are unexpected in both the function and its derivatives. To overcome this problem smoothness is imposed by requiring $(d-1)$ derivatives to match at all the knots. The set of all possible piece-wise polynomials with this imposed continuity form the *spline space*.

Definition 1.3.1 *Spline Space*

Let $\xi_o = l < \xi_1 < \cdots < \xi_k < u = \xi_{k+1}$ and an integer $d \geq 1$ be given then

$$\mathbf{S}_{d,\boldsymbol{\xi}} = \mathbf{S}_d(\xi_1, \ldots, \xi_k) = \left\{ f \in C^{d-1}[l, u] : f|_{[\xi_i, \xi_{i+1}]} \in \mathbf{P}_d, i = 0, \ldots, k \right\}$$

where

$$C^r[a, b] = \left\{ f : the\ r^{th}\ derivative\ f^{(r)}\ is\ continuous\ on\ [a, b] \right\}$$

is the space of polynomial splines of degree d with k fixed knots $\xi_1, \ldots, \xi_k$.

As with the polynomial space, the spline space is linear (de Boor, 1978). Bases constructed for this space will have dimension $d + k + 1$. Thus $d + k + 1$ functions are needed to span the space: $d + 1$ basis functions to specify the d-degree polynomial in the first interval and one extra function for each of the other k intervals as they all have d restrictions due to the $d - 1$ continuous derivatives. Two common bases $\{b_1(t), \ldots, b_{d+k+1}(t)\}$ for $\mathbf{S}_{d,\boldsymbol{\xi}}$ are the:

a) Truncated Power Basis

$$\left\{ 1, t, t^2, \ldots, t^d, (t - \xi_1)_+^d, \ldots, (t - \xi_k)_+^d \right\},$$

where $(x)_+ = x\mathbf{1}(x > 0)$ with $\mathbf{1}$ the indicator function. Though a very intuitive basis, this basis suffers from poor numerical conditioning when the knots ξ_i are dense in [l,u].

b) B-spline Basis

$$\left\{ B_{d,-d}(t), B^d_{d,-d+1}(t), \ldots, B_{d,0}(t), B_{d,1}(t), \ldots, B_{d,k}(t) \right\},$$

a numerically stable basis constructed by defining $2d + 2$ exterior knots

$$\xi_{-d} \leq \cdots \leq \xi_0 = l < \xi_1 < \cdots < \xi_k < u = \xi_{k+1} \leq \cdots \leq \xi_{d+k+1}$$

and taking the $(d + 1)^{th}$ divided differences of $(t - \cdot)^d_+$ giving

$$B_{d,h}(t) = (\xi_{h+d+1} - \xi_h)[\xi_h, \ldots, \xi_{h+d+1}](t - \cdot)^d_+$$

where

$$(t - \xi)^d_+[\xi_h, \ldots, \xi_{h+d+1}] = \frac{\det \begin{bmatrix} 1 & \xi_h & \cdots & \xi_h^d & (t - \xi_h)^d_+ \\ 1 & \xi_{h+1} & \cdots & \xi_{h+1}^d & (t - \xi_{h+1})^d_+ \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \xi_{h+d+1} & \cdots & \xi_{h+d+1}^d & (t - \xi_{h+d+1})^d_+ \end{bmatrix}}{\det \begin{bmatrix} 1 & \xi_h & \cdots & \xi_h^d & \xi_h^{d+1} \\ 1 & \xi_{h+1} & \cdots & \xi_{h+1}^d & \xi_{h+1}^{d+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \xi_{h+d+1} & \cdots & \xi_{h+d+1}^d & \xi_{h+d+1}^{d+1} \end{bmatrix}}$$

(Note: strict inequalities are required for the interior knots else the number of continuous derivatives drops by one and hence the function f would no longer be a spline by definition 1.3.1). The numerical stability attributed to B-splines arises from the divided differences since $B_{d,h}(t) = 0$ for $t \notin (\xi_h, \xi_{h+d+1}]$. An equivalent recursive relationship that is numerically efficient and useful in practice (de Boor, 1978) can be derived as follows:

Let $B_{0,0}, \ldots B_{0,k}$ be

$$B_{0,i}(t) = \mathbf{1}\left\{ t \in [\xi_h, \xi_{h+1}) \right\}$$

Then for $d \geq 1$ the basis functions can be calculated recursively as

$$B_{d,h}(t) = \frac{t - \xi_h}{\xi_{h+d} - \xi_h} B_{d-1,h}(t) + \frac{\xi_{h+d+1} - t}{\xi_{h+d+1} - \xi_{h+1}} B_{d-1,h+1}(x).$$

Transforming between these two equivalent bases for $\mathcal{S}_{d,k}$; B-spline basis, useful for numerical computing, and the Truncated Power basis, useful for testing, can be accomplished via a linear transform constructed from the following useful relationships. If $f \in \mathcal{S}_{d,\xi}$ then

$$f(t) = \sum_{i=0}^{d} f^{(i)}(0)t^d/i! + \sum_{j=1}^{k}[f^{(d)}(\xi_j^+) - f^{(d)}(\xi_j^-)](t - \xi_j)_+^d/d! \qquad (1.5)$$

where $f^{(i)}(t) = \partial^i f(t)/\partial t^i$. So if we express f as a B-spline, $f(t) = \sum_{h=1}^{k+4} \psi_h B_{d,h-4}(t)$ then an equivalent representation from (1.5) in Truncated Power form

$$f(t) = \sum_{h=0}^{d} \varphi_{h+1}t^h + \sum_{h=1}^{k} \varphi_{h+d+1}(t - \xi_h)_+^d$$

is obtained via the transformation $\boldsymbol{\varphi} = \mathbf{T}\boldsymbol{\psi}$ where

$$\mathbf{T} = \begin{bmatrix} B_{d,-d}(0) & B_{d,-d+1}(0) & \cdots & B_{d,k}(0) \\ B_{d,-d}^{(1)}(0) & B_{d,-d+1}^{(1)}(0) & \cdots & B_{d,k}^{(1)}(0) \\ B_{d,-d}^{(2)}(0)/2! & B_{d,-d+1}^{(2)}(0)/2! & \cdots & B_{d,k}^{(1)}(0)/2! \\ \vdots & \vdots & \ddots & \vdots \\ B_{d,-d}^{(d)}(0)/d! & B_{d,-d+1}^{(d)}(0)/d! & \cdots & B_{d,k}^{(d)}(0)/d! \\ D_{d,-d}^{(d)}(\xi_1)/d! & D_{d,-d+1}^{(d)}(\xi_1)/d! & \cdots & D_{d,k}^{(d)}(\xi_1)/d! \\ \vdots & \vdots & \ddots & \vdots \\ D_{d,-d}^{(d)}(\xi_k)/d! & D_{d,-d+1}^{(d)}(\xi_k)/d! & \cdots & D_{d,k}^{(d)}(\xi_k)/d! \end{bmatrix} \qquad (1.6)$$

$\boldsymbol{\psi} = [\psi_1, \ldots, \psi_{d+k+1}]^T$, $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_{d+k+1}]^T$, $D_{d,h}^{(i)}(\xi_i) = B_{d,h}^{(i)}(\xi_i^+) - B_{d,h}^{(i)}(\xi_i^-)$ and $B_{d,h}^{(i)}(t) = \partial^i B_{d,h}(t)/\partial t^i$ (Note: The $B_{d,h}^{(d)}(t)$'s are step functions with jumps at the ξ_i's so the $D_{d,h}^{(d)}(\xi_i)$'s are simply the vertical difference in the $B_{d,h}^{(d)}(t)$'s at the knots ξ_i).

Fitting the data (Y_i, t_i) using splines is very simple. Assume that $f \in \mathbf{S}_{d,\xi}$ so that f can be expressed as a linear combination of the basis functions

$$f(t_i) = \sum_{h=1}^{d+k+1} \psi_h b_h(t_i)$$

then $f(\cdot)$ can be estimated by solving

$$\min ||\mathbf{Y} - \mathbf{f}||^2 = \min_{\boldsymbol{\psi} \in \mathbb{R}^{d+k+1}} ||\mathbf{Y} - \mathbf{B}\boldsymbol{\psi}||^2$$

where

$$\mathbf{B} = \begin{bmatrix} b_1(t_1) & \dots & b_{d+k+1}(t_1) \\ \vdots & \ddots & \vdots \\ b_1(t_n) & \dots & b_{k+d+1}(t_n) \end{bmatrix}. \tag{1.7}$$

The problem reduces to least squares estimation so the solution is

$$\hat{\psi} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y} \tag{1.8}$$

and thus the estimated fit given by

$$\hat{\mathbf{f}} = \mathbf{B}\hat{\psi} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y}.$$

This is an attractive feature of using splines as they provide a closed form method of estimation. Unfortunately the problem is complicated by recalling that constructing $\mathbf{B}$ requires specifying the number and placement of knots as well as the degree of the spline. All of these have an affect on the estimated fit and thus requires some consideration. Linear splines form piece-wise linear segments which are not suited for capturing curvature. Splines of degree $d \geq 2$ will give smooth fits to the data. The cubic spline ($d = 3$) is often used as it is sufficiently flexible and higher order splines typically give overly erratic fits. To improve the fit of a spline it is better to increase the number of knots rather than its degree. The number of knots used reflects the amount of smoothing as increasing the number of knots in any region of the domain of f reduces the amount of smoothing that is being applied in that region. The number and position of the knots turns out to have a strong effect on the resulting estimates and so careful attention needs to be taken if one desires a high order of approximation. Two main approaches to this problem have been proposed in the literature. The first approach is to consider the number and position of the knots as variables that need to be estimated or the so-called free-knot approach (Jupp, 1978; Kooperberg et al., 1997; Mao and Zhao, 2003). The second is the roughness penalty approach (Wahba, 1990; Green and Silverman, 1994; Eilers and Marx, 1996; Ruppert et al., 2003) which uses a large number of knots and a penalty term to control the oscillatory behavior of the estimated function. Several hybrid approaches that combine both penalty terms and careful knot selection have also been considered (Schwetlick and Schütze, 1995; Luo and Wahba, 1997; Lindstrom, 1999). In the following two sub-sections the two main approaches will be discussed briefly.

1.3.1 Free Knot Splines

In this approach the knots $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_k]^T$ are considered parameters that require estimation. First assume that the number of knots k is known so the problem becomes

$$\min_{\boldsymbol{\psi}, \boldsymbol{\xi}} ||\mathbf{Y} - \mathbf{B}_{\boldsymbol{\xi}} \boldsymbol{\psi}||^2 \tag{1.9}$$

subject to $\boldsymbol{\xi} \in G_k(l, u)$ where $G_k(l, u) = \left\{ \boldsymbol{\xi} = [\xi_1, \ldots, \xi_k]^T : l < \xi_1 < \cdots < \xi_k < u \right\}$. From (1.8) $\hat{\boldsymbol{\psi}}$ is the unique solutions to the least squares problem in $\boldsymbol{\psi}$ and hence $\boldsymbol{\psi}$ will be profile out of (1.9). This is the so-called variable projection method (Golub and Pereyra, 2003) and results in the following constrained nonlinear least squares problem:

$$\min_{\boldsymbol{\xi} \in G_k(l,u)} ||\mathbf{e}(\boldsymbol{\xi})||^2 = \min_{\boldsymbol{\xi} \in G_k(l,u)} ||\mathbf{Y} - \mathbf{B}_{\boldsymbol{\xi}} \hat{\boldsymbol{\psi}}(\boldsymbol{\xi})||^2 = \min_{\boldsymbol{\xi} \in G_k(l,u)} ||\mathbf{Y} - \mathbf{B}_{\boldsymbol{\xi}} (\mathbf{B}_{\boldsymbol{\xi}}^T \mathbf{B}_{\boldsymbol{\xi}})^{-1} \mathbf{B}_{\boldsymbol{\xi}}^T \mathbf{Y}||^2. \tag{1.10}$$

Unfortunately the objective function in (1.10) is rather irregular which was first noted in Jupp (1978) where he refers to this behavior as "lethargy". This problem is caused by two different properties of the objective (1.10). The first has to do with a symmetry resulting from constraining the knots to lie in the simplex $G_k(l, u)$. To see why this occurs consider the case in which $k = 2$ and take two points $\boldsymbol{\xi}_1 = [\xi_1, \xi_2]^T \in G_2(l, u)$ and $\boldsymbol{\xi}_2 = [\xi_2, \xi_1]^T \notin G_2(l, u)$. Clearly $||\mathbf{e}(\boldsymbol{\xi}_1)||^2 = ||\mathbf{e}(\boldsymbol{\xi}_2)||^2$ and hence the objective surface of (1.10) in $G_2[l, u]$ is a reflection of the one in $(l, u) \times (l, u) \bigcap G_2(l, u)^c$. The result of this is that the gradient of (1.10) at any point on the boundary of $G_2(l, u)$ points along this boundary. So as two knots move closer together the attraction between them becomes strong and draws them onto the boundary where many local optima occur. This is undesirable as the solution may be quite far from the global minimum of (1.10) and will not preserve the assumed level of "smoothness" of the underlying function since coincident knots reduce the number of absolutely continuous derivatives. Also, if the solution is near the boundary this "boundary effect" causes a great deal of instability in optimization algorithms. The second problem is a result of the fact that small perturbations in ξ result in values of $\mathbf{e}(\boldsymbol{\xi})$ that are similar. A direct consequence is that the objective function in (1.10) is not smooth and has many plateaus which makes the solution difficult to obtain using standard methods. Using computationally demanding global optimization methods such as branch and bound can provide the true global solution (Beliakov, 2004). Using such a method the optimal number and position of the knots can be obtained by solving (1.10) for different values of k and selecting the best model using criteria such as the C_p Mallows statistic or AIC.

Another approach is to employ a stochastic type optimization algorithm on a dense grid of possible candidate knots such as using insertion-deletion (Hansen and Kooperberg, 2002) or a genetic algorithm (Pittman, 2002) but neither of these techniques will guarantee a global optima.

The free knot approach results in very flexible estimators that have a high order of approximation (Schumaker, 1981), without requiring too many degrees of freedom, making them very attractive in many situations. In the context of normal linear regression presented here the computational methods described can be implemented fairly efficiently. However extensions of such approaches to more complicated situations such as the recurrent event analysis considered in this thesis may be computationally burdensome.

1.3.2 Penalized Splines

The knot selection problem can be avoided by choosing a maximal set of knots, such as a knot at each data point so that $\xi_i = t_i$, $i = 1, \ldots, n$ or alternatively some dense subset of times. To control the complexity and the inherently ill-posed nature of the resulting estimation problem regularization is used. This regularization takes the form of some assumption on the class of functions one would like to approximate. For instance, if one assumes that $f \in \mathcal{S}_{d,\xi}$ an estimate can be found by minimizing

$$||\mathbf{Y} - \mathbf{B}\boldsymbol{\psi}||^2 + \delta\boldsymbol{\psi}^T\mathbf{P}\boldsymbol{\psi} \tag{1.11}$$

where $\mathbf{B}$ is a design matrix of basis functions (1.7) on the dense knot grid, δ is the smoothing parameter and $\mathbf{P} = \{p_{ij}\}$ is the penalty matrix where $p_{ij} = \int_l^u b_i^{(d-1)}(t)b_j^{(d-1)}(t)dt$. Clearly here the form of the penalty is chosen to regulate the assumed amount of curvature of f. The point that should be noted is that use of regularization has effectively reduced the problem of selecting the number and position of knots to finding an appropriate value of δ. If $\delta \to \infty$ then (1.11) reduces to $d - 2$ degree polynomial regression and if $\delta = 0$ then f becomes an interpolator. For any fixed value of δ the solution to (1.11) is given by $\hat{\boldsymbol{\psi}} = (\mathbf{B}^T\mathbf{B}+\delta\mathbf{P})^{-1}\mathbf{B}^T$ so that

$$\hat{\mathbf{f}} = \mathbf{B}\hat{\boldsymbol{\psi}} = \mathbf{B}(\mathbf{B}^T\mathbf{B} + \delta\mathbf{P})^{-1}\mathbf{B}^T\mathbf{Y} = \mathbf{S}_\delta\mathbf{Y}$$

where $\mathbf{S}_\delta$ is the so-called linear smoother operator (Buja et al., 1989). The trace of $\mathbf{S}_\delta$ is referred to as the effective degrees of freedom (Wahba, 1990) and is conceptually analogous to degrees of freedom in regression, however, it will not be integer valued in general due to

the regularization. For computational convenience the smoothing operator can always be expressed as

$$\mathbf{S}_\delta = \mathbf{A}(\mathbf{I} + \delta \mathbf{\Gamma})^{-1} \mathbf{A}^T$$

by taking the singular value decomposition of $\mathbf{B} = \mathbf{UDV}^T$, forming the matrix $\mathbf{G} = \mathbf{D}^{-1}\mathbf{V}^T\mathbf{PVD}^{-1}$ and taking its eigen decomposition $\mathbf{G} = \mathbf{E}\mathbf{\Gamma}\mathbf{E}^T$ where $\mathbf{U}$ a $n \times (d + k + 1)$ matrix with orthogonal columns, $\mathbf{V}$ a $(k + d + 1) \times (d + k + 1)$ orthogonal matrix, $\mathbf{D} = \text{diag}\{s_1, \ldots, s_{d+k+1}\}$, $\mathbf{E}$ the orthonormal eigenvectors of $\mathbf{G}$, $\mathbf{\Gamma} = \text{diag}\{\gamma_1, \ldots, \gamma_{d+k+1}\}$ its eigenvalues and $\mathbf{A} = \mathbf{UE}$ so that $\mathbf{A}^T\mathbf{A} = \mathbf{I}$.

Selection of the smoothing parameter δ is crucial in the penalized spline approach to smoothing. For a review and assessment of the many proposed approaches such as cross validation or unbiased risk see Gu (2002). One of the most popular approaches is the so-called generalized cross validation (GCV) which selects δ by minimizing

$$n \frac{\mathbf{Y}^T \left[\mathbf{I} - \mathbf{B}(\mathbf{B}^T\mathbf{B} + \delta\mathbf{P})^{-1}\mathbf{B}^T\right]\mathbf{Y}}{(n - \text{tr}\{\mathbf{S}_\delta\})^2}$$

where $\text{tr}\{\mathbf{S}_\delta\} = \sum_{i=1}^{d+k+1} 1/(1 + \delta\gamma_i)$. This was originally proposed by Craven and Wahba (1978). Another approach is to minimize the generalized maximum likelihood

$$\frac{\mathbf{Y}^T \left[\mathbf{I} - \mathbf{B}(\mathbf{B}^T\mathbf{B} + \delta\mathbf{P})^{-1}\mathbf{B}^T\right]\mathbf{Y}}{\prod_{i=1}^{k-1} [1/(1 + \delta\gamma_i)]^{\frac{1}{n-d-2}}} \tag{1.12}$$

which is a restricted maximum likelihood (REML) estimator in the normal linear model context with a Gaussian process prior assumption. For penalized spline smoothing with normally distributed residuals, GCV is preferred, however in the case of non-normal error distributions a REML/GML approach is recommended (Gu, 2002). In (1.12) the last $d - 2$ eigenvalues $\gamma_1 > \gamma_2 > \cdots > \gamma_{d+k+1}$ are equal to zero reflecting the unpenalized $d - 2$ polynomial component.

It should be noted that to be consistent with the rest of the presentation in this section a B-spline approach has been used; however, other approaches to penalized smoothing could also be considered, such as smoothing splines (Green and Silverman, 1994; Gu, 2002), and would result in a similar framework. The difference in practice would be the choice of basis and form of the penalty with smoothing splines being derived via reproducing Hilbert kernel spaces. The estimating procedure that is proposed herein does not depend on the choice of the basis-penalty combination, whether by linear differential arguments (Heckman

and Ramsay, 2000) or otherwise, and would be carried out equivalently independent of this decision.

1.4 Estimating Equations

In statistics, methods such as least-squares, maximum likelihood and minimum chi-squared estimation have a common property that they involve constructing and solving a system of estimating equations:

$$\mathbf{g}(\tilde{\boldsymbol{\theta}}, \mathbf{y}) = [g_1(\tilde{\boldsymbol{\theta}}, \mathbf{y}), \ldots, g_p(\tilde{\boldsymbol{\theta}}, \mathbf{y})]^T = \mathbf{0}$$

where $\mathbf{y} = [y_1, \ldots, y_n]^T$ is observed data with distribution governed by some unknown parameters $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$. Due to this commonality among methods it is logical to consider properties of $\mathbf{g}$, the estimating function

$$\mathbf{g} : \Theta \times \mathcal{Y} \to \mathbb{R}^p$$

where Θ and $\mathcal{Y}$ are the parameter and sample spaces respectively. It will also be assumed that $\det \left\{ E[\mathbf{g}(\boldsymbol{\theta}, \mathbf{Y}) \mathbf{g}(\boldsymbol{\theta}, \mathbf{Y})^T \mid \boldsymbol{\theta}] \right\} < \infty$ for all $\boldsymbol{\theta}$ which ensures that the second moments conditional on $\boldsymbol{\theta}$ of each $g_q(\boldsymbol{\theta}, \mathbf{Y})$, $q = 1, \ldots, p$ are finite. The goal is to determine a estimator $\tilde{\boldsymbol{\theta}}(\mathbf{Y})$ which is the solution to $\mathbf{g}(\tilde{\boldsymbol{\theta}}, \mathbf{Y}) = \mathbf{0}$. It is required that the estimating function be unbiased in the sense that $E[\mathbf{g}(\boldsymbol{\theta}, \mathbf{Y}) \mid \boldsymbol{\theta}] = \mathbf{0}$ for all $\boldsymbol{\theta}$ in the sample space Θ. This extends the notion of unbiasedness since every unbiased estimator with finite variance is the root of an unbiased estimating function (Kendall, 1951; Kendall, 1952). It should be noted however that not every root of an unbiased estimating function is itself unbiased.

An idea that is central to estimating equation theory is that of Godambe efficiency (Godambe, 1960) which is conceptually analogous to the idea of a minimum variance unbiased estimator in that we would like to find an unbiased estimating equation that has the smallest variance. It can be shown (Chandrasekar and Kale, 1984) that the following three conditions are equivalent for determining Godambe efficiency of an estimating function $\mathbf{g}^*$:

1. $\mathrm{eff}_{\boldsymbol{\theta}}(\mathbf{g}^*) - \mathrm{eff}_{\boldsymbol{\theta}}(\mathbf{g})$ is non-negative definite for all $\mathbf{g}$ and $\boldsymbol{\theta}$

2. $\mathrm{trace}\left\{\mathrm{eff}_{\boldsymbol{\theta}}(\mathbf{g})\right\} \leq \mathrm{trace}\left\{\mathrm{eff}_{\boldsymbol{\theta}}(\mathbf{g}^*)\right\}$ for all $\mathbf{g}$ and all $\boldsymbol{\theta}$.

3. $\det\left\{\mathrm{eff}_{\boldsymbol{\theta}}(\mathbf{g})\right\} \leq \det\left\{\mathrm{eff}_{\boldsymbol{\theta}}(\mathbf{g}^*)\right\}$ for all $\mathbf{g}$ and all $\boldsymbol{\theta}$.

where

$$\mathrm{eff}_{\boldsymbol{\theta}}(\mathbf{g}) = E[\partial\mathbf{g}(\boldsymbol{\theta},\mathbf{Y})/\partial\boldsymbol{\theta}^T \mid \boldsymbol{\theta}]^T E\left[\mathbf{g}(\boldsymbol{\theta},\mathbf{Y})\mathbf{g}(\boldsymbol{\theta},\mathbf{Y})^T \mid \boldsymbol{\theta}\right]^{-1} E[\partial\mathbf{g}(\boldsymbol{\theta},\mathbf{Y})/\partial\boldsymbol{\theta}^T \mid \boldsymbol{\theta}].$$

An intuitive interpretation of the Godambe efficiency $\mathrm{eff}_{\boldsymbol{\theta}}(\mathbf{g})$ is as the inverse of the variance-covariance of the standardized estimating function

$$
\begin{aligned}
[\mathrm{eff}_{\boldsymbol{\theta}}(\mathbf{g})]^{-1} &= E\left(\left\{ E[\partial\mathbf{g}(\boldsymbol{\theta},\mathbf{Y})/\partial\boldsymbol{\theta}^T \mid \boldsymbol{\theta}]^{-1}\mathbf{g}(\boldsymbol{\theta},\mathbf{Y}) \right\} \left\{ E[\partial\mathbf{g}(\boldsymbol{\theta},\mathbf{Y})/\partial\boldsymbol{\theta}^T \mid \boldsymbol{\theta}]^{-1}\mathbf{g}(\boldsymbol{\theta},\mathbf{Y}) \right\}^T \Big| \boldsymbol{\theta} \right) \\
&= E[\partial\mathbf{g}(\boldsymbol{\theta},\mathbf{Y})/\partial\boldsymbol{\theta}^T \mid \boldsymbol{\theta}]^{-1} E\left[\mathbf{g}(\boldsymbol{\theta},\mathbf{Y})\mathbf{g}(\boldsymbol{\theta},\mathbf{Y})^T \mid \boldsymbol{\theta}\right] E[\partial\mathbf{g}(\boldsymbol{\theta},\mathbf{Y})^T/\partial\boldsymbol{\theta} \mid \boldsymbol{\theta}]^{-1}(1.13)
\end{aligned}
$$

which is useful for developing asymptotics for the equations.

A special type of estimation function used extensively in this thesis is the so-called quasi-likelihood (Wedderburn, 1974). The quasi-likelihood approach is quite useful in that it only requires partial parametric specification, that both the mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ and variance-covariance $\mathbf{V}(\boldsymbol{\theta})$ are some known functions of $\boldsymbol{\theta}$. Given these limited assumptions and considering the class of unbiased estimating functions of the form

$$\mathbf{A}(\boldsymbol{\theta})\left[\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})\right]$$

the only value of $\mathbf{A}$ that results in a Godambe efficient estimating function is $\mathbf{A}(\boldsymbol{\theta}) = \mathbf{D}(\theta)^T\mathbf{V}(\theta)^{-1}$ (Heyde, 1997) where $\mathbf{D}(\theta) = \partial\boldsymbol{\mu}(\theta)/\partial\boldsymbol{\theta}^T$. An estimator can then be found by solving

$$\mathbf{D}(\theta)^T\mathbf{V}(\theta)^{-1}\left[\mathbf{Y} - \boldsymbol{\mu}(\theta)\right] = \mathbf{0}.$$

It can be show that $\sqrt{n}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to N(\mathbf{0},\boldsymbol{\Sigma})$ (White, 1982) where $\boldsymbol{\Sigma}$ is given in (1.13) providing a framework for inference robust against model misspecification as only low moment assumptions are required.

1.5 Continuous and Discrete Mixture Models

The term *mixture distribution* refers to a distribution arising from a hierarchical structure. More clearly, a random variable Y is said to have a mixture distribution if the distribution of Y depends on a quantity that itself has a distribution. So if $Y|\Theta \sim H(\Theta)$ and $\Theta \sim M$ then Y has a mixture distribution with density

$$f(y) = \int h(y;\theta)\, \mathrm{d}M(\theta) \tag{1.14}$$

where M is the mixing distribution and h denotes the probability density or probability mass function of the distribution H. If the mixing distribution M is continuous then (1.14) becomes

$$f(y) = \int h(y; \theta) m(\theta) \, d\theta$$

where m is the probability density function of M. As an example, consider the case where $Y|\Theta = \theta$ is distributed as a Poisson random variable with mean θ and Θ is a gamma distribution depending on parameters τ and μ with density

$$m(\theta; \tau) = \frac{1}{\Gamma(\frac{1}{\tau})(\tau\mu)^{\frac{1}{\tau}}} \theta^{\frac{1}{\tau}-1} \exp\left\{-\frac{\theta}{\tau\mu}\right\} \mathbf{1}(\theta \geq 0, \tau > 0, \mu > 0).$$

The unconditional distribution of Y in this case becomes negative binomial

$$f(y; \mu, \tau) = \frac{\Gamma(y + \tau^{-1})}{(y!)\Gamma(\tau^{-1})} \left(\frac{\tau\mu}{1+\tau\mu}\right)^y \left(\frac{1}{1+\tau\mu}\right)^{\tau^{-1}} \tag{1.15}$$

$y = 0, 1, \ldots$ with mean μ and variance $\mu + \tau\mu^2$. The negative binomial is a common method of modeling heterogeneity in count data. It occurs when the standard Poisson assumption cannot accommodate the variance exhibited by the data and is commonly referred to as extra-Poisson variation or overdispersion (Dean, 1998). If overdispersion is ignored the resulting analysis can be misleading as standard errors can be severely underestimated. Given a random sample $Y_1, \ldots, Y_n$ from (1.15) with covariate information $\mathbf{x}_i$, related to the means through a monotone function g such that $\mu_i = g(\mathbf{x}_i^T \beta)$, β the covariate effects, one obtains a regression model for analyzing overdispersed count data (Lawless, 1987b). This model belongs to a class of models arising from continuous mixture distributions referred to as mixed models (Demidenko, 2004). Such models are prevalent throughout the discipline of statistics as they provide an intuitive framework for the incorporation of correlation structures into standard methods such as generalized linear models (McCullagh and Nelder, 1999) that assume independence.

In the case where the mixing distribution M is discrete, say a discrete distribution with c points of support such that $p_j = P(\Theta = \theta_j)$, $j = 1, \ldots, c$ then (1.14) is given by

$$f(y) = \sum_{j=1}^{c} p_j h(y; \theta_j). \tag{1.16}$$

Such a mixing distribution is referred to as a finite mixture distribution. Finite mixture distributions are used primarily for modeling unobserved clusters and have many practical

applications (McLachlan and Peel, 2000). To illustrate, following the example given in the continuous mixture case, assume that $Y_1, \ldots, Y_n$ are a random sample with probability mass function (1.16) where $Y_i | \Theta = \theta_j$ are assumed to be independent Poisson random variables with mean θ_j. The log likelihood function can then be expressed as

$$l(\boldsymbol{\theta}, \mathbf{p}; \mathbf{y}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{c} p_j \frac{\theta_j^{y_i} \exp\{-\theta_j\}}{y_i!} \right) \tag{1.17}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_c]^T$ and $\mathbf{p} = [p_1, \ldots, p_c]^T$. The interpretation here is that each Y_i belongs to one of c distinct Poisson distributions with some specific probability. Maximizing (1.17) over $\boldsymbol{\theta}$ and $\mathbf{p}$ is not a simple optimization problem as the log likelihood is not additive. Estimation would be trivial if the cluster to which each observation belonged was known since in this case the likelihood would have a simple multinomial form

$$\prod_{i=1}^{n} \prod_{j=1}^{c} \left(p_j \frac{\theta_j^{y_i} \exp\{-\theta_j\}}{y_i!} \right)^{z_{ij}}$$

where $z_{ij} = \mathbf{1}(\Theta = \theta_j)$ and $\sum_{j=1}^{c} z_{ij} = 1$ so that the log likelihood becomes proportional to

$$\sum_{i=1}^{n} \sum_{j=1}^{c} z_{ij} \left[\log(p_j) + y_i \log(\theta_j) - \theta_j \right].$$

The maximum likelihood estimates given the z_{ij}'s have closed form solutions

$$\begin{aligned} \hat{\theta}_j &= \frac{\sum_{i=1}^{n} z_{ij} \, y_i}{\sum_{i=1}^{n} z_{ij}} \\ \hat{p}_j &= \frac{\sum_{i=1}^{n} z_{ij}}{n} \end{aligned} \tag{1.18}$$

in this simple case.

More generally the z_{ij}'s are unknown and need to be imputed. This can be accomplished by computing the posterior expectation of cluster membership via Bayes rule

$$z_{ij}^* = \frac{p_j \, \theta_j^{y_i} \exp\{-\theta_j\}}{\sum_{j=1}^{c} p_j \, \theta_j^{y_i} \exp\{-\theta_j\}}. \tag{1.19}$$

Estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{p}}$ that maximize (1.17) can then obtained by iteratively updating (1.18) and (1.19) until convergence which is the so-called Expectation-Maximization (EM) algorithm (McLachlan and Krishnan, 1997). Although the EM algorithm is simple in this case it still

requires specifying the number of components c which requires some consideration. Further, finding the true global maximum of (1.17) is not trivial for a fixed value of c due to the presence of multiple local maxima (Böhning, 2003) and special modifications to the basic EM algorithm need to be made to handle this problem. The EM algorithm can also be fairly slow to converge compared to direct Newton type methods which converge quadratically to the nearest local maximum (Nocedal and Wright, 1999).

1.6 Outline of Thesis

This thesis consists of three projects. Each of Chapters 2, 3 and 4 is written in a style similar to that for publication. As a result some introductory material is repeated as well as the description of a motivating data set.

1.6.1 Chapter 2

In this Chapter we consider a semiparametric model for the analyses of longitudinal studies where data are collected as panel counts. The model is a nonhomogeneous Poisson process with a multiplicative intensity incorporating covariates through a proportionality assumption. Heterogeneity is accounted for in the model through subject-specific random effects. The key feature of the model is the use of regression splines to model the distribution of recurrences over time. This provides a flexible and robust method of relaxing parametric assumptions. In addition, quasi-likelihood methods are proposed for estimation, requiring only first and second moment assumptions to obtain consistent estimates. Simulations demonstrate that the method produces estimators and standard errors of the intensity with low bias and whose distributions are well-approximated by the normal. The usefulness of this approach, especially as an exploratory tool, is illustrated by analyzing a study designed to assess the effectiveness of a pheromone treatment in disturbing the mating habits of the Cherry Bark Tortrix moth.

1.6.2 Chapter 3

An adaptive semiparametric model for analyzing longitudinal panel count data extending the model of Chapter 2 is discussed. The counts are assumed to arise from a mixed nonhomogenous Poisson process where frailties account for heterogeneity common to this type of

data. The generating intensity of the counting process is assumed to be a smooth function modeled with penalized splines. A main feature is that the penalization used to control the amount of smoothing, usually assumed to be time homogeneous, is allowed to be time dependent so that the spline can more easily adapt to sharp changes in curvature regimes. Penalized splines are also used to model covariate effects relaxing the proportional intensity assumption. Penalized quasi-likelihood (PQL; Breslow and Clayton (1993)) is used to derive estimating equations for this adaptive spline model so that only low moment assumptions are required for inference. Both jackknife and bootstrap variance estimators are developed. The finite sample properties of the proposed estimating functions are investigated empirically by simulation. Comparisons with a model assuming a time homogeneous penalty are made. The methods are used in an analysis of data from an experiment to test the effectiveness of pheromones in disrupting the mating pattern of the cherry bark tortrix moth. Recommendations are provided on when the simpler model with a time homogeneous penalty may provide a fair approximation to data and where such an approach will be lacking, calling for the more complicated adaptive methods.

1.6.3 Chapter 4

A flexible semiparametric model for analyzing longitudinal panel count data arising from mixtures is presented. The model assumes that the counts for each subject are generated by mixtures of nonhomogeneous Poisson processes with smooth intensity functions modeled with penalized splines. Time dependent covariate effects are also incorporated into the process intensity using splines. Discrete mixtures of these nonhomogeneous Poisson process spline models extract functional information from underlying clusters representing hidden subpopulations. The motivating application is an experiment to test the effectiveness of pheromones in disrupting the mating pattern of the cherry bark tortrix moth. Mature moths arise from hidden, but distinct, subpopulations and monitoring the subpopulation responses was of interest. Within-cluster random effects are used to account for correlation structures and heterogeneity common to this type of data. An estimating equation approach to inference requiring only low moment assumptions is developed and the finite sample properties of the proposed estimating functions are investigated empirically by simulation. The method is also illustrated on several additional examples.

1.6.4 Chapter 5

The thesis closes with a discussion of topics for future exploration.

Chapter 2

Mixed NHPP Regression Spline Model

Recurrent events arise where subjects experience multiple occurrences of an event of interest. In many situations, and especially for chronic diseases, continuous followup providing the times of occurrence of events is either too invasive or not possible and repeated counts are collected as the number of events which have occurred within specific time periods. Such longitudinal count data is referred to as panel data. For each individual these counts are stochastically ordered and typically correlated. In addition, they are subject to censoring, where data may be missing within specific time periods or panels, or data within consecutive panels may be aggregated. Also, the times of followup for different individuals may not be coincident, and are often subject-specific.

Parametric regression methods for analyzing longitudinal data have been developed by Lawless (1987a), Lawless and Nadeau (1995), Pepe and Cai (1993), Jiang et al. (1999), and Thall (1988), for example. Lawless (1987a) provides a thorough analysis of mixed Poisson process data using a gamma-frailty and a Weibull rate. Jiang et al. (1999) use a similar approach and account for overdispersion and measurement error through the use of robust variance estimates instead of the gamma frailty. Thall (1988) provides a likelihood analysis of panel data using gamma subject-specific effects and with the form of the rate being a specified function of time. Although parametric methods are very useful, especially in that they can accommodate all the features of recurrent event data mentioned above, they require parametric specification of the rate over time. This can be difficult in practice

unless the process under investigation is well-understood. Pepe and Cai (1993) and Lawless and Nadeau (1995) instead model the mean and rate functions for recurrent event data and provide methods for regression analyses. The theory behind their developments is refined in a more recent paper by Lin et al. (2000).

For the analysis of panel data, Staniswalis et al. (1997) use a semiparametric estimator of the rate function, keeping a gamma frailty for random effects. They use the generalized profile likelihood method of Severini and Wong (1992). Lawless and Zhan (1998) and Balshaw and Dean (2002) also use a semiparametric estimator for the rate function, modeling the rate as a piece-wise constant. Sun and Wei (2000) and Hu et al. (2003) model the mean function of the cumulative number of recurrent events. Importantly, Sun and Wei consider the situation where both observation and censoring times depend on covariates, and develop estimating equations for regression parameters. Balshaw and Dean (2002) use quasi-likelihood methods for estimation so only mean and variance assumptions regarding the subject-specific random effects are required. Lawless and Zhan (1998) use estimating equations based on a Poisson process, and account for heterogeneity using robust variance estimates.

Non-parametric estimates of the rate function derived from methods above generate a step-function estimator. Here we discuss the semiparametric analysis of recurrent event panel data using regression splines to model the rate function. Regression splines are a natural extension of the piece-wise constant models described above and have been used recently in survival analysis using the Cox model. For example, Sleeper and Harrington (1990) model covariate effects using fixed-knot regression splines, Rosenberg (1995) applied cubic regression splines for estimating the baseline hazard function and Kooperberg et al. (1995) use linear splines and their tensor products to estimate one or more covariate effects in the log-hazard function. More recently, Molinari et al. (2001) use low order regression splines and one or two free knots, which are estimated, for determining thresholds in the covariate space. Through the use of splines, we retain here simultaneously the advantage of non-parametric estimation, in that a flexible, non-parametric form for the rate is assumed, and the advantage of parametric estimators, in that a continuous estimate of the rate is obtained.

The main contributions of the chapter are two-fold: the use of the flexible, robust and efficient combination of spline smoothing for the rate function along with the retention of

simple quasi-likelihood methods for estimation, the latter achieved through the parameter-
ization adopted and through straightforward adaptations of usual estimating equations for
the analysis of recurrent event data. The model proposed herein is similar to the propor-
tional intensity overdispersed Poisson parametric model except that the baseline is estimated
via regression splines. In addition, distributional assumptions concerning random effects are
relaxed to second order moment assumptions and quasi-likelihood estimation is adopted for
covariate effects. Our objective is to provide a tool that is simple to use yet able to provide
smooth estimates and reliable inference for panel data exhibiting all of the features of re-
current event data mentioned above. We show that the method requires simple adaptations
of quasi-likelihood regression functions conventionally used for count data analysis.

In Section 2.1, quasi-likelihood estimation and inference using regression splines is de-
scribed. An analysis of an entomological experiment which motivated the development of
these methods is presented in Section 2.2. In the example, which considers the mating ac-
tivities of the Cherry Bark Tortrix moth, the rate function is of prime scientific significance.
None of the usual parametric models provide a reasonable approximation since the rate is bi-
modal. We have investigated the accuracy of the proposed spline estimators of the rate and
cumulative rate functions and the coverage of point-wise confidence intervals, and results of
our studies are presented in this section. The results show that, unless there is substantial
local curvature, the regression splines are able to recover a variety of smooth shapes typical
of rate functions, and the proposed quasi-likelihood estimators perform well with moderate
sample sizes. The chapter concludes with a discussion of the proposed method.

2.1   Semiparametric Analysis of Recurrent Event Panel Data with Regression Splines

Let $N_1(t), N_2(t), \ldots, N_I(t)$ be independent random variables where $N_i(t)$ is a counting pro-
cess denoting the number of events experienced by the ith individual during $(T_{io}, T_{ie_i}]$. Given
subject specific random effects ν_i, the conditional intensity, $\lambda_i(t)$, governing each $N_i(t)$ is
assumed to be multiplicative and is of the form

$$\lambda_i(t) = \nu_i \lambda_o^*(t) e^{\beta_o + \mathbf{x}_i^T \boldsymbol{\beta}}$$

where $\lambda_o(t) = e^{\beta_o} \lambda_o^*(t)$ is the baseline intensity, $\boldsymbol{\beta}$ are regression parameters, and $\mathbf{x}_i$ is a
vector of covariates for individual i. Conditional on ν_i, $N_i(t)$ follows a Poisson process. The

random variables ν_i account for heterogeneity; $\nu_i > 0$ is assumed to follow a distribution with density $p(\nu_i; \tau)$; $E(\mathcal{V}_i) = 1$ without loss of generality and variance $V(\mathcal{V}_i) = e^\tau$, a measure of the overdispersion.

In many studies it is either too costly, overly intrusive or simply impossible to continuously monitor individuals in the study. Instead, individuals i are examined at a set of follow-up times, $T_{i1} < T_{i2} < T_{i3} < \cdots < T_{ie_i}$ where $T_{ij} \in (T_{io}, T_{ie_i}]$, $j = 1, 2, \ldots, e_i$ and the number of occurrences, N_{ij}, within each of these follow-up periods $(T_{i(j-1)}, T_{ij}]$ is recorded. Data collected in this fashion is said to have a panel structure. A likelihood based on these panel counts $\{N_{ij} = N_i(T_{i(j-1)}, T_{ij})\}$, $j = 1, \ldots, e_i$ has the form

$$\mathcal{L} = \prod_{i=1}^{I} \left\{ \binom{N_i}{N_{i1}, \ldots, N_{ie_i}} \prod_{j=1}^{e_i} \left(\frac{\mu_{ij}}{\mu_i}\right)^{N_{ij}} \right\} \int_0^\infty [\nu_i \mu_i]^{N_i} \exp\left\{-\nu_i \mu_i\right\} p(\nu_i; \tau) \mathrm{d}\nu_i \qquad (2.1)$$

where $\mu_{ij} = E[N_{ij}] = [\Lambda_o(T_{ij}) - \Lambda_o(T_{i(j-1)})]e^{\mathbf{x}_i^T \boldsymbol{\beta}}$, $\Lambda_o(t) = \int_0^t \lambda_o(u)\mathrm{d}u$, the cumulative baseline rate, and $\mu_i = E[N_i] = \sum_{j=1}^{e_i} \mu_{ij} = \Lambda_o(T_{ie_i})e^{\mathbf{x}_i^T \boldsymbol{\beta}}$. This likelihood is a product of two terms (Lawless, 1987a): the first is the conditional kernel of the panel counts N_{ij} given N_i, the total number of events observed for the ith individual, which follows a multinomial distribution; the second term is the mixed Poisson kernel for the total number of events N_i which, for example, is negative binomial under a gamma frailty distribution for ν_i. If event times are recorded, then the likelihood (2.1) can easily be modified to reflect this by simply changing the form of the first term, the conditional kernel, so that it becomes the distribution of order statistics from a sample of size N_i from a distribution with density $\lambda_o(t)/\Lambda_o(T_{ie_i})$, $0 \leq t \leq T_{ie_i}$. If data are unavailable in certain panels so the followup period for the ith individual is not continuous through $(T_{io}, T_{ie_i}]$ but a union of disjoint intervals, then μ_i above would be modified accordingly and the first term would become a product over the panels for which data are recorded. Finally, time dependent covariates are straightforward to handle under certain simplifying assumptions on the nature of the covariate process; in this case, the representation of the likelihood as a product of mixed Poisson distributions over panels and individuals (c.f. Dean, 1991) is more convenient.

It is common to assume a Weibull parametric form (Lawless, 1987a; Dean and Balshaw, 1997), for the baseline rate so $\lambda_o^*(t) = \alpha t^{\alpha-1}$. The Weibull admits a variety of monotone increasing and decreasing shapes for the rate as well as the constant case corresponding to a homogeneous process. Although the Weibull is indeed a flexible distribution there may be situations, for example when the rate is bimodal as in the illustration considered later, where

it does not provide an adequate fit. In this situation, Lawless and Zhan (1998) and Balshaw and Dean (2002) suggest modeling the rate as piece-wise constant. To provide a smooth yet highly flexible estimator of the rate, we suggest here that splines offer a very convenient alternative. Splines are piece-wise polynomials of degree d with continuity imposed at the k specified joining points, referred to as knots. They form linear spaces spanned by $d + k + 1$ basis functions (Schumaker, 1981). A convenient basis for this space is the B-spline basis whose elements are easily computed via a recursive relationship (de Boor, 1978). Choice of the spline order and the sequence of knots will complete the specification of the model. Cubic splines ensure continuity of the cumulative rate and its derivatives and are often sufficiently flexible to handle a wide variety of shapes. Because of the inherent smoothness of most cumulative rate functions, knot placement is not usually crucial for modeling the cumulative rate and a small number of interior knots should suffice. These should be placed so as to ensure a moderately large number of counts between knots and more knots where steep changes are detected in the empirical cumulative rate function.

The cumulative baseline rate is then modeled as

$$\Lambda_o(t) = \exp\{\beta_o + \gamma(t)\} = \exp\left\{\beta_o + \sum_{h=1}^{3+k} \psi_h B_{h-3}(t)\right\}$$

$\gamma(t)$ being an intercept-free cubic B-spline where the $B_{h-3}(t)$'s are the B-spline basis functions and the ψ_h's are the B-spline coefficients. This parameterization is convenient because the logarithm of the cumulative rate and therefore the mean function for the ith individual is linear in both the covariate and spline parameters. Hence, with a small number of knots, estimation can proceed in a traditional fashion using maximum likelihood based on (2.1) or more robust quasi-likelihood methods. Quasi-likelihood will be adopted here as it requires few moment assumptions, has been shown to perform well (Lawless, 1987b; McCullagh and Nelder, 1999) and is commonly used. Implementation of such regression splines in usual statistical packages can easily be incorporated into the regression component of the model, as will be seen shortly.

Note that the likelihood factorization in (2.1) can be written as $\mathcal{L}_1(\psi)\mathcal{L}_2(\beta, \psi, \tau)$, where

$$\mathcal{L}_1 = \prod_{i=1}^{I} \left\{ \binom{N_i}{N_{i1}, \ldots, N_{ie_i}} \prod_{j=1}^{e_i} \left(\frac{\mu_{ij}}{\mu_i}\right)^{N_{ij}} \right\},$$

and

$$\mathcal{L}_2 = \int_0^\infty [\nu_i \mu_i]^{N_i} \exp\{-\nu_i \mu_i\} p(\nu_i; \tau) d\nu_i.$$

All the information about β and τ is provided in the total counts $N_i, i = 1, \ldots, I$, i.e., in $\mathcal{L}_2$. Usual quasi-likelihood estimating equations for estimating the covariate effects β based on the number of events observed over an interval may therefore be employed. These take the form $\mathbf{D}_\beta^T \mathbf{V}^{-1}(\mathbf{N} - \boldsymbol{\mu})$ where $\mathbf{N} = \{N_i\}_{I \times 1}$, $\boldsymbol{\mu} = \{\mu_i\}_{I \times 1}$, the end-of-followup counts and their means respectively, $\mathbf{V} = \text{Var}(\mathbf{N}) = \text{diag}\{\mu_i(1 + e^\tau \mu_i)\}_{I \times I}$ and $\mathbf{D}_\beta = \partial \boldsymbol{\mu} / \partial \boldsymbol{\beta}^T$. Since the logarithm of the cumulative rate is modeled as a cubic B-spline, the mean μ_i can be written $\mu_i = \exp\left\{\beta_o + \sum_{h=1}^{3+k} \psi_h B_{h-3}(t) + \mathbf{x}_i^T \boldsymbol{\beta}\right\}$ and the logarithm of μ_i is therefore linear in the spline parameters ψ_j, an advantageous feature of the use of B-splines for modeling the cumulative means. Based only on $\mathcal{L}_2$, the quasi-likelihood estimating equation for ψ would then similarly take the form $\mathbf{D}_\psi^T \mathbf{V}^{-1}(\mathbf{N} - \boldsymbol{\mu})$ where $\mathbf{D}_\psi = \partial \boldsymbol{\mu} / \partial \boldsymbol{\psi}^T$. This is the basis for our quasi-likelihood approach; for the estimation of ψ we add the component $\partial \log \mathcal{L}_1 / \partial \psi$ which arises from the conditional kernel of the panel counts given $\mathbf{N}$, and does not depend on the distribution of random effects ν_i. Hence estimation of β and ψ requires assumptions only on the mean and variance of the ν_i's. The quasi-likelihood estimating equations for β and ψ simplify to

$$g_\beta = \mathbf{X}^T \mathbf{Q}^{-1}(\mathbf{N} - \boldsymbol{\mu}) \tag{2.2}$$

$$g_\psi = \mathbf{r} + \mathbf{B}^T \mathbf{Q}^{-1}(\mathbf{N} - \boldsymbol{\mu}) \tag{2.3}$$

where $\mathbf{Q} = \text{diag}\{1 + e^\tau \mu_i\}_{I \times I}$; $\mathbf{X} = \{x_{ij}\}_{I \times p}$, the design matrix; $\mathbf{B} = \{B_{ij}\}_{I \times (3+k)}$, $B_{ij} = B_{j-3}(T_{ie_i})$, the elements of the B-spline design matrix; and $\mathbf{r} = \partial \log \mathcal{L}_1 / \partial \psi = \{r_h\}_{(3+k) \times 1}$,

$$r_h = \sum_{i=1}^{I} \sum_{j=1}^{e_i} N_{ij} \left\{ \frac{B_{h-3}(T_{ij}) e^{\gamma(T_{ij})} - B_{h-3}(T_{i(j-1)}) e^{\gamma(T_{i(j-1)})}}{e^{\gamma(T_{ij})} - e^{\gamma(T_{i(j-1)})}} - B_{h-3}(T_{ie_i}) \right\}.$$

The estimating equations for β (2.2) and ψ (2.3) are also the maximum likelihood estimating equations under a gamma frailty. However they are used here without the assumption of a gamma frailty and so standard errors will be different from those obtained under such an assumption. For estimating τ, the pseudo-likelihood equation proposed by Davidian and Carroll (1987) may be used:

$$g_\tau = \sum_{i=1}^{I} \frac{(N_i - \mu_i)^2 - (1 - h_i)\mu_i(1 + e^\tau \mu_i)}{(1 + e^\tau \mu_i)^2} e^\tau. \tag{2.4}$$

Here h_i is the ith diagonal of the hat matrix, $\mathbf{W}^{\frac{1}{2}}\mathbf{C}(\mathbf{C}^T\mathbf{W}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{W}^{\frac{1}{2}}$ where $\mathbf{W} = \mathrm{diag}\{\mu_i/(1+$
$\mathrm{e}^\tau\mu_i)\}_{I\times I}$ and $\mathbf{C} = [\mathbf{X}\ \mathbf{B}]_{I\times(p+k+3)}$; representing a small sample correction. This equation
is derived under the assumption of normally distributed residuals $N_i - \mu_i$. The estimat-
ing equations for β and ψ are asymptotically independent of that chosen for τ. Jørgensen
and Knudsen (2004) terms this type of combination of estimating functions as 'nuisance
parameter insensitive'. This has several important properties in common with conventional
parameter orthogonality, such as the nuisance parameter causing no loss of efficiency for
estimating the parameters of interest, and a simplified estimation algorithm. Alternative
estimating equations for τ may also be employed. For example, Breslow (1984) suggests
setting the Pearson statistic equal to its degree of freedom.

The quasi-likelihood estimates $\tilde{\theta}$ are obtained by solving the system of equations $\boldsymbol{g_\theta} = \mathbf{0}$
where $\boldsymbol{g_\theta} = [\boldsymbol{g_\beta}^T, \boldsymbol{g_\psi}^T, g_\tau]^T$. It can be shown (White, 1982) that as $I \to \infty$ under mild
regularity conditions $\sqrt{I}(\tilde{\theta} - \theta)$ has a multivariate normal distribution with mean $\mathbf{0}$ and
covariance

$$E\left[-\lim_{I\to\infty}\frac{1}{I}\frac{\partial \boldsymbol{g_\theta}}{\partial\theta^T}\right]^{-1} E\left[\lim_{I\to\infty}\frac{1}{I}\boldsymbol{g_\theta}\boldsymbol{g_\theta}^T\right] E\left[-\lim_{I\to\infty}\frac{1}{I}\frac{\partial \boldsymbol{g_\theta}^T}{\partial\theta}\right]^{-1}.$$

Finite sample empirical variance estimates are obtained by substituting θ with $\tilde{\theta}$ and omit-
ting the limits so that

$$\widetilde{\mathrm{Var}}(\tilde{\theta}) = E\left[-\frac{\partial \boldsymbol{g_\theta}}{\partial\theta^T}\right]^{-1} E\left[\boldsymbol{g_\theta}\boldsymbol{g_\theta}^T\right] E\left[-\frac{\partial \boldsymbol{g_\theta}^T}{\partial\theta}\right]^{-1}\Bigg|_{\theta=\tilde{\theta}}.$$

This covariance is generally referred to as the sandwich variance estimator (Liang and
Zeger, 1986) and protects against variance misspecification as well as reduces the number
of moments required for inference. The matrix of expected negative derivatives of the
estimating equations $\boldsymbol{g_\theta}$ is

$$E\left[-\frac{\partial \boldsymbol{g_\theta}}{\partial\theta^T}\right] = \begin{bmatrix} \mathbf{X}^T\mathbf{W}\mathbf{X} & \mathbf{X}^T\mathbf{W}\mathbf{B} & \mathbf{0} \\ \mathbf{B}^T\mathbf{W}\mathbf{X} & \mathbf{R}+\mathbf{B}^T\mathbf{W}\mathbf{B} & \mathbf{0} \\ \mathbf{z}^T\mathbf{X} & \mathbf{z}^T\mathbf{B} & E\left[-\frac{\partial g_\tau}{\partial\tau}\right] \end{bmatrix},$$

partitioned conforming to the partitioning of $\boldsymbol{g_\theta}$, with

$$\mathbf{W} = \mathrm{diag}\left\{\frac{\mu_i}{1+\mathrm{e}^\tau\mu_i}\right\}_{I\times I};$$

$$\mathbf{z} = \left\{\frac{1+2\mathrm{e}^\tau\mu_i}{1+\mathrm{e}^\tau\mu_i}\cdot\mathrm{e}^\tau\right\}_{I\times 1};$$

$$E\left[-\frac{\partial g_\tau}{\partial \tau}\right] = \sum_{i=1}^{I}\left(\frac{e^\tau \mu_i}{1+e^\tau \mu_i}\right)^2;$$

and

$$\boldsymbol{R} = \{R_{hu}\}_{(3+k)\times(3+k)},$$

$$R_{hu} = \left\{\sum_{i=1}^{I}\sum_{j=1}^{e_i} e^{\mathbf{x}_i^T \boldsymbol{\beta}}\left[B_{h-3}(T_{ij})B_{u-3}(T_{ij})e^{\gamma(T_{ij})} - B_{h-3}(T_{i(j-1)})B_{u-3}(T_{i(j-1)})e^{\gamma(T_{i(j-1)})}\right.\right.$$

$$\left.\left.- \frac{\left[B_{h-3}(T_{ij})e^{\gamma(T_{ij})} - B_{h-3}(T_{i(j-1)})e^{\gamma(T_{i(j-1)})}\right]\left[B_{u-3}(T_{ij})e^{\gamma(T_{ij})} - B_{u-3}(T_{i(j-1)})e^{\gamma(T_{i(j-1)})}\right]}{e^{\gamma(T_{ij})} - e^{\gamma(T_{i(j-1)})}}\right]\right\}.$$

A test for a specific parametric form can easily be constructed using a likelihood ratio (LR) statistic based on the conditional multinomial kernel of (2.1). Similarly a likelihood ratio test, once again based on the conditional multinomial kernel of (2.1) similar to that of Lawless (1987a), can be constructed to test the proportional intensity assumption. This assumption can also be checked heuristically by fitting the model to different strata in the sample separately and plotting the resulting cumulative baseline rates. Finally, a quasi-score test for overdispersion (Balshaw and Dean, 2002) can be constructed and has the following asymptotically standard normal test statistic, $S_o = \sum_{i=1}^{I}\left[(n_i - \tilde{\mu}_i^*)^2 - \tilde{\mu}_i^*\right]\Big/\sqrt{2\sum_{i=1}^{I}\tilde{\mu}_i^{*2}}$ where $\tilde{\mu}_i^* = \exp\left\{\tilde{\beta}_o^* + \sum_{h=1}^{3+k}\tilde{\psi}_h^* B_{h-3}(t) + \mathbf{x}_i^T\tilde{\boldsymbol{\beta}}^*\right\}$, asterisks denoting quasi-likelihood estimates under the limiting case hypothesis of no overdispersion with $V(\mathcal{V}_i) = 0 \ \forall i$, i.e. a nonhomogeneous Poisson process model. In the next section we illustrate the flexibility and efficiency of the proposed methods.

2.2 Illustration and Simulation

2.2.1 Cherry Bark Tortrix Moth Study

This experiment was designed to test the effectiveness of pheromones in disrupting the mating patterns of the Cherry Bark Tortrix Moth (*Enarmonia fomosana*). The pheromone in question had been shown to be competitive with caged virgin females in luring males into traps. It was hypothesized that the release of the pheromone would confuse mate seeking males from locating females in the trees. To test this hypothesis, 20 cherry trees were outfitted with pheromone-baited traps, attached in similar locations in each tree. All

of the trees were fitted with scent dispensers, but ten were selected at random and their dispensers were filled with female pheromones (treatment trees), the remaining ten were used as controls. Males were attracted to the female bait in all of the traps but were frustrated in their attempts to find the traps in the treatment trees. Approximately once a week the traps were emptied and the number of moths caught were counted. Thus the data arise in a panel structure of 19 followup intervals generating approximately weekly counts over the followup period of 18 weeks. At three week intervals the baits were refreshed.

To analyze this data the semiparametric quasi-likelihood regression spline model was fit, as well as the fully parametric quasi-likelihood Weibull model (Dean and Balshaw, 1997) for comparison. A non-parametric estimator of the cumulative rate function (Balshaw and Dean, 2002) was computed to determine knot sequencing. Knots were placed at 9.0, 36.0, 54.0, 72.0, 90.0, and 119.0, where time is in days. Two covariates were included: the first to represent the treatment, and the second being a time-dependent covariate reflecting the age of the bait. The regression spline and Weibull models both result in identical values of the treatment effect $\tilde{\beta}_1$, 3.68 (s.e. 0.354), bait refreshing effect $\tilde{\beta}_2$, -0.191 (s.e 0.0696) reflecting the decreasing potency of the pheromone through time and the overdispersion parameter $\tilde{\tau}$, -0.473 (s.e. 0.427). For 'balanced' designs as observed here with no dropouts and with identical scheduled followup times for each individual, Dean and Balshaw (1997) show that with a proportional rate model and regardless of the form of the baseline (Weibull, or spline, for example), the estimate of the treatment effect is the logarithm of the ratio of the mean number of events observed over all individuals in the treatment group to that in the control group. The quasi-score test for overdispersion results in a test statistic with observed significance level $p < 0.0001$ implying that heterogeneity is present; the proportional rates assumption is supported by its LR test with $p > 0.4$ (9 d.f.). Although it is interesting to note the high effectiveness of the pheromone, the shape of the rate function is of scientific interest here. It depicts the wave or cycle of adult moth prevalence over the summer and is discussed further below. We focus on estimation of the rate noting that the LR test for a Weibull form yields an observed significance level $p < 0.001$ (8 d.f.) indicating severe inappropriateness of that form.

Figure 2.1 plots the estimated baseline rates and cumulative baseline rates along with their associated 95% point-wise confidence bands (obtained using the delta method) for both the regression spline and Weibull models. The baseline rates, $\lambda_o(t) = \partial \Lambda_o(t)/\partial t$ are easily computed using the parameterization offered here as the time dependent components are the

B-spline basis functions $B_h(t)$'s which are polynomials of degree 3. The fitted Weibull model suggests a constant rate over time, $\tilde{\alpha} = 0.999$ (s.e. 0.029). The regression spline model, on the other hand, indicates that Cherry Bark Tortrix moths mate in two consecutive waves. These waves have biological significance as the smaller first wave corresponds to larvae who overwinter in the developmental state ("instar" stage) and are ready to breed earlier than larvae hatched in early summer corresponding to the second larger wave. The Weibull model is unable to match these changes in the rate over time. If a full parametric model is preferred, the resulting spline fit can provide direction for choices of parametric functional forms for the baseline rate that can accommodate multiple peaks. Here, a two-point mixture model would likely be reasonable as it would reflect the biological interpretation of the two peaks. In particular, a mixture of two lognormal distributions seems to provide a fair approximation to the estimated baseline rate function. Figure 2.2 shows the least squares fit of the lognormal mixtures to the baseline rate estimated from the spline fit. The corresponding cumulative rates are also displayed.

2.2.2 Knot Placement

Kooperberg et al. (1995, 1997) have suggested some useful procedures for selecting knots. Their algorithm includes insertion and deletion steps using Akaike's Information Criterion, for example, for assessment of the fit. Implementing their insertion-deletion algorithm for random effects non-homogeneous Poisson processes requires sophisticated computational algorithms which are quite complex. However, by using an adaptation of their knot placement strategy we can evaluate knot placement for our panel study. Let S, the set of all panel followup times T_{ij}, as well as the midpoint of panel intervals $(T_{i(j-1)} + T_{ij})/2$, be considered as the set of candidate knot values. Under the constraint of using 6 knots, and having at least 2 panel times between each knot (inclusive of the knots themselves), there are 715 possible sets of knot placements within S. The second constraint above ensures that there are sufficient data values between knots to avoid extreme local smoothing with the spline fitting the data exactly between knots. Each of these 715 candidates was used to fit the spline model to the data. Optimality of knot placement was ranked using the log likelihood based on the gamma distribution for the random effects. Table 2.1 displays summary statistics for the 715 values of the maximized log likelihoods. The log-likelihood corresponding to the simple strategy for knot placement adopted in the previous section is -3657.86 which is the 97th percentile.
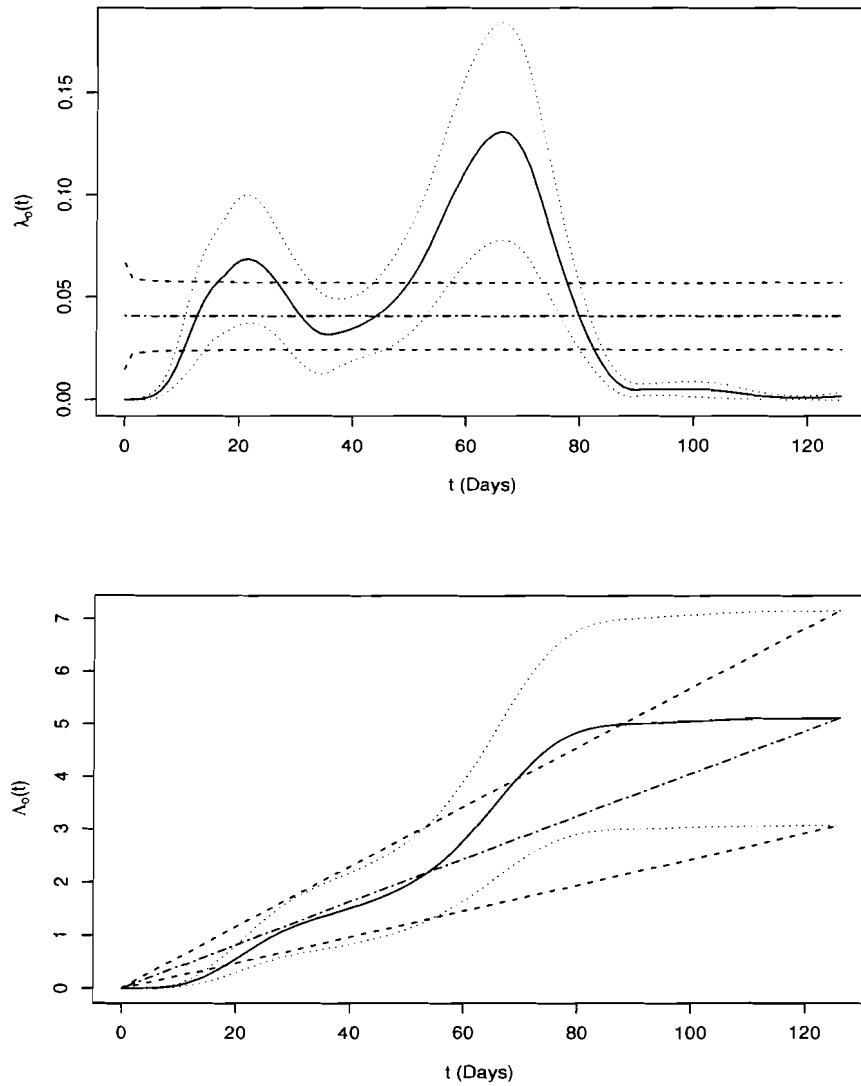
Figure 2.1: Plots of the B-spline estimates (——) and 95% point-wise confidence bands
(· · · · ·) of the cumulative baseline rate $\Lambda_o(t)$ and baseline rate $\lambda_o(t)$ as well as the cor-
responding fully parametric Weibull rate estimates (—·—) and 95% point-wise confidence
bands (- - -) for the Cherry Bark Tortrix Data.
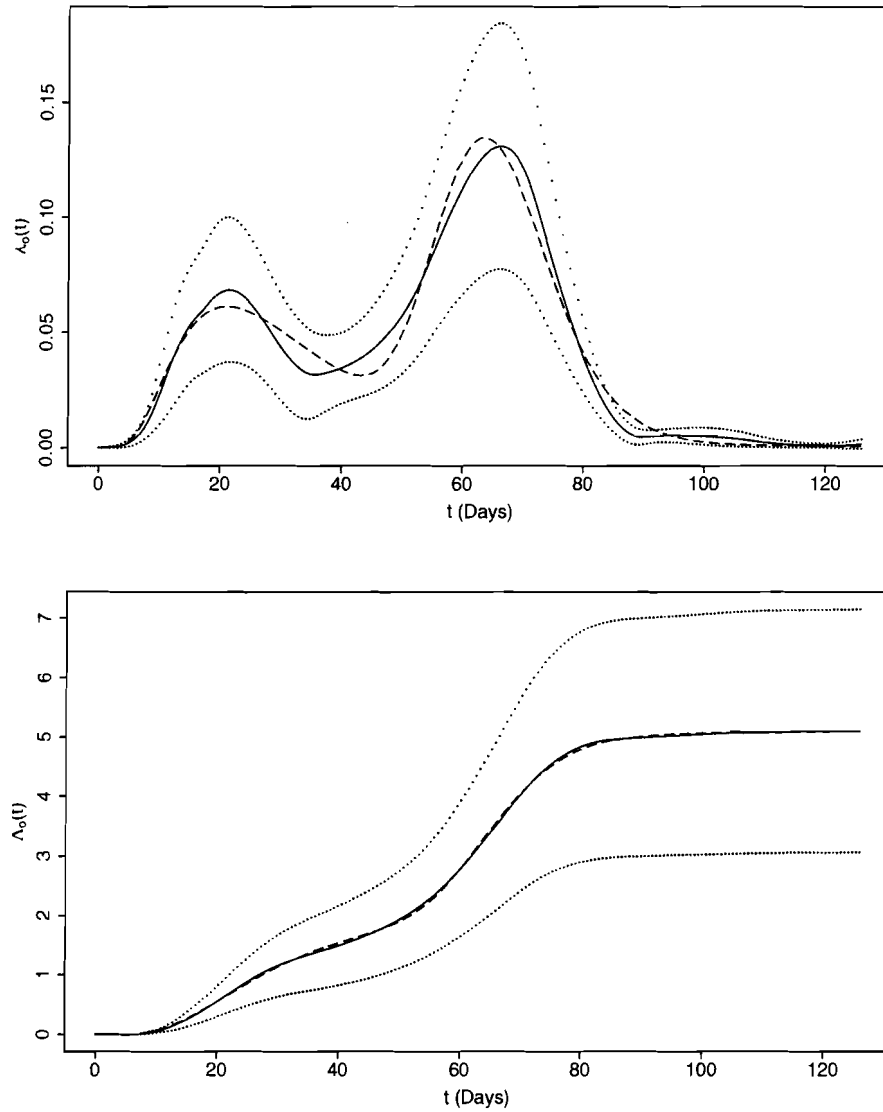
Figure 2.2: The B-spline estimates (——) and 95% point-wise confidence bands (······) of the cumulative baseline rate $\Lambda_o(t)$ and baseline rate $\lambda_o(t)$ as well as the least squares fit of a lognormal mixture (- - -) to the baseline rate estimated from the B-spline fit.

Minimum	-3780.425
1st Quartile	-3723.941
Median	-3695.592
Mean	-3702.043
3rd Quartile	-3678.084
Maximum	-3648.645
Std. Dev.	31.507

Table 2.1: Summary statistics of the log-likelihoods of the 715 possible knot sequences.

Figure 2.3 compares the fitted spline from Section 2.2.1 with that using the knot placement yielding the highest log-likelihood (with knots at 9.0, 21.0, 45.0, 70.0, 87.5 and 119.0 days). If the rate function is fairly smooth, a wide variety of knot placements generally lead to satisfactory estimates for the examples we have studied. Plots of the non-parametric step function estimate of the rate function (Balshaw and Dean, 2002) are very helpful in suggesting knot placement and how many knots are required for an adequate fit. Using Kooperberg's criterion or investigating all possible placements within some reasonable subset of knots provides a fair strategy for optimal allocation.

2.2.3   Small Sample Properties of Estimators

Simulation studies were conducted to investigate how well standard errors and confidence intervals are estimated for the spline methods proposed and the ability of the semiparametric quasi-likelihood regression spline analysis to recover different smooth shapes of cumulative baseline rates. The first simulation was designed to mimic the Cherry Bark Tortrix data and focuses on evaluating biases in the estimation procedure and the performance of usual asymptotic normal confidence intervals. Five thousand data sets were generated using the "thinning" algorithm (Ross, 1990) assuming a conditional Poisson intensity with only one covariate, the treatment, with the same spline form as in the analysis of the Cherry Bark Tortrix data, and with gamma subject-specific effects. The true values of all parameters were assigned to be close to those obtained in the Cherry Bark Tortrix analysis. The true rate function is displayed in Figure 2.4. Two cases were considered: Case I: 20 trees ($I = 20$) similar to the original data, 10 treatment and 10 controls, and Case II: 200 trees ($I = 200$), 100 treatment and 100 controls, to investigate the properties of larger samples. The data
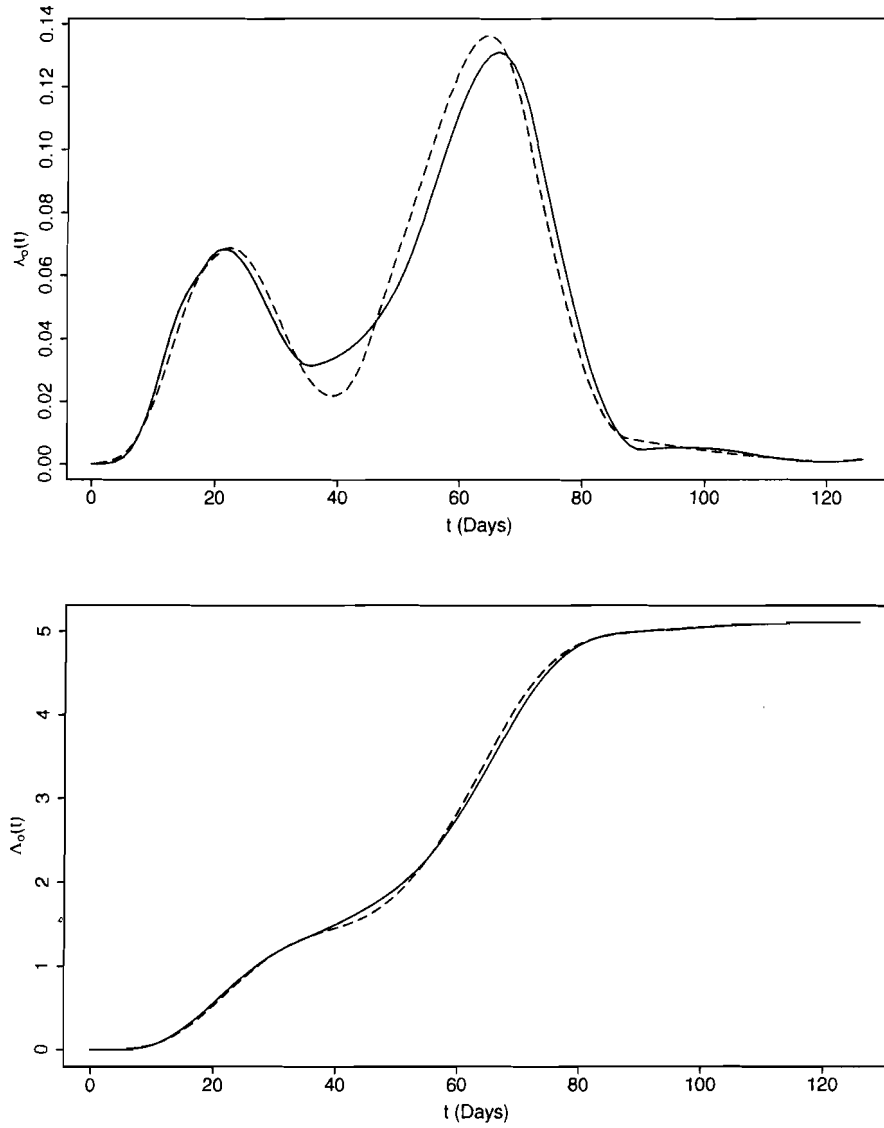
Figure 2.3: Plots of the B-spline estimates (——) of the cumulative baseline rate $\Lambda_o(t)$ and baseline rate $\lambda_o(t)$ of the original model with knots at (9.0,36.0,54.0,72.0,90.0,119.0) as well those of the B-spline fit yielding the highest log-likelihood (- - -) with knots at (9.0,21.0,45.0,70.0,87.5,119.0).

were binned according to the panel structure for the Cherry Bark Tortrix data with 19 intervals over 18 weeks of follow-up, generating the panel counts. Each dataset was then analyzed as in Section 2.2.1 using a cubic spline with identical knots as previously. The quasi-likelihood estimates of θ were computed as well as their robust standard errors.

Table 2.2 contains the the true generating values (True), simulated mean values (Mean), simulated standard errors (Simulated SE), robust standard errors (Robust SE) computed using the sandwich variance estimator and 95% coverage probabilities (CP) for β_1, τ and $\Lambda_o(t)$ at $t = 10, 50, 90, 126$ for both cases. Even for Case I the estimates are quite close to their target values. The robust standard errors for this case are somewhat lower than the simulated ones resulting in undercoverage. When I is increased to 200 the differences between the simulated and robust standard errors are negligible. Coverage probabilities are quite close to the nominal 95% level except for the parameter τ which displays undercoverage. Similar results for estimation of overdispersion parameters have been reported elsewhere (Dean and Lawless, 1989; Breslow, 1990). Figure 2.4 displays the simulated mean values for the baseline and cumulative rates along with their associated simulated 95% point-wise confidence bands. This plot indicates that the estimators of the baseline and cumulative rates have low bias, though the bias increases when $t > 80$ for Case I. Several other simulations have been conducted to investigate the use of regression splines to model recurrent events collected as both event times and panel data. Briefly, the methods discussed herein seem to be able to capture a variety of simple smooth shapes including constant rates or rates which are increasing linearly with time.

2.3 Discussion

This chapter presents a flexible and robust method of incorporating regression splines into the quasi-likelihood analysis of recurrent event data to provide smooth estimates of the mean and baseline rates. Quasi-likelihood only requires low order moment assumptions and also allows for efficient methods of dealing with the heterogeneity often encountered when analyzing counts. In the framework developed, regression splines are used to model the cumulative baseline rate so that the parameters β and ψ enter the logarithm of the mean in a linear fashion leading to a straightforward interpretation. Only simple adaptations of traditional quasi-likelihood estimating functions for the analysis of longitudinal counts are required with this parameterization; it also seems fairly robust with regard to knot

Case I: $I = 20$					
Parameter	True	Mean	Simulated SE	Robust SE	CP
β_1	3.321	3.335	0.390	0.351	0.909
τ	-0.478	-0.638	0.421	0.383	0.908
$\Lambda_o(10)$	0.050	0.049	0.019	0.018	0.866
$\Lambda_o(50)$	1.928	1.942	0.568	0.529	0.870
$\Lambda_o(90)$	4.993	5.033	1.464	1.361	0.872
$\Lambda_o(126)$	5.100	5.142	1.495	1.390	0.871
Case II: $I = 200$					
Parameter	True	Mean	Simulated SE	Robust SE	CP
β_1	3.321	3.320	0.118	0.119	0.953
τ	-0.478	-0.494	0.143	0.133	0.915
$\Lambda_o(10)$	0.050	0.050	0.006	0.006	0.944
$\Lambda_o(50)$	1.928	1.930	0.174	0.174	0.940
$\Lambda_o(90)$	4.993	4.999	0.447	0.449	0.942
$\Lambda_o(126)$	5.100	5.106	0.457	0.459	0.942

Table 2.2: Mean values of estimates, simulated standard errors, robust standard errors and coverage probabilities for 95% confidence intervals of the Cherry Bark Tortrix simulation for both $I = 20$ and $I = 200$.
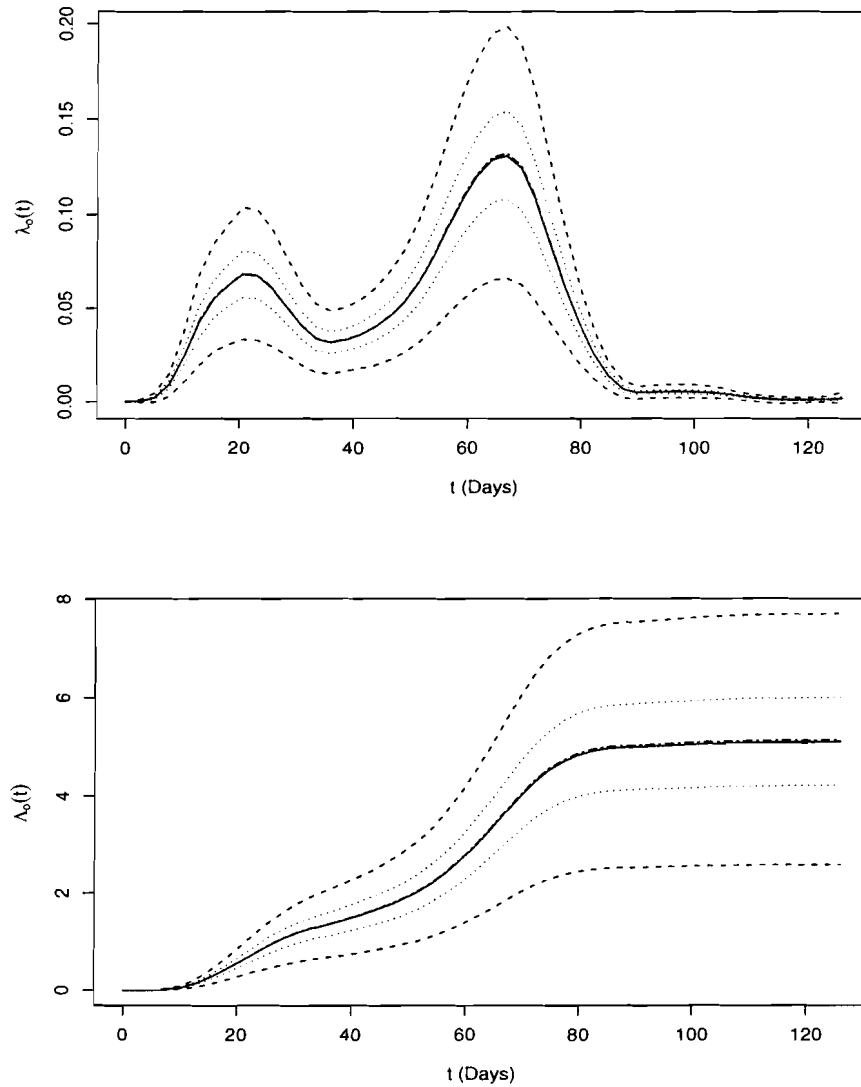
Figure 2.4: The true baseline rate $\lambda_o(t)$ and cumulative baseline rate $\Lambda_o(t)$ denoted by solid lines (——), for the Cherry Bark Tortrix simulation; the average corresponding simulated estimates denoted by dot-dash lines (—·—) for $I = 20$ and long-dash lines (- -) for $I = 200$, along with their associated 95% confidence bands denoted by dashed lines (- - -) and dotted lines (......) respectively.

placement. The flexibility offered by this approach can provide considerable insight into the underlying recurrence rate.

In cases where the sample size is quite small and recurrences are very sparse estimates of the cumulative rate may not be strictly monotone. In the following chapters monotonicity is imposed by modeling the baseline intensity with a log linear form and numerical quadrature is used to evaluate the cumulative rate. This adds computational complexity and requires the development of specialized software but eliminates the possibility of this behaviour in the sparse data scenario.

Chapter 3

Adaptive Functional Mixed NHPP Model

In longitudinal studies it is not unusual that each object or individual under study may experience multiple occurrences of some primary event of interest and where the number of such events between scheduled followup times is recorded. A common approach for analysis of such recurrent events is the use of counting process models, such as the Poisson (Andersen et al., 1995) or related quasi-likelihood techniques (Heyde, 1997). The nonhomogeneous Poisson process (Lawless, 1987a) is also popular for handling situations when its intensity is not constant over time. In addition, random effects or frailty models (Hougaard et al., 1997) have been developed to handle the heterogeneity common to count data analysis. These methods have wide application in econometrics (Hausman et al., 1984), environmetrics (Burnett et al., 1994) and medicine (Abu-Libdeh et al., 1990). For a thorough review of methods for recurrent event analysis see Cook and Lawless (2002).

In some situations modeling the intensity function generating the counting process as a smooth function is of interest. Here we consider the use of penalized spline smoothing with a penalty that varies over time permitting the spline to adapt to temporal changes in smoothness. In addition, we also develop methods for covariate effects which vary smoothly over time or are incorporated in a simple proportional intensity framework, resulting in a highly flexible spline estimator. We offer comparisons with the special case of the methodology where the penalty is constant, which is the traditional approach in smoothing methods for

continuous data (Wahba, 1990; Green and Silverman, 1994). In the context of the regression analysis of normal data the use of an adaptive penalty has been considered by Ruppert and Carroll (2000) who model the penalty as a piece-wise linear function and Pintore et al. (2006) who consider a piece-wise constant penalty.

In Section 3.1, the adaptive spline method is developed. The penalized quasi-likelihood estimating functions are derived in Section 3.2. In Section 3.3, we discuss relaxing the proportional intensity assumption used in Section 3.2. Inference based on both jackknife and bootstrap procedures is presented in Section 3.4. A simulation study to evaluate the performance of the proposed estimating equations is discussed in Section 3.5 where comparisons are made with the estimator derived under a simpler constant penalty assumption. Section 3.6 considers the use of the methods in an analysis of an entomological study to illustrate the proposed methodology. The chapter concludes with a discussion of the techniques developed.

3.1 Adaptive Functional Mixed NHPP Model

Suppose that data are collected as a vector of counts in contiguous non-overlapping time intervals or panels for each member of a group of subjects under observation. Let $\mathbf{N}_i = (N_{i1}, \ldots, N_{ie_i})^T$, $i = 1, \ldots, I$ denote the independent random vector of these counts for subject i where N_{ij} is the number of events observed to occur in the jth time panel $[T_{i,j-1}, T_{ij})$, $j = 1, \ldots, e_i$ with $T_{i0} < T_{i1} < \cdots < T_{ie_i}$ and set $T_{i0} = 0$. In order to simplify the development of the model we express each of these followup times as a multiple of some common unit of time denoted by t_c, so that the panel $T_{ij} - T_{i,j-1} = c_{ij}t_c$, with c_{ij} being the number of these time units occurring in the jth time panel. Panel times are often neatly structured and are always recorded in a common unit of measure such as days or weeks in which case t_c could represent any sub-unit of measurement such as a minute or second for example. These counts are assumed to be generated from a underlying counting process, $\{N_i(t), t > 0\}$, with proportional intensity

$$\nu_i \lambda_i (t; \mathbf{x}_i) = \nu_i \lambda_0(t) \exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}\right\} = \nu_i \exp\left\{\beta_0(t) + \mathbf{x}_i^T \boldsymbol{\beta}\right\} \qquad (3.1)$$

where $\lambda_0(t) = \exp\left\{\beta_0(t)\right\}$ is the baseline intensity with $\beta_0(t)$ being an unknown smooth function, $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]^T$ is an known vector of covariates, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$ is its coefficient and the subject-specific effects ν_i, $i = 1, \ldots, I$, are frailties that are assumed

to be independent and identically distributed random variables from an unknown mixing distribution with mean 1 and variance τ. These frailties account for possible heterogeneity reflecting all the information not contained in $\mathbf{x}_i$. Denote N_{ijk}, $k = 1, \ldots, c_{ij}$ as the number of events that occur in $(T_{i,j-1,k-1}, T_{i,j-1,k}]$ where $T_{ijk} = T_{i,j} + k \cdot t_c$ so that $N_{ij} = \sum_{k=1}^{c_{ij}} N_{ijk}$. Conditional on the unknown frailties $N_i(t)$ is assumed to be a nonhomogeneous Poisson process (NHPP) so that conditionally N_{ijk} is Poisson distributed with mean

$$\nu_i \ [\Lambda_i(T_{i,j-1,k}) - \Lambda_i(T_{i,j-1,k-1})] = \nu_i \exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}\right\} \int_{T_{i,j-1,k-1}}^{T_{i,j-1,k}} \lambda_0(t) \ \mathrm{d}t \qquad (3.2)$$

where $\Lambda_i(t) = \int_0^t \lambda_i(u) \ \mathrm{d}u$ and similarly define $\Lambda_0(t) = \int_0^t \lambda_0(u) \ \mathrm{d}u$. The unconditional mean and variance of the vector of responses for subject i are given by $E[\mathbf{N}_i] = \boldsymbol{\mu}_i = [\mu_{i1}, \ldots, \mu_{ie_i}]^T$ and $V[\mathbf{N}_i] = \text{diag}\{\boldsymbol{\mu}_i\} + \tau \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T$ where

$$\mu_{ij} = t_c \exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}\right\} \sum_{k=1}^{c_{ij}} \int_{T_{i,j-1,k-1}}^{T_{i,j-1,k}} \lambda_0(t) \ \mathrm{d}t. \qquad (3.3)$$

To complete the specification of the model it will be assumed that $\beta_0(t)$ exists in the space of quadratic splines (de Boor, 1978; Ruppert et al., 2003) on the defining knot sequence

$$0 < \xi_1 < \cdots < \xi_s < \xi_{s+1} = \max\{T_{ie_i}\} \qquad (3.4)$$

assumed here to be an equally spaced grid. The spline space is linear and so can be represented as a linear combination of basis functions that span this space

$$\beta_0(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \sum_{j=1}^{s} a_j(t - \xi_j)_+^2$$

where $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2]^T$ and $\mathbf{a} = [a_i, \ldots, a_s]^T$ are vectors of spline coefficients, and $x_+ = x1(x > 0)$; 1 is the indicator function. The truncated power basis is used here for ease of presentation but in practice any basis that spans this spline space defined on the knot sequence (3.4) can be used; a popular choice is the numerically fast and stable B-spline basis (de Boor, 1978; Eilers and Marx, 1996). The number of interior knots s will be chosen so that knots are relatively dense, for example close to the upper limit of the number of followup time $\max\{e_i\}$. As a rule of thumb we suggest $s \approx 0.75 \cdot \max\{e_i\}$ which ensures an overparameterized model.

In order to control the degrees of freedom and hence the flexibility of the baseline spline β_0 a quadratic form penalty, $\mathbf{a}^T \Delta \mathbf{a}$, is used to regularize the fit. The simplest and most

common penalty (Ruppert et al., 2003) when using the truncated power basis would be $\Delta = \delta \mathbf{I}_s$ leading to the standard penalized spline. The assumption of a time homogeneous penalty term may not be reasonable in some situations. For example, if the function β_0 oscillates rapidly in the first half of its domain and is quadratic or linear in the second half, then using a fixed penalty will likely lead to over-smoothing in the first part and under-smoothing in the second part. With a constant penalty, the amount of flexibility allowed is determined over the full domain of the function (Wand, 2000). Here we allow the penalty term to vary over time giving the spline the flexibility to adapt to differing degrees of smoothness/curvature over its domain. This is in the spirit of Ruppert and Carroll (2000) who permit the penalty term to vary over time as a linear interpolation of a set of initial penalty terms at a subset of times. Here, however, the log of $\delta(t)$, the time dependent smoothing term will be assumed to be a spline itself with

$$\log [\delta(t)] = \gamma_0 + \gamma_1 t + \sum_{j=2}^{s-1} b_{j-1}(t - \xi_j)_+$$

where $\boldsymbol{\gamma} = [\gamma_0, \gamma_1]^T$ the non-regularized and $\mathbf{b} = [b_1, \ldots, b_{s-2}]^T$ the regularized spline coefficients with penalty $\rho \, \mathbf{I}_s$. The penalty for the baseline spline $\beta_0(t)$ in this case is then of the form $\boldsymbol{\Delta} = \text{diag}\{\delta(\xi_1), \ldots \delta(\xi_s)\}$. It will be assumed that ρ is known and will be fixed at a value so that few effective degrees of freedom are used to estimate $\delta(t)$. The goal here is to allow the penalty to vary over time permitting local adaptation of the baseline spline and so only determining the general temporal trend of $\delta(t)$ is of interest. Fixing ρ is not unreasonable in this situation and is akin to using a fixed knot regression spline for $\log[\delta(t)]$ with a small number of knots. Using many knots with a penalty dampens the importance of knot position while simultaneously keeping the effective number of parameters low, if the value of ρ is suitably large. Though an estimate of $\delta(t)$ is not of primary interest it will give an indication of where $\lambda_0(t)$ is deviating from the assumption of a uniform level of smoothness.

3.2 Penalized Quasi-likelihood Estimation

Estimates of $\theta = [\beta^T, \alpha^T]^T$ and $\mathbf{a}$ are obtained by solving the following estimating equations

$$\mathbf{g}_\theta = \sum_{i=1}^{I} \mathbf{D}_{\theta,i}^T \mathbf{V}_i^{-1} (\mathbf{N}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

$$\mathbf{g}_\mathbf{a} = \sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},i}^T \mathbf{V}_i^{-1} (\mathbf{N}_i - \boldsymbol{\mu}_i) - \boldsymbol{\Delta}\mathbf{a} = \mathbf{0} \tag{3.5}$$

where $\mathbf{D}_{\mathbf{y},i} = \partial\boldsymbol{\mu}_i / \partial\mathbf{y}^T$, $\mathbf{V}_i^{-1} = \mathrm{diag}\{1/\mu_{ij}\}_{e_i \times e_i} - \tau/(1 + \tau\mu_{i+}) \mathbf{J}_{e_i}$, $\mathbf{J}_{e_i}$ a $e_i \times e_i$ matrix of ones and $\mu_{i+} = \sum_{j=1}^{e_i} \mu_{ij}$. For a simple homogeneous Poisson model where $\beta_0(t) = \alpha_0$, $\mathbf{g}_\theta$ reduces to the Poisson maximum likelihood estimating equations; with overdispersion introduced through the frailty terms, these are the usual quasi-likelihood estimating equations (Thall, 1988).

To compute the expected number of events per panel, μ_{ij}, the non-analytic integral values in (3.3) are evaluated as $t_c \exp\{\beta_{0k}\}$ for suitably small t_c with $\beta_{0k} = \beta_0(T_{i,j-1,k})$ so that in the limit as $t_c \to 0$, (3.3), becomes

$$\mu_{ij} = t_c \exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}\right\} \sum_{k=1}^{c_{ij}} \exp\left\{\beta_{0k}\right\}. \tag{3.6}$$

How small a value of t_c is used depends on the precision desired which can be verified during computation by halving the value of t_c, re-computing μ_{ij}, and evaluating the absolute difference between the two values. This process can be repeated until the specified level of precision is obtained. Solutions to the set of equations (3.5) can be obtained by using Fisher's scoring by iteratively solving

$$\begin{bmatrix} \sum_{i=1}^{I} \mathbf{D}_{\theta,i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\theta,i} & \sum_{i=1}^{I} \mathbf{D}_{\theta,i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\mathbf{a},i} \\ \sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\theta,i} & \sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\mathbf{a},i} + \boldsymbol{\Delta} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{I} \mathbf{D}_{\theta,i}^T \mathbf{V}_i^{-1} \mathbf{N}_i^* \\ \sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},i}^T \mathbf{V}_i^{-1} \mathbf{N}_i^* \end{bmatrix} \tag{3.7}$$

where $\mathbf{N}_i^* = \mathbf{N}_i - \boldsymbol{\mu}_i + \mathbf{D}_{\theta,i}\boldsymbol{\theta} + \mathbf{D}_{\mathbf{a},i}\mathbf{a}$, with $\boldsymbol{\mu}_i$, $\mathbf{V}_i$, $\mathbf{D}_{\theta,i}$ and $\mathbf{D}_{\mathbf{a},i}$ all evaluated at the current value of $\boldsymbol{\theta}$ and $\mathbf{a}$. The linear system (3.7) can be recognized as the linear mixed model equations (Harville, 1977). Connections between penalized quasi-likelihood and linear mixed models have been discussed by Breslow and Clayton (1993) while similar links between penalized splines and linear mixture models has been well documented by Lin and Zhang (1999) and Ruppert et al. (2003) for example. The use of penalized splines and penalized

quasi-likelihood creates a structure where estimators are obtained within similar frameworks and which mimic well known linear mixed model estimators.

To estimate the variance parameters, the overdispersion parameter τ and the penalty spline coefficients γ and $\mathbf{b}$ we develop estimators again within the penalized quasi-likelihood framework (Breslow and Clayton, 1993) where the unknown likelihood component of this function is replaced with Pearson's chi-squared statistic. An approximate penalized likelihood or penalized quasi-likelihood function used for estimation is obtained by taking a Laplace approximation (Tierney et al., 1989) with respect to θ and $\mathbf{a}$ where θ is assumed to have a diffuse prior:

$$ql(\tilde{\theta}, \tilde{\mathbf{a}}, \gamma, \mathbf{b}) = -\sum_{i=1}^{I}(\mathbf{N}_i - \tilde{\mu}_i)^T \mathbf{V}_i^{-1}(\mathbf{N}_i - \tilde{\mu}_i) + \sum_{j=1}^{s} \log [\delta(\xi_j)] - \tilde{\mathbf{a}}^T \Delta \tilde{\mathbf{a}} - \rho \mathbf{b}^T \mathbf{b} - \log |\mathbf{H}| \quad (3.8)$$

where $\tilde{\theta}$ and $\tilde{\mathbf{a}}$ are the solutions to the system of equations in (3.7), $\tilde{\mu}_i$ is μ_i evaluated at this solution and $\mathbf{H} = \sum_{i=1}^{I} \mathbf{H}_i$ with

$$\mathbf{H}_i = \left[\begin{array}{cc} \mathbf{D}_{\theta,i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\theta,i} & \mathbf{D}_{\theta,i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\mathbf{a},i} \\ \mathbf{D}_{\mathbf{a},i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\theta,i} & \mathbf{D}_{\mathbf{a},i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\mathbf{a},i} + \Delta \end{array} \right];$$

for notational convenience denote the inverse of $\mathbf{H}$ as

$$\mathbf{H}^{-1} = \left[\begin{array}{cc} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{array} \right]^{-1} = \left[\begin{array}{cc} \mathbf{H}^{11} & \mathbf{H}^{12} \\ \mathbf{H}^{21} & \mathbf{H}^{22} \end{array} \right] \quad (3.9)$$

with $\mathbf{H}_{12} = \mathbf{H}_{21}^T$ where $\mathbf{H}$ conforms to the partitioning of $\mathbf{H}_i$. From (3.8) the estimating equation for τ is given by

$$g_\tau = \sum_{i=1}^{I} \frac{(N_{i+} - \mu_{i+})^2 - \mu_{i+}(1 + \tau \ \mu_{i+}) + r_i}{(1 + \tau \ \mu_{i+})^2} = 0. \quad (3.10)$$

where

$$r_i = \text{tr}\left\{ \mathbf{H}^{-1} \left(\mathbf{H}_i - \left[\begin{array}{cc} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Delta \end{array} \right] \right) \right\}$$

a first-order correction term, $N_{i+} = \sum_{j=1}^{e_i} N_{ij}$ and $\mu_{i+} = \sum_{j=1}^{e_i} \mu_{ij}$. This estimating equation is equivalent to the so-called 'pseudo-likelihood' equation of Davidian and Carroll (1987). The effective degrees of freedom of the baseline spline estimate, a quantity used

in the evaluations of Section 3.5, can be obtained from the r_i's and is given by $\sum_{i=1}^{I} r_i - p$.
Estimating equations for γ and $\mathbf{b}$ are similarly derived and given by

$$\mathbf{g}_\gamma = \begin{bmatrix} s - \tilde{\mathbf{a}}^T \boldsymbol{\Delta} \tilde{\mathbf{a}} - \text{tr}\left\{\mathbf{H}^{22}\boldsymbol{\Delta}\right\} \\ \sum_{j=1}^{s} \xi_j - \tilde{\mathbf{a}}^T \boldsymbol{\Delta}_{\gamma_1} \tilde{\mathbf{a}} - \text{tr}\left\{\mathbf{H}^{22}\boldsymbol{\Delta}_{\gamma_1}\right\} \end{bmatrix} = \mathbf{0}$$

$$\mathbf{g}_\mathbf{b} = \begin{bmatrix} \sum_{j=1}^{s}(\xi_k - \xi_j)_+ - \tilde{\mathbf{a}}^T \boldsymbol{\Delta}_{b_k} \tilde{\mathbf{a}} - \text{tr}\left\{\mathbf{H}^{22}\boldsymbol{\Delta}_{b_k}\right\} \end{bmatrix}_{(s-2)\times 1} - 2\rho\,\mathbf{b} = \mathbf{0} \qquad (3.11)$$

respectively where $\boldsymbol{\Delta}_y = \partial\boldsymbol{\Delta}/\partial y = \text{diag}\left\{\delta(\xi_k)\cdot\partial\log\left[\delta(\xi_k)\right]/\partial y\right\}_{(s-2)\times(s-2)}$.

To control the amount of smoothing of the penalty spline, $\delta(t)$, the smoothing parameter ρ is selected such that

$$d = \text{tr}\left\{\begin{bmatrix} \mathbf{K}_{\gamma,\gamma} & \mathbf{K}_{\gamma,\mathbf{b}} \\ \mathbf{K}_{\gamma,\mathbf{b}}^T & \mathbf{K}_{\mathbf{b},\mathbf{b}} + 2\,\rho\mathbf{I}_{s-2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{K}_{\gamma,\gamma} & \mathbf{K}_{\gamma,\mathbf{b}} \\ \mathbf{K}_{\gamma,\mathbf{b}}^T & \mathbf{K}_{\mathbf{b},\mathbf{b}} \end{bmatrix}\right\} = \text{tr}\left\{\mathbf{K}^{-1}\mathbf{L}\right\} \qquad (3.12)$$

where $\mathbf{K}_{\mathbf{y}_1,\mathbf{y}_2} = \{k_{ij}\}$ with

$$k_{ij} = \tilde{\mathbf{a}}^T\left(\partial^2\boldsymbol{\Delta}/\partial y_{1i}\,\partial y_{2j}\right)\tilde{\mathbf{a}} + \text{tr}\left\{\mathbf{H}^{22}\left(\partial^2\boldsymbol{\Delta}/\partial y_{1i}\,\partial y_{2j}\right)\right\} - \text{tr}\left\{\mathbf{H}^{22}\boldsymbol{\Delta}_{y_{1i}}\mathbf{H}^{22}\boldsymbol{\Delta}_{y_{2j}}\right\};$$

the effective degrees of freedom is set to d, some value much less than $s-2$ and is determined using bracketing and bisection as there is a one-to-one correspondence between ρ and d (i.e. for fixed values of ρ one can solve for all the other parameters, compute the effective degrees of freedom (3.12) for each of these values, find a bracketing interval for the desired d, and obtain the solution via bisection).

Alternatively, if a large amount of data is available, an estimate of ρ can also be obtained if desired by taking another Laplace approximation about the solution to the equations in (3.11) and results in the following updating step

$$\rho = \frac{s-2}{\mathbf{b}^T\mathbf{b} + \text{tr}\left\{\mathbf{K}^{-1}\right\}} \qquad (3.13)$$

which is iterated with solutions to the estimating equations (3.5), (3.10) and (3.11) until convergence.

3.3 Relaxing the Assumption of Proportional Intensities

The proportional intensity assumption for incorporating covariate effects is not always reasonable. Where time varying coefficients are postulated, model (3.1) may be replaced with

$$\nu_i \lambda_i(t; \mathbf{x}_i) = \nu_i \exp \left\{ \beta_0(t) + \sum_{r=1}^{q} x_{ir} \beta_r(t) + \mathbf{x}_i^T \boldsymbol{\beta} \right\}$$

where $\mathbf{x}_i$ is a known $p - q$ vector of covariates, $\boldsymbol{\beta}$ the associated time homogeneous covariate effects, $\beta_0(t)$ is an adaptive spline and $\beta_r(t)$ are quadratic splines

$$\beta_r(t) = \alpha_{r0} + \alpha_{r1}t + \alpha_{r2}t^2 + \sum_{j=1}^{s} a_{rj}(t - \xi_j)_+^2 \tag{3.14}$$

with $\boldsymbol{\alpha}_r$ and $\mathbf{a}_r$ being the spline coefficients. Here again, the baseline spline β_0 will be considered to have a time varying penalty. This spline models the overall trend in the counting process intensity; hence determining deviation from the assumption of global smoothness is most important for this spline term. The treatment spline effects $\beta_r(t)$ will be modulated by simple penalties of the form $\rho_r \mathbf{I}_s$. Estimates of $\boldsymbol{\theta} = [\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_q]^T$ and $\mathbf{a} = [\mathbf{a}_0, \ldots, \mathbf{a}_q]^T$ can then be obtained in a similar fashion as in the previous section by solving the system of equations

$$\mathbf{H} \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{I} \mathbf{D}_{\boldsymbol{\theta},i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\boldsymbol{\theta},i} & \sum_{i=1}^{I} \mathbf{D}_{\boldsymbol{\theta},i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\mathbf{a},i} \\ \sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\boldsymbol{\theta},i} & \sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},i}^T \mathbf{V}_i^{-1} \mathbf{D}_{\mathbf{a},i} + \mathbf{P} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{I} \mathbf{D}_{\boldsymbol{\theta},i}^T \mathbf{V}_i^{-1} \mathbf{N}_i^* \\ \sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},i}^T \mathbf{V}_i^{-1} \mathbf{N}_i^* \end{bmatrix} \tag{3.15}$$

where $\mathbf{D}_{\mathbf{a},i} = [\mathbf{D}_{\mathbf{a},i1} \ldots \mathbf{D}_{\mathbf{a},iq}]$ with $\mathbf{D}_{\mathbf{y},is} = \partial \boldsymbol{\mu}_i / \partial \mathbf{y}_s^T$ and

$$\mathbf{P} = \begin{bmatrix} \boldsymbol{\Delta} & \mathbf{0} \\ \mathbf{0} & \text{diag}\{\rho_1, \ldots, \rho_q\} \otimes \mathbf{I}_s \end{bmatrix}.$$

The smoothing parameters ρ_r for each of the time varying coefficient splines $\beta_r(t)$, $r = 1, \ldots, q$ are obtained iteratively by using the following updates

$$\rho_r = \frac{s}{\mathbf{a}_r^T \mathbf{a}_r + \mathbf{1}^T \mathbf{u}_r} \tag{3.16}$$

where $\mathbf{u} = [\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_q]^T = \text{diag}\{\mathbf{H}^{22}\}$ with $\mathbf{H}^{22}$ being the $s(q+1) \times s(q+1)$ lower right block matrix of of the inverse of $\mathbf{H}$. These PQL estimates of the smoothing parameters are equivalent to the generalized maximum likelihood estimates proposed by Gu (2002). Estimation of the penalty spline for $\beta_0(t)$, $\delta(t)$, is carried out as previously.

With very large datasets solving (3.15) can become prohibitive. In such cases the Gauss-Seidel method can be used by cyclically solving

$$\left(\sum_{i=1}^{I} \mathbf{D}_{\boldsymbol{\theta},i}^{T} \mathbf{V}_{i}^{-1} \mathbf{D}_{\boldsymbol{\theta},i}\right) \boldsymbol{\theta} = \sum_{i=1}^{I} \mathbf{D}_{\boldsymbol{\theta},i}^{T} \mathbf{V}_{i}^{-1} (\mathbf{N}_{i}^{*} - \mathbf{D}_{i}\mathbf{a})$$

$$\left(\sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},ir}^{T} \mathbf{V}_{i}^{-1} \mathbf{D}_{\mathbf{a},i} + \mathbf{P}\right) \mathbf{a}_{r} = \sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},i}^{T} \mathbf{V}_{i}^{-1} \left(\mathbf{N}_{i}^{*} - \mathbf{D}_{\boldsymbol{\theta},ir}\boldsymbol{\theta} - \sum_{s=1,\ s\neq r}^{q} \mathbf{D}_{\mathbf{a},is}\mathbf{a}_{s}\right), \ r = 1,\ldots,q$$

often referred to as the so-called backfitting algorithm (Buja et al., 1989). The backfitting procedure, although computationally less burdensome, can lead to fairly slow convergence. To improve this situation only a few complete cycles of the backfitting algorithm are performed followed by quasi-newton secant updates (Nocedal and Wright, 1999); secant updating is initialized by

$$\mathrm{diag}\left\{\left(\sum_{i=1}^{I} \mathbf{D}_{\boldsymbol{\theta},i}^{T} \mathbf{V}_{i}^{-1} \mathbf{D}_{\boldsymbol{\theta},i}\right)^{-1}, \left(\sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},i0}^{T} \mathbf{V}_{i}^{-1} \mathbf{D}_{\mathbf{a},i0} + \boldsymbol{\Delta}\right)^{-1}, \left(\sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},i1}^{T} \mathbf{V}_{i}^{-1} \mathbf{D}_{\mathbf{a},i1} + \rho_{1}\ \mathbf{I}_{s}\right)^{-1}, \right.$$
$$\left. \ldots, \left(\sum_{i=1}^{I} \mathbf{D}_{\mathbf{a},iq}^{T} \mathbf{V}_{i}^{-1} \mathbf{D}_{\mathbf{a},iq} + \rho_{q}\ \mathbf{I}_{s}\right)^{-1}\right\}$$

and provides sequentially improving approximations to the inverse of $\mathbf{H}$ leading to fast quadratic convergence near the solution that is computationally tractable for very large systems. This resulting inverse approximation can by used to compute $\mathbf{H}^{22}$ for the update steps (3.16) so that inverting the potentially large matrix $\mathbf{H}$ is avoided.

3.4 Inference using Resampling Methods

Let $\boldsymbol{\varphi} = [\boldsymbol{\theta}^{T}, \mathbf{a}^{T}, \tau, \boldsymbol{\alpha}^{T}, \mathbf{b}^{T}]^{T}$, $\mathbf{g}_{\boldsymbol{\varphi}} = [\mathbf{g}_{\boldsymbol{\theta}}^{T}, \mathbf{g}_{\mathbf{a}}^{T}, \mathbf{g}_{r}, \mathbf{g}_{\gamma}^{T}, \mathbf{g}_{\mathbf{b}}]^{T}$ and denote $\tilde{\boldsymbol{\varphi}}$ as the solution to $\mathbf{g}_{\boldsymbol{\varphi}} = \mathbf{0}$. It can be shown that $\sqrt{I}(\tilde{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) \xrightarrow{\mathfrak{D}} N(\mathbf{0}, \boldsymbol{\Sigma})$. An estimator for $\boldsymbol{\Sigma}$ is the one-step Jackknife

$$\tilde{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{I}(e_{i} - r_{i}) - d - 1}{\sum_{i=1}^{I} e_{i}} \sum_{i=1}^{I} \left\{ E\left[\frac{\partial \mathbf{g}_{\boldsymbol{\varphi}}^{(-i)}}{\partial \boldsymbol{\varphi}^{T}}\right]^{-1} \left(\mathbf{g}_{\boldsymbol{\varphi}}^{(-i)} \mathbf{g}_{\boldsymbol{\varphi}}^{(-i)\ T}\right) E\left[\frac{\partial \mathbf{g}_{\boldsymbol{\varphi}}^{(-i)\ T}}{\partial \boldsymbol{\varphi}}\right]^{-1}\right\}\Bigg|_{\boldsymbol{\varphi}=\tilde{\boldsymbol{\varphi}}} \tag{3.17}$$

proposed by Lipsitz et al. (1994) which is asymptotically equivalent to the robust variance estimator of White (1982); $\mathbf{g}_{\boldsymbol{\varphi}}^{(-i)}$ denotes the estimation equations evaluated with data from the ith individual removed.

Bootstrap estimators (Wu, 1986; Efron, 1987) may also be computed. The following bootstrap procedure uses a one-step estimator for estimation in each bootstrap sample and hence offers computational efficiency:

1) Draw a sample of size I without replacement $(\mathbf{N}_1^{(b)}, \mathbf{x}_i^{(b)}), \ldots, (\mathbf{N}_I^{(b)}, \mathbf{x}_I^{(b)})$ from $(\mathbf{N}_1, \mathbf{x}_i), \ldots, (\mathbf{N}_I, \mathbf{x}_I)$.

2) Compute:

$$\tilde{\varphi}^{(b)} = \tilde{\varphi} + E \left[\frac{\partial \mathbf{g}_\varphi^{(b)}}{\partial \varphi^T} \right]^{-1} \mathbf{g}_\varphi^{(b)} \Bigg|_{\varphi = \tilde{\varphi}} \tag{3.18}$$

where $\mathbf{g}_\varphi^{(b)}$ are the estimating equations evaluated on the bootstrap re-sample $(\mathbf{N}_1^{(b)}, \mathbf{x}_i^{(b)}), \ldots, (\mathbf{N}_I^{(b)}, \mathbf{x}_I^{(b)})$.

3) Repeat 1) and 2) for $b = 1, \ldots, B$.

The bootstrap distribution Φ^* of φ obtained from this process can then be used to test hypotheses and construct confidence intervals for φ. The main benefit of this method over standard bootstrapping techniques is that the one-step linearization (3.18) does not require fully solving $\mathbf{g}_\varphi^{(b)} = \mathbf{0}$ at each bootstrap iteration. We evaluate these estimators in the following section.

3.5 Simulation Study

In order to evaluate the performance of the adaptive spline model a simulation study was performed by generating five thousand data sets from mixed NHPP models with intensity functions $f_1(t)$ and $f_2(t)$ as displayed in Figure 3.1. The forms of the intensity were selected to reflect a situation where the constant penalty should perform well enough, $f_1(t)$, as well as one where the use of a constant penalty may not be appropriate, $f_2(t)$, as this intensity has several regimes in which the curvature of the function is changing. The simulation was performed as a two by two factorial experiment with factors being the number of individuals with levels $I = 30, 50$ and panels with levels of equally spaced panels $e = e_i = 40, 80$ $i = 1, \ldots, I$; the amount of overdispersion was set to a moderate amount, $\tau = 0.3$. To each generated dataset both the adaptive spline mixed NHPP model of Section 3.1 and the constant penalty model were fit for comparison. In addition, both resampling methods of inference discussed in Section 3.4 were performed.
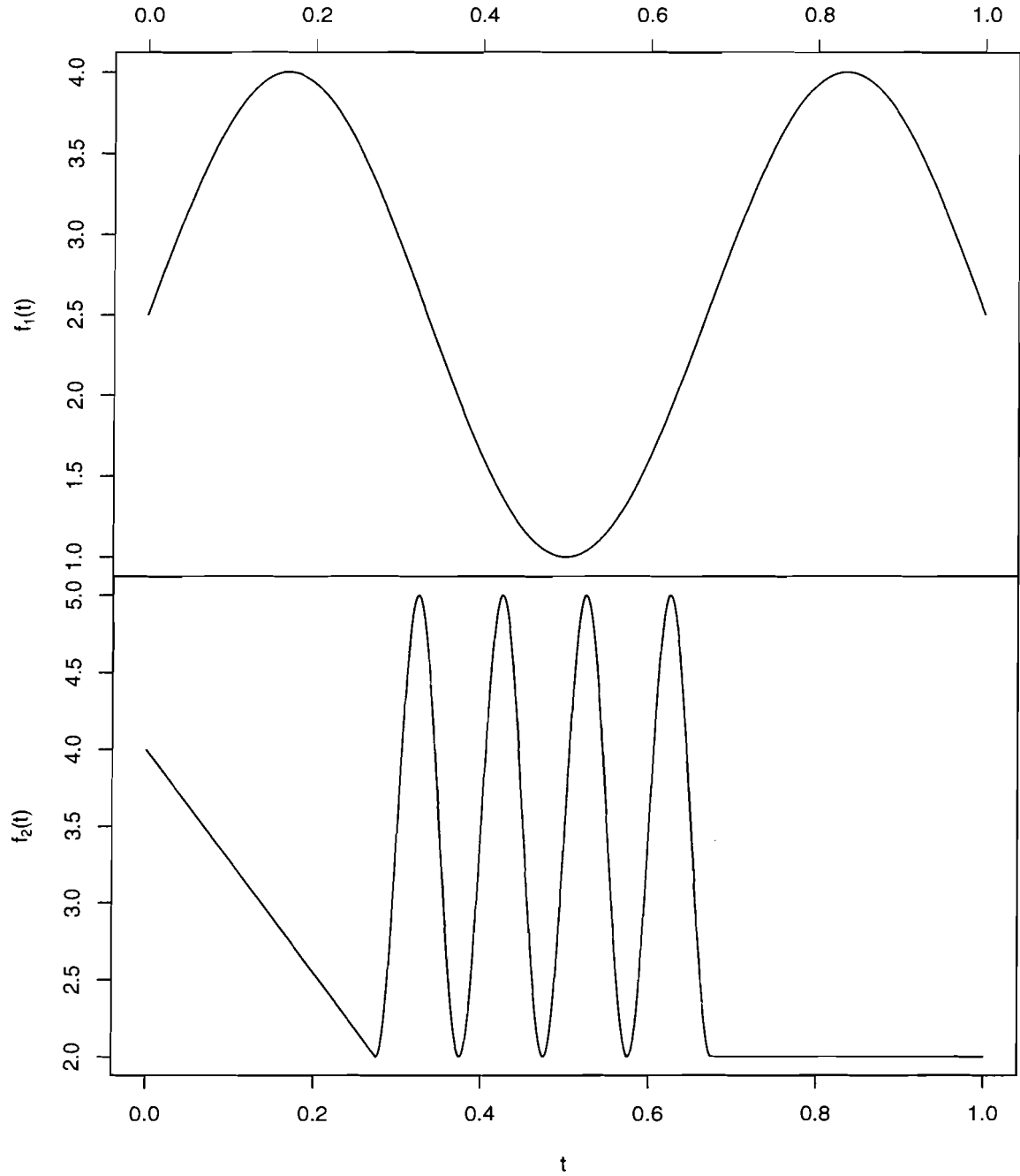
Figure 3.1: Intensity functions considered in the simulation study: $f_1(t)$ - top panel and $f_2(t)$ - bottom panel.

Figures 3.2 and 3.3 present the mean of the simulated estimates of both the adaptive and the constant penalty model overlaying each of the two true generating intensities respectively at all levels of number of individuals (I) and panels (e). The results for generating intensity $f_1(t)$ suggest that both methods are doing a reasonable job of estimating the underlying functional form for all levels of the experiment. It was originally postulated that the adaptive penalty method may be susceptible to over-fitting when the extra flexibility of this approach is unnecessary; results for $f_1(t)$ suggest that this is not the case. Figure 3.3 indicates that the adaptive method is better able to detect regions of high curvature particularly when less data are available. As the amount of data increases however both methods seem able to capture the form of the underlying generating intensity. These results are an overall summary and provide the mean behavior over all fits.

Figure 3.4 displays estimates for both the adaptive (dashed line) and constant penalty (solid grey line) models overlaying the true generating intensity (solid line) $f_2(t)$ as well as plots of the log of their corresponding penalty functions, $\log[\delta(t)]$, of a single run of the experiment at two levels with the least ($I = 30$, $e = 40$) and most ($I = 50$, $e = 80$) amount of data. This figure shows more clearly how the adaptive method is better able to handle changes between regions of low and high curvature, particularly in the case with more data, as the constant penalty fit oscillates in the smooth sections since it is unable to adapt. In the low data scenario the constant penalty method always fails to detect the sharp changes and simply smoothes them out while the adaptive approach is occasionally able to pick up this trend as illustrated. Note a value of $\delta(t)$ above the estimate from the constant penalty model indicates more smoothing relative to the fixed penalty model at that time and that below the estimate from the constant penalty model corresponds analogously to less smoothing. The plots of $\log[\delta(t)]$ show clearly how the adaptive method modulates the flexibility of the spline intensity.

An overall summary of these observations are shown in Figure 3.5 which presents box plots of the integrated squared error, $ISE[f_i(t)] = \int_0^1 [f_i(t) - \tilde{f}_i(t)]^2 \, dt$, $i = 1, 2$ for both the adaptive and constant penalty models at all levels of number of individuals $I = 30, 50$ and numbers of panels $e = 40, 80$. It indicates that the adaptive and constant penalty models perform equivalently under generating intensity $f_1(t)$ while the adaptive approach has superior performance for $f_2(t)$. It is interesting to note that use of an adaptive penalty, although requiring extra degrees of freedom to estimate $\delta(t)$, on average uses fewer effective degrees of freedom to fit the intensity functions. Table 3.1 presents the average estimated

Figure 3.2: Plots of the simulated mean values for estimates based on the adaptive spline (dashed line) and constant penalty spline (solid gray line) overlaying the true value (solid line) of the intensity $f_1(t)$ for all levels of number of individuals $I = 30, 50$ and numbers of panels $e = 40, 80$.
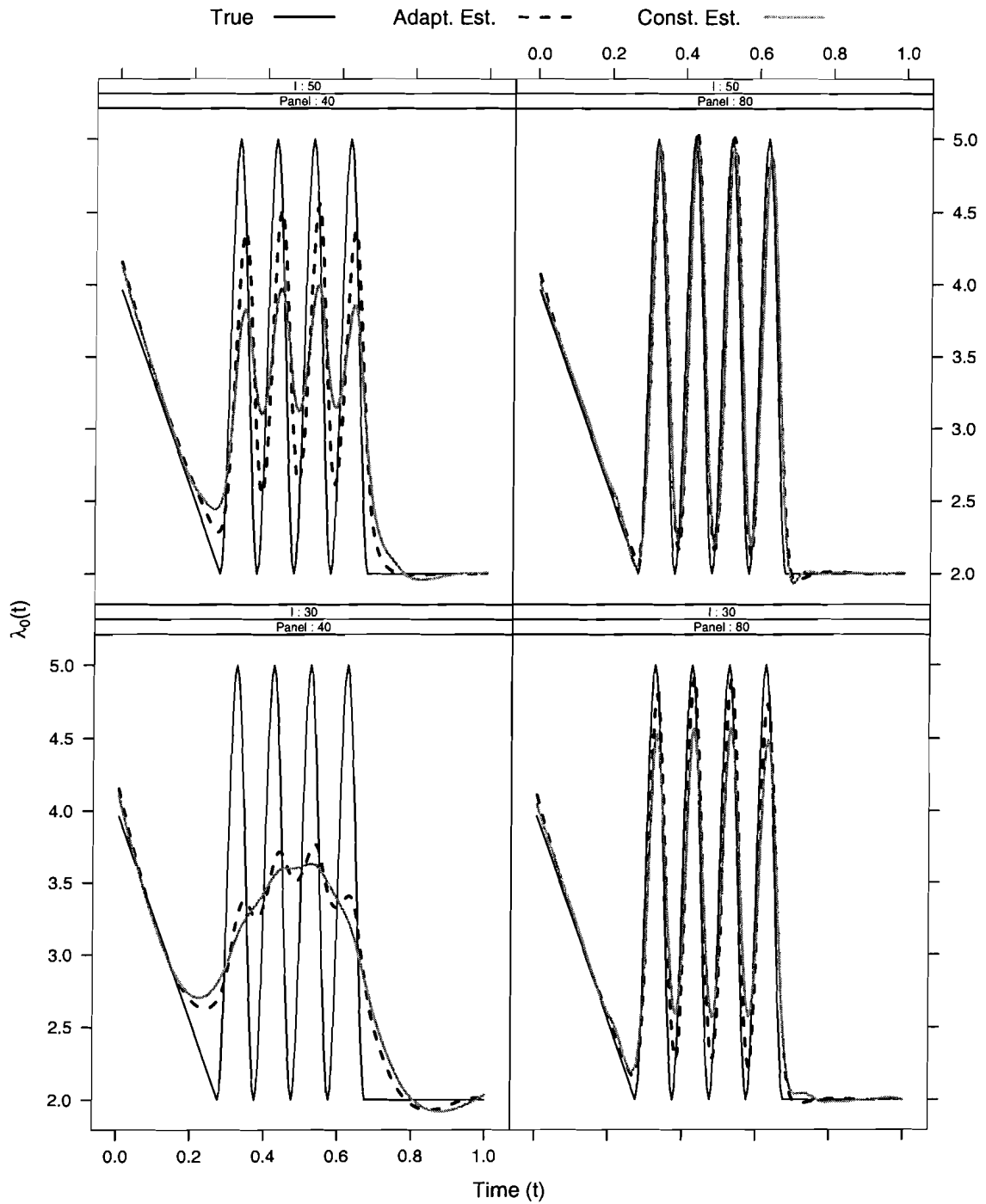
Figure 3.3: Plots of the simulated mean values for estimates based on the adaptive spline (dashed line) and constant penalty spline (solid gray line) overlaying the true value (solid line) of the intensity $f_2(t)$ for all levels of number of individuals $I = 30, 50$ and numbers of panels $e = 40, 80$.

			Penalty	
	Indiv.	Panel	$\delta(t)$	$\delta(t) = \delta$
$f_1(t)$	30	40	6.84	7.58
$f_2(t)$	30	40	7.23	7.00
$f_1(t)$	30	80	7.29	8.36
$f_2(t)$	30	80	16.13	22.31
$f_1(t)$	50	40	7.18	8.14
$f_2(t)$	50	40	13.98	14.95
$f_1(t)$	50	80	7.69	9.03
$f_2(t)$	50	80	17.54	28.15

Table 3.1: Average estimated effective degrees of freedom for the baseline spline from the fits of the adaptive and the constant penalty spline ($\delta(t) = \delta$) models over the simulations for both functional forms of the intensity (Figure 3.1) at all levels of individuals $I = 30, 50$ and number of panels $e = 40, 80$.

effective degrees of freedom for the baseline spline of the fitted models for all levels of the simulation experiment. It seems that the adaptive method is able to perform as well as, and for generating intensity $f_2(t)$ better than, the constant penalty model while generally using fewer effective degrees of freedom.

A comparison of the coverage probabilities based on the one-step jackknife and one-step bootstrap variance estimators are displayed in Figures 3.6 and 3.7 for generating intensities $f_1(t)$ and $f_2(t)$ respectively. The coverage probabilities for generating intensity $f_1(t)$ are for the most part on target, particularly for larger numbers of panels ($e = 80$). The jackknife and bootstrap intervals tend to perform similarly with the jackknife having the slight advantage in this case. For $f_2(t)$ the coverage probabilities are fairly poor when $e = 40$ which can be largely attributed to the large bias. As the number of panels increases however the performance of the coverage probabilities improves and for the case with $I = 50$ individuals and $e = 80$ coverage probabilities approach the nominal level in general. Note that the coverage is going to be poor for very steep increases or decreases in the presence of even small amounts of bias which can be seen clearly in this plot.

Simulations were also run at various levels of the overdispersion parameter but only the case of $\tau = 0.3$ is presented here since the spline estimates of the intensity are fairly robust to changes in this parameter. Of interest however is that the bootstrap intervals for this nuisance parameter seem to provide better performance, in terms of coverage, than the
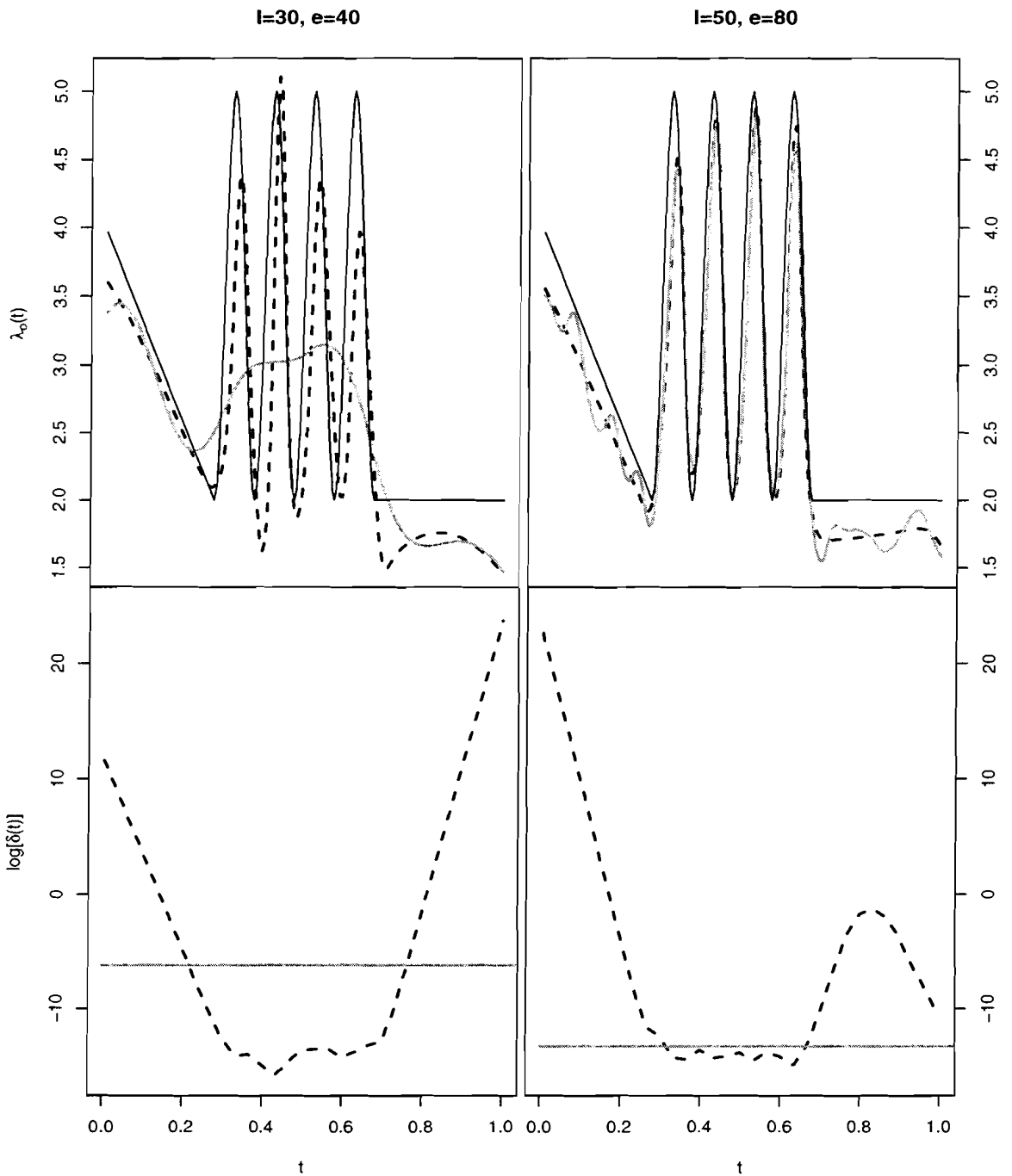
Figure 3.4: [Top Panels] Plots of estimates for one sample run under two scenarios (left column $I = 30$ and $e = 40$; right column $I = 50$ and $e = 40$) for both the adaptive spline (dashed lines) and constant penalty spline (solid gray lines) models overlaying the true value (solid lines) of the intensity $f_2(t)$. [Bottom Panels] Estimates of the log of the penalty spline $\log[\delta(t)]$ (dashed lines) and the log of the estimate under the assumption that $\delta(t) = \delta$ (solid gray lines) for both scenarios.
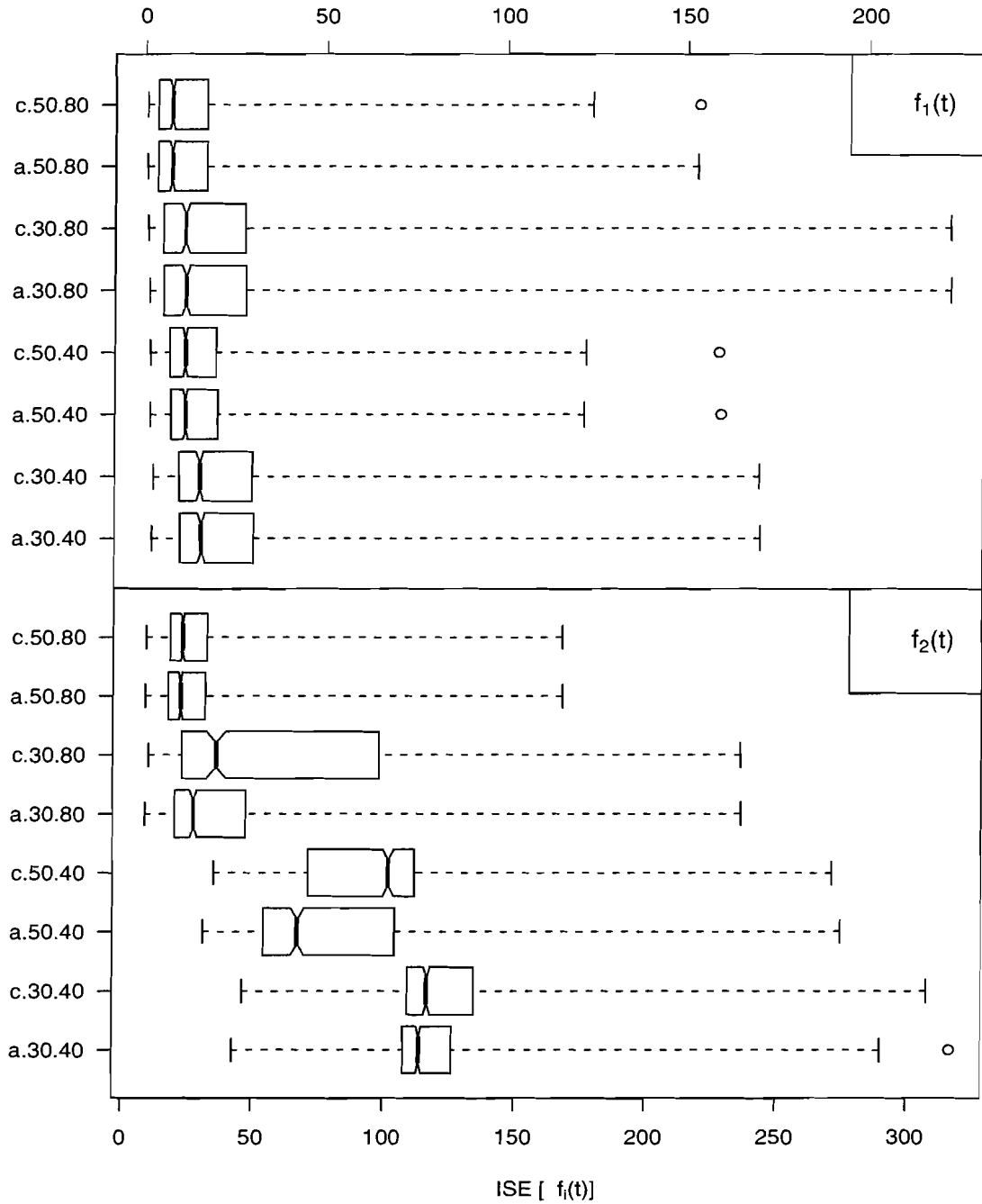
Figure 3.5: Box plots of the integrated squared error, $ISE[f_i(t)] = \int_0^1 [f_i(t) - \tilde{f}_i(t)]^2 \, dt$, for both of the generating functions; $f_1(t)$ in the top panel and $f_2(t)$ in the bottom panel. Results are displayed for both the adaptive spline (a) and constant penalty spline (c) models, all levels of number of individuals $I = 30, 50$ and numbers of panels $e = 40, 80$. Labels are of the form $a/c.I.e$.
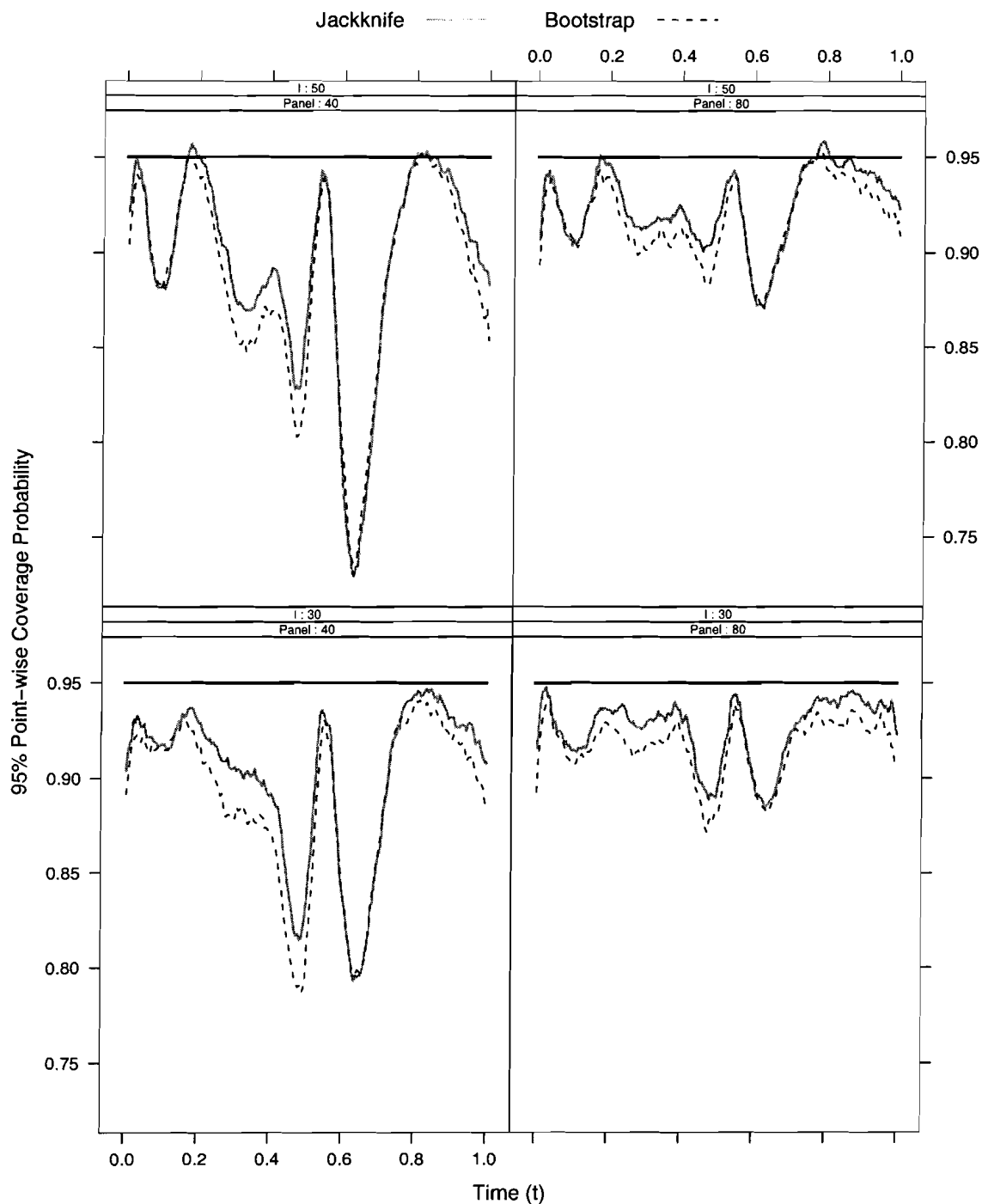
Figure 3.6: Point-wise coverage probabilities for 95% confidence intervals of the intensity $f_1(t)$ constructed from the one-step jackknife and one-step bootstrap procedures from the simulated data analysis for the adaptive model. All levels of numbers of individuals $I = 30, 50$ and numbers of panels $e = 40, 80$ are displayed.
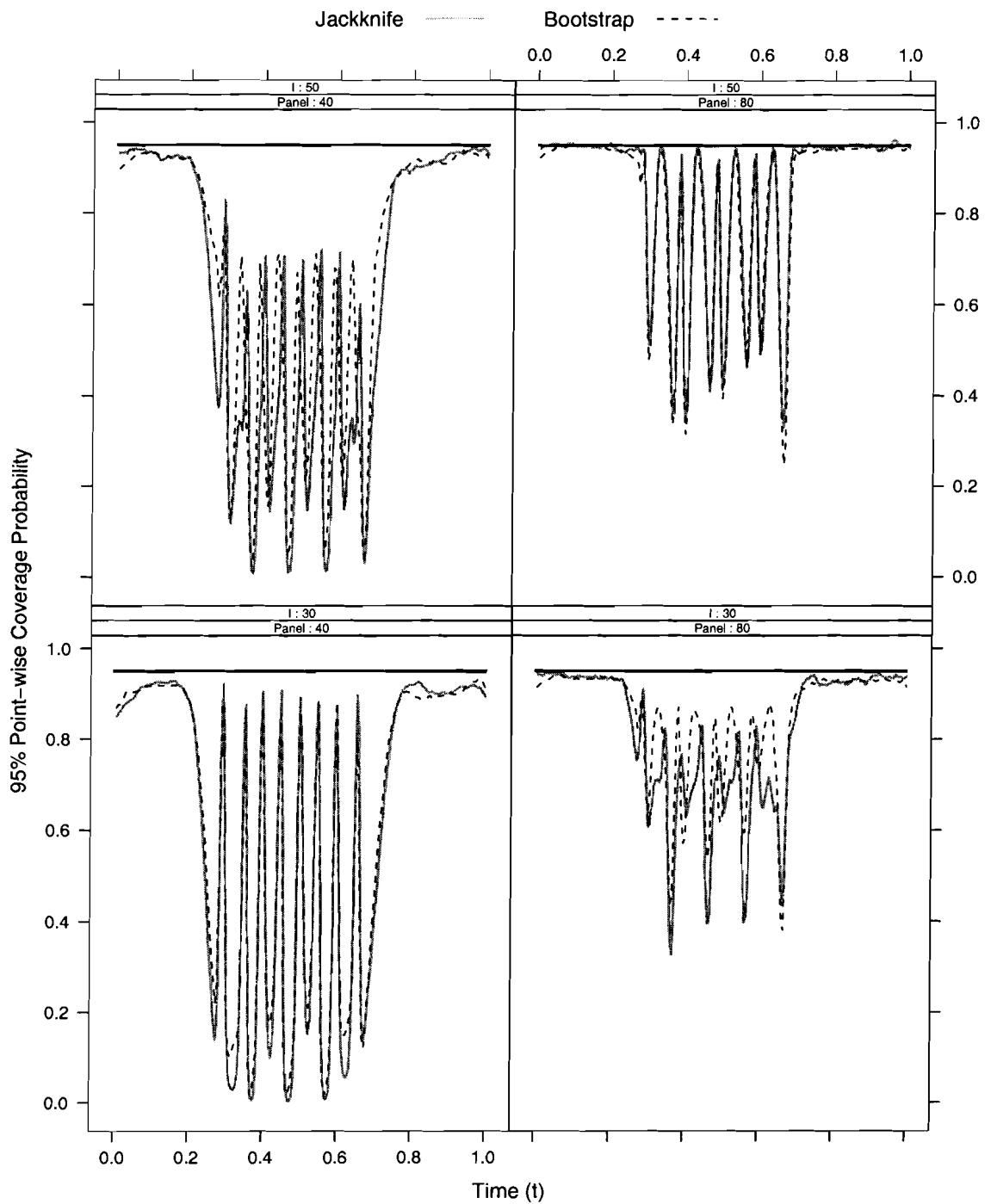
Figure 3.7: Point-wise coverage probabilities for 95% confidence intervals of the intensity $f_2(t)$ constructed from the one-step jackknife and one-step bootstrap procedures from the simulated data analysis for the adaptive model. All levels of numbers of individuals $I =$ 30, 50 and numbers of panels $e =$ 40, 80 are displayed.

jackknife interval, which tend to be too large. This is likely the result of the skewness of the distribution of the overdispersion estimator which the bootstrap approach is better able to accommodate.

3.6 Cherry Bark Tortrix Experiment

This experiment was designed to test the effectiveness of pheromones in disrupting the mating patterns of the cherry bark tortrix moth (*Enarmonia fomosana*). It was postulated that the release of the pheromone would confuse males from locating and mating with females in the trees thereby hindering reproduction. Twenty cherry trees were outfitted with pheromone-baited traps, attached in similar locations in each tree. All of the trees were fitted with scent dispensers, but ten were selected at random (treatment trees) and their dispensers were filled with female pheromones; the remaining ten were used as controls. Approximately once a week for 19 weeks the traps were emptied and the number of moths caught were counted. The baits were refreshed at three-week intervals.

The model presented in Section 3.1 was fit to the data using both an adaptive spline penalty $\delta(t)$ with five effective degrees of freedom as well as with a constant penalty for comparison. Increasing the effective degrees of freedom provides little differences in the estimated fits. Estimates for the treatment effect, 3.321 (s.e. 0.343), indicate the the treatment is effective and overdispersion parameter, 0.662 (s.e. 0.114), that heterogeneity is present; these estimates are identical for the two fits. Figure 3.8 displays plots of the baseline intensity for both the adaptive spline, the solid line, and the constant penalty model, the dashed line, along with their 95% point-wise confidence intervals as well as plots of $\log[\delta(t)]$ for both models. These plots are also overlaid with empirical step function estimates as described in Lawless and Zhan (1998)). These authors model the intensity function as a proportional intensity model with a random frailty term and with the baseline intensity being piecewise constant. This figure shows that moths mate in two bimodal waves, the first corresponding to moths that over-winter in a "instar" pupal state and are ready to mate earlier than moths that hatch in the early summer represented by the second wave. The model with the constant penalty seems to provide a reasonably flexible fit in this situation with regard to the differences in the three fits. The adaptive penalty allows the intensity estimate somewhat more flexibility for the first wave and less in the last third of the time period where the intensity approaches zero. This is clear in the bottom panel, a plot of the

estimate of $\log[\delta(t)]$ overlaying the log of the estimate of the penalty when held constant, $\delta(t) = \delta$. All three methods are in general agreement.

3.7 Discussion

This chapter presents a method of incorporating both adaptive and regular penalized spline smoothing into the quasi-likelihood analysis of recurrent event panel data and compares these two approaches under various scenarios. The adaptive approach provides some gains when there is sharp curvature or abrupt changes between regions of high oscillation to regions that are fairly flat. In many of the cases considered however the use of a constant penalty provides a fairly good approximation to the underlying functional form. In situations where it is believed that sharp changes in the derivative of the intensity function exist then the added complexity may be beneficial. This approach might be useful for monitoring extreme events in climate change, for example, where accurate definitions of peaks are important or in extensions for spatio-temporal analysis where cliffs in the intensity may occur in the spatial domain.

An alternative approach to adaptive penalized splines for modeling mixed NHPP investigated here would be to optimally select the number and position of the knots, breakpoints between connecting polynomial sections, commonly refereed to as the free knot approach (Jupp, 1978; Mao and Zhao, 2003). This approach, even in the normal regression context, requires far more computationally complex methods such as insertion-deletion (Kooperberg et al., 1997) algorithms or stochastic optimization (Pittman, 2002) because of many local optima. Hybrid approaches that combine careful knot selection and penalties have also been studied (Luo and Wahba, 1997; Lindstrom, 1999). The development of such methods may be useful for the analysis of massive datasets where the overparameterized basis required for the penalty methods may become prohibitive and lower dimensional adaptive analogue to the penalized approach may be required.
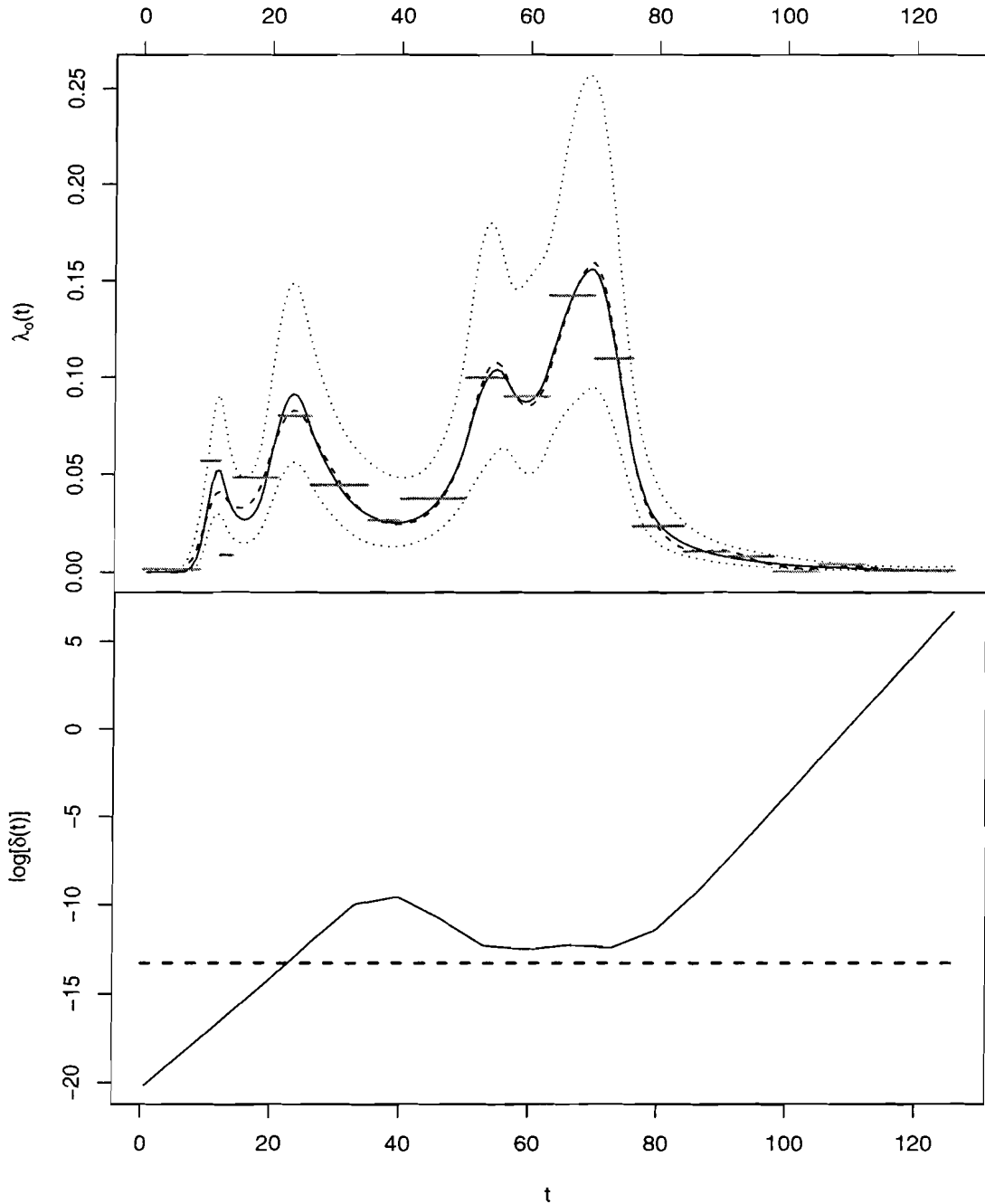
Figure 3.8: [Top panel] Estimated overall baseline intensity $\lambda_0(t)$ (———) with corresponding 95% point-wise confidence intervals ($\cdots\cdots$) overlaid with the estimates under the assumption that $\delta(t) = \delta$ (– – ––) along with piecewise constant estimates. [Bottom panel] Estimated log of the penalty spline $\log[\delta(t)]$ (———) as well as the log of the estimate under the assumption that $\delta(t) = \delta$ (– – ––).

Chapter 4

Clustered Mixed NHPP Spline Model

Basic methods for the analysis of panel count data are fairly well-established; Cook and Lawless (2002) provide an excellent review. A common and flexible approach uses an assumption of a proportional intensities model — the intensity function governing the rate of occurrence of events has the form $\lambda(t) = \lambda_0(t) \exp(\mathbf{x}^T \beta)$ where λ_0 is the so-called 'baseline intensity function', $\mathbf{x}$ is a vector of covariates and β the regression coefficients. Parametric forms for the baseline intensity function may be used; for example, the Weibull (Lawless, 1987a) is popular. Semi-parametric methods have also been employed: Lawless and Zhan (1998) and Balshaw and Dean (2002) adopt a piece-wise constant baseline intensity.

Several recent applications by the authors exhibit a number of features for which methods are currently unavailable. In this chapter, we extend the current methodology in two main directions within a unified framework. The extensions emerge out of, and impact, collaborative work in medicine and biological sciences which drive our conceptual positioning. We model heterogeneity through discrete mixtures of mixed Poisson distributions, accommodating clusters as well as individual frailties. The discrete mixtures can be viewed as hidden sub-populations of individuals whose counting processes are being generated via an unknown group-specific intensity process. Indeed, for several applications we have considered there is scientific reasoning that hypothesizes such hidden clusters generating the counts which may react differentially to treatments. The motivating application described in this chapter is an experiment to test the effectiveness of pheromones in disrupting the

mating patterns of the cherry bark tortrix moth. Moths mate in the summer in two distinct phases: the first corresponds to those moths that overwinter in a developmental pupal state, while the second correspond to those that overwinter as eggs. Mature moths from these sub-populations cannot be distinguished, but it is of interest to determine whether the treatment has differential effects on them. More broadly, the authors have used the methods herein on analysis of medical data where there is evidence of differential response to a treatment with constant and with decreasing response counts over time for distinct sub-populations. In addition, the focus here is on joint estimation of the baseline intensity and the effects of covariates, both of which are modeled as smooth functions using splines. Hence the simple case of a single-cluster model derived herein also provides a flexible, useful approach for the analysis of panel count data. Note that outside of the mixture context and for data arising from continuous followup, splines have been used previously for the analysis of longitudinal data and have been shown to offer great flexibility for such modeling (see, for example Lin and Zhang (1999) and Ruppert et al. (2003)).

With an emphasis in the literature on quasi-likelihood methods for analysis of count data, and the benefits and flexibility of such estimation schemes, the approach taken here imposes fairly weak distributional assumptions to provide robustness to model misspecification. Since we focus on methods for the analysis of panel data, this brings the need to handle additional features such as missing data over panels, aggregation of counts over panels, and non-identical times of follow-up for individuals or clusters; all of these are accommodated by the methodology developed.

In Section 4.1 we develop the clustered mixed nonhomogeneous process spline model. Section 4.2 develops an algorithm for quasi-likelihood estimation utilizing a scheme analogous to EM (Estimation-Maximization; McLachlan and Krishnan, 1997) which is a special case of the Expectation-Solution (ES) algorithm of Rosen et al. (2000). Inference and a brief discussion on identifiability is considered in Section 4.3; robust variance estimates are provided. Section 4.4 applies the proposed methodology to several data sets beginning with the motivating example. Simulation studies in Section 4.5 probe how well components are recovered when there is high variability and few panels, and how well variance estimates and their small sample distributions perform in their ability to provide confidence interval coverage. The chapter closes with a discussion of the methods and the rich opportunities for future work in recurrent event analysis.

4.1 Spline Model for Clustered Overdispersed Longitudinal Counts

Let $\{N_i(t), t \in [0, T_{ie_i})\}$ be a counting process governing the number of events experienced by individual $i = 1, \ldots, I$ and assume that for each individual the process, $N_i(t) = \sum_{g=1}^{G} z_{gi} C_{gi}(t)$, arises from one of several subprocesses $C_{gi}(t)$ where $z_{gi} = 1\{i \in \mathcal{G}_g\}$; $1\{\cdot\}$ being the indicator function and $\mathcal{G}_g$, $g = 1, \ldots, G$ denoting the unobservable cluster from which the subprocesses are generated. Given a mixing distribution m_g, each $C_{gi}(t)$ is assumed to be a nonhomogeneous Poisson process (NHPP) with conditional intensity denoted by $\lambda_{gi}(t|\nu_{gi})$ where ν_{gi} are within cluster random effects, $\nu_{gi} \overset{i.i.d.}{\sim} m_g$, that account for possible within cluster heterogeneity. Let $\Lambda(\cdot)$ represent the cumulative intensity function, $\Lambda(t) = \int_0^t \lambda(u) du$. Data for each individual are assumed to be collected at follow-up times $0 \leq T_{i1} < T_{i2} < \cdots < T_{ie_i}$; let $N_{ij} = N_i[T_{i(j-1)}, T_{ij})$, the number of events experienced by individual i between $T_{i(j-1)}$ and T_{ij}, $j = 1, \ldots, e_i$ and define $T_{io} = 0$.

The conditional intensity governing the generation of counts $N_{ij} = N_i[T_{i(j-1)}, T_{ij})$ for individual i, given group membership, $z_{gi} = 1$, and the individual-specific sub-group frailty term ν_{gi}, is $\nu_{gi} \lambda_{gi}(t) \mathbf{1}\{t \in [T_{i(j-1)}, T_{ij})\}$ where

$$\lambda_{gi}(t) = \lambda_g(t|\mathbf{x}_i) = \exp\left\{\gamma_{g0}(t) + \sum_{q=1}^{p} x_{iq}\,\gamma_{gq}(t)\right\} \tag{4.1}$$

with the x_{iq}'s representing covariate information for the ith individual, and γ_{gq} being group-specific spline effects, $q = 0, \ldots, p$, assumed to have absolutely continuous second derivatives. The term $\gamma_{g0}(t)$ reflects what could be called in this context the logarithm of the baseline spline function. Several possible spline basis functions could be effectively utilized; cubic B-splines are used here for their numerical stability (de Boor, 1978) yielding the representations:

$$\gamma_{gq}(t) = \sum_{h=1}^{4+k} \psi_{gqh} B_{h-4}(t) = \mathbf{b}(t)^T \boldsymbol{\psi}_{gq}$$

where $\boldsymbol{\psi}_{gq}$ is the vector of B-spline coefficients $\psi_{gq\cdot}$, $q = 0, \ldots, p$, corresponding to the qth spline effect in the gth group, $\mathbf{b}(t)$ is the vector of cubic B-spline basis functions $B_{\cdot}(t)$, and k is the number of breakpoints or knots defining the basis. The unconditional distribution of the vector of observations $\mathbf{N}_i = (N_{i1}, \ldots, N_{ie_i})^T$ for individual i is $P(\mathbf{N}_i) = \sum_{g=1}^{G} p_g P_g(\mathbf{N}_i)$ with expectation $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ie_i})^T$ where $p_g = E[\mathbf{1}\{i \in \mathcal{G}_g\}] =$

$P(i \in \mathcal{G}_g)$, $P_g(\mathbf{N}_i) = \int \prod_{j=1}^{e_i} P_g(N_{ij}|\nu_{gi})m_g(\nu_{gi})\mathrm{d}\nu_{gi}$, $P_g(N_{ij}|\nu_{gi})$ is distributed as Poisson with mean $\nu_{gi}\mu_{gij}$, $\mu_{gij} = \Lambda_{gi}(T_{ij}) - \Lambda_{gi}(T_{i(j-1)})$, $\mu_{ij} = \sum_{g=1}^{G} p_g \mu_{gij}$ and Λ_{gi} is the cumulative intensity corresponding to λ_{gi}. The random multiplicative components ν_{gi} model within-cluster overdispersion. We assume that ν_{gi} has mean 1, without loss of generality, and denote its variance as τ_g. As scaling terms, they also ensure that clustering identifies subgroups with intensities of varying functional form rather than simply differences in magnitude.

The number and position of the knots defining the B-spline basis have a strong impact on the possible shapes that the spline can approximate and hence require careful consideration. The approach taken here is to use penalized splines where a large number of knots are used, such as all unique panel midpoints $\{(T_{ij} + T_{i(j-1)})/2\}$, and a penalty term is added to control the amount of fitting of the resulting over-parameterized model. In the case of a few coincident panels over individuals, the naive method of using a few fixed knots usually works sufficiently well in practice (Nielsen and Dean, 2005). Alternatively, one could use a small number of well-chosen knots or a "free knot" approach (Jupp, 1978; Mao and Zhao, 2003). The advantage of such an approach is that an adaptive spline basis can adjust to local changes in variability of the underlying function resulting in better approximations in cases where such heterogeneity is present. These adaptive methods do require more intensive computational algorithms such as knot insertion-deletion (Kooperberg et al., 1997; Hansen and Kooperberg, 2002). Adaptive splines are particularly useful when event times are recorded. For the analysis of panel data, the rough time grid over which the panel process is observed can dampen the gains in flexibility afforded by such an approach since this aggregation of the data into panels can effectively mask any sharp changes that may exist in the intensity.

There are advantages to the form of the "link" function chosen in (4.1) above. It reduces to the usual log linear form when the observations consist of end-of-follow-up counts. It also ensures that $\lambda_{gi}(\cdot)$ remains positive providing the required monotonicity in the cumulative intensity. However, this representation results in a non-analytic cumulative intensity function. In practice, even simple approximations to the integral perform well in this context of panel data analysis. Here, the integral is approximated using a 4-point Legendre-Gauss quadrature rule which allows exact computation of the integral of up to a 7-degree polynomial for each panel. Such an approximation seems sufficient for applications that are encountered in practice. Other possible "links" for the intensity function (4.1) could be

envisioned. For example, with the identity link, the resulting model would be additive instead of multiplicative which may aid in the interpretation of the estimated effects in some situations. The cumulative intensity is also analytic in this case and if one were to ignore the need for constraints on the spline coefficients to impose monotonicity in the cumulative intensities, inference using the estimating equations developed in the next section would be straightforward. In fact, such a simple approach will often yield satisfactory results provided there are a moderate number of events per individual and that these events do not oscillate too rapidly from panel to panel.

4.2   Estimation

Let $\theta = (\phi^T, \mathbf{p}^T)^T$ denote all the parameters of interest where $\phi = (\phi_1, \ldots, \phi_G)^T$, $\phi_g = (\psi_g^T, \tau_g)^T$, the parameters of the gth subpopulation, $\psi_g = (\psi_{g0}^T, \psi_{g1}^T, \ldots, \psi_{gp}^T)^T$ and $\mathbf{p} = (p_1, \ldots, p_G)$.

Case 1: G fixed

Given initial values for the component indicators z_{gi}, estimates of ψ_g are obtained by solving penalized quasi-likelihood equations (Breslow and Clayton, 1993)

$$\mathbf{g}_{\psi_g} = \sum_{i=1}^{I} z_{gi} \mathbf{D}_{gi}^T \mathbf{V}_{gi}^{-1} (\mathbf{N}_i - \boldsymbol{\mu}_{gi}) - (\text{diag}\{\boldsymbol{\delta}_g\} \otimes \mathbf{P}) \psi_g = 0, \quad g = 1, \ldots, G \qquad (4.2)$$

where $\boldsymbol{\mu}_{gi} = (\mu_{gi1}, \ldots, \mu_{gie_i})^T$, $\mathbf{D}_{gi} = \partial \boldsymbol{\mu}_{gi} / \partial \psi_g^T$, $\mathbf{V}_{gi}^{-1} = \text{diag}\{1/\mu_{gij}\}_{e_i \times e_i} - \tau_g \mathbf{J}_{e_i}/(1 + \tau_g \mu_{gi+})$, $\mathbf{J}_{e_i} = \{1\}_{e_i \times e_i}$, $\mu_{gi+} = \sum_{j=1}^{e_i} \mu_{gij}$, $\boldsymbol{\delta}_g = (\delta_{g0}, \delta_{g1}, \ldots, \delta_{gp})^T$, the smoothing parameters for the baseline and spline covariate effects, and

$$\mathbf{P} = \int_0^{\max\{T_{ie_i}\}} \mathbf{b}'(t) \mathbf{b}'(t)^T dt, \qquad (4.3)$$

the penalty matrix, which shrinks the spline covariate effects to a constant as $\delta_{gq} \to \infty$ or results in an interpolating spline as $\delta_{gq} \to 0$. Hence, if $\delta_{gq} \to \infty$ so that $\gamma_{gq}(t) \to \gamma_{gq}$, for all $q \in \{1, \ldots, p\}$, then $\lambda_{gi}(t)$ is of the proportional intensity form; if, in addition, $\delta_{g0} \to \infty$, then the model for the gth component becomes overdispersed Poisson regression. Note that (4.2) can be derived equivalently using penalized negative binomial maximum likelihood under the assumption of gamma distributed frailties.

The vector of smoothing parameters $\boldsymbol{\delta}_g$ are updated using the so-called GML estimator of Gu (2002),

$$\tilde{\delta}_{gq} = \frac{\text{tr}\,[\mathbf{R}_{gqq}]}{\tilde{\psi}_{gq}^T \mathbf{P} \tilde{\psi}_{gq}}, \tag{4.4}$$

where $\mathbf{R}_{gqq}$ are the $(k+4) \times (k+4)$ block diagonal elements of

$$\mathbf{R}_g = \{\mathbf{R}_{gsv}\} = \left[\sum_{i=1}^{I} z_{gi}\mathbf{D}_{gi}^T\mathbf{V}_{gi}^{-1}\mathbf{D}_{gi} + \text{diag}\{\boldsymbol{\delta}_g\} \otimes \mathbf{P}\right]^{-1}\left(\sum_{i=1}^{I} z_{gi}\mathbf{D}_{gi}^T\mathbf{V}_{gi}^{-1}\mathbf{D}_{gi}\right), \tag{4.5}$$

$s, v = 1, \ldots, p$. This is a restricted maximum likelihood estimator (REML) for $\boldsymbol{\delta}_g$. The update of the smoothing parameter on the right-hand side of (4.4) is an empirical Bayes estimator, evaluated at the current value of the estimates. It is a ratio of the overall residual variation within subgroup, which under the model is one, to the estimated variance of the contrast induced by the penalty $\mathbf{P}$ on ψ_{gq}. This contrast is expressible as $\mathbf{U}^-\psi_{gq}$ where $\mathbf{U}^-$ corresponds to the non-zero rows of the matrix $\mathbf{U} = \mathbf{E}^{\frac{1}{2}}\mathbf{S}^T$, $\mathbf{E}^{\frac{1}{2}}$ being the diagonal of the square root of the eigenvalues, and $\mathbf{S}$, the normalized eigenvectors, of $\mathbf{P}$.

To estimate τ_g, we use restricted maximum likelihood under the assumption of normally distributed residuals, $N_{ij} - \mu_{gij}$, leading to the estimating equations

$$g_{\tau_g} = \sum_{i=1}^{I} z_{gi}\frac{(N_{i+} - \mu_{gi+})^2 - \mu_{gi+}(1 + \tau_g\mu_{gi+}) + r_{gi}}{(1 + \tau_g\mu_{gi+})^2} = 0,\ g = 1, \ldots, G, \tag{4.6}$$

where

$$r_{gi} = \text{tr}\left[z_{gi}\left(\sum_{i=1}^{I} z_{gi}\mathbf{D}_{gi}^T\mathbf{V}_{gi}^{-1}\mathbf{D}_{gi} + \text{diag}\{\boldsymbol{\delta}_g\} \otimes \mathbf{P}\right)^{-1}\mathbf{D}_{gi}^T\mathbf{V}_{gi}^{-1}\mathbf{D}_{gi}\right]$$

and $N_{i+} = \sum_{j=1}^{e_i} N_{ij}$. The estimating equation (4.6) is the so-called 'pseudo-likelihood' equation of Davidian and Carroll (1987) with the inclusion of a first-order correction term, r_{gi}.

To update z_{gi}, we adapt a scheme proposed by Rosen et al. (2000) called the ES algorithm, which uses the posterior estimator

$$z_{gi}^* = E[\mathbf{1}\{i \in \mathcal{G}_g\}\,|\mathbf{N}_i] = \frac{p_g P_g(\mathbf{N}_i)}{\sum_{g=1}^{G} p_g P_g(\mathbf{N}_i)},$$

where $P_g(\mathbf{N}_i)$ is computed assuming $m_g(\nu_{gi}) \overset{iid}{\sim} \Gamma(1/\tau_g, \tau_g)$. Lastly, the estimates of the probability of group membership are computed by solving

$$g_{p_g} = \sum_{i=1}^{I}\left(\frac{z_{gi}^*}{p_g} - \frac{z_{Gi}^*}{p_G}\right) = 0,\ g = 1, \ldots, G-1, \tag{4.7}$$

which has an explicit solution $\tilde{p}_g = \sum_{i=1}^{I} z_{gi}^*/I$, $g = 1, \ldots, G - 1$. This process of quasi-likelihood estimation and imputation with empirical posterior Bayesian estimates for the unknown component membership quantities is iterated until convergence. The ES algorithm is not guaranteed to converge from arbitrary initial configurations like the EM algorithm; however, if the estimating scheme does converge then the method will result in unbiased estimating equations (Rosen et al., 2000). In practice it has been found that the ES algorithm is quite robust to initial parameter specification and converges relatively quickly for fair starting values.

Case 2: G unknown

To obtain an estimate of the number of components we develop an algorithm leveraging theory from non-parametric maximum likelihood (NPML) estimation which states that for a finite number of components $G < I$ the log-likelihood of the finite mixture distribution will be convex (Lindsay, 1983). It has been found to be efficient and reliable for a variety of applications considered.

Algorithm:

1) Fit a One-Component Model. Set $c = 1$ and compute $\tilde{\phi}_1$, the estimates from the one-component model, by solving equations (4.2) and (4.6) with $z_{1i} = 1$, $i = 1, \ldots, I$. Evaluate $P_1(\mathbf{N}_i|\tilde{\phi}_1)$ and set f_{1i} to $P_1(\mathbf{N}_i|\tilde{\phi}_1)$, $i = 1, \ldots, I$.

2) Obtain Initial Values for the $(c + 1)$-Component Model. Perform a univariate grid search over α ($0 < \alpha < 1$) to obtain a new mixing weight α^* and component parameters $\tilde{\phi}_{c+1}(\alpha^*)$ which maximize

$$\sum_{i=1}^{N} \log \left\{ (1 - \alpha)f_{ci} + \alpha P_{c+1}(\mathbf{N}_i|\phi_{c+1}) \right\} - \frac{1}{2} \sum_{q=0}^{p} \delta_{(c+1)q} \ \psi_{(c+1)q}^{T} \mathbf{P} \psi_{(c+1)q}$$

where P_{c+1} is the postulated $(c + 1)$-component of the model. Set $f_{c+1,i} = (1 - \alpha^*)f_{ci} + \alpha^* P_{c+1}(\mathbf{N}_i|\tilde{\phi}_{c+1}(\alpha^*))$. These values are used as initial values for estimating the parameters of the $(c + 1)$-component mixture.

4) Fit a $(c + 1)$-Component Model. With the number of components set as $c + 1$, the parameters in the $(c+1)$-component mixture model are estimated using the ES algorithm

detailed in Case 1 with initial values from the previous step. Set f_{gi} to $P_g(\mathbf{N}_i | \tilde{\boldsymbol{\phi}}_g)$, $g = 1, \ldots, c+1$ and $i = 1, \ldots, I$.

5) Stopping Rule. When $p_{c+1} \to 0$, set the number of components to $G = c$, update the c-component model parameters using the ES algorithm and exit; else return to step 2 and set $c = c + 1$.

In general, for the application considered here and several others not presented, the algorithm tends to select models with few components, less than four and in some cases only one component.

4.3 Inference

Inference here is developed for the case of G fixed and under usual assumptions that the design matrix $\mathbf{X} = \{x_{iq}\}_{I \times p}$ is of full rank and the Jacobian matrix $\partial \mathbf{g}_\theta / \partial \boldsymbol{\theta}^T$ is non-singular. These assumptions are discussed in 4.3.1 below. With estimates $\tilde{\boldsymbol{\theta}}$ obtained by solving $\mathbf{g}_\theta = (\mathbf{g}_{\phi_1}^T, \ldots, \mathbf{g}_{\phi_G}^T, \mathbf{g}_\mathbf{p}^T)^T = \mathbf{0}$ where $\mathbf{g}_{\phi_g} = (\mathbf{g}_{\psi_g}^T, g_{\tau_g})^T$ it can be shown (White, 1982) that, $\sqrt{I}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\tilde{\boldsymbol{\Sigma}} = E \left[\frac{\partial \mathbf{g}_\theta}{\partial \boldsymbol{\theta}^T} \right]^{-1} \left(\sum_{i=1}^{I} \mathbf{g}_{i,\theta} \mathbf{g}_{i,\theta}^T \right) E \left[\frac{\partial \mathbf{g}_\theta^T}{\partial \boldsymbol{\theta}} \right]^{-1} \Bigg|_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}}.$$

This is the "sandwich" variance estimate (Liang and Zeger, 1986) where $\mathbf{g}_{i,\theta}$ is the contribution of the ith individual to the estimating function $\mathbf{g}_\theta$.

An asymptotically equivalent estimator of $\boldsymbol{\Sigma}$ (Lipsitz et al., 1994) can be constructed using a jackknife approach to improve its finite sample properties. Let $\tilde{\boldsymbol{\theta}}_{(-i)}$ be the estimates obtained by the method described in the previous section with the ith individual removed, $i = 1, \ldots, I$. The jackknife estimate of $\boldsymbol{\Sigma}$ would then be given by

$$\tilde{\boldsymbol{\Sigma}}_* = \frac{\sum_{i=1}^{I} e_i - r}{\sum_{i=1}^{I} e_i} \sum_{i=1}^{I} \left(\tilde{\boldsymbol{\theta}}_{(-i)} - \tilde{\boldsymbol{\theta}} \right) \left(\tilde{\boldsymbol{\theta}}_{(-i)} - \tilde{\boldsymbol{\theta}} \right)^T \tag{4.8}$$

where $r = \sum_{g=1}^{G} \text{tr}\{\mathbf{R}_g\} + 2G - 1$ with $\mathbf{R}_g$ given in (4.5). A simple one-step approximation to (4.8) can be obtained by taking one Newton step from the overall estimate $\tilde{\boldsymbol{\theta}}$ as

$$\tilde{\boldsymbol{\theta}}_{(-i)} = \tilde{\boldsymbol{\theta}} + E \left[\sum_{j=1,i \neq j}^{I} \frac{\partial \mathbf{g}_{j,\theta}}{\partial \boldsymbol{\theta}^T} \right]^{-1} \left(\sum_{j=1,i \neq j}^{I} \mathbf{g}_{j,\theta} \right) \Bigg|_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}} = \tilde{\boldsymbol{\theta}} - E \left[\sum_{j=1,i \neq j}^{I} \frac{\partial \mathbf{g}_{j,\theta}}{\partial \boldsymbol{\theta}^T} \right]^{-1} \mathbf{g}_{i,\theta} \Bigg|_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}},$$

which results in the expression

$$\frac{\sum_{i=1}^{I} e_i - r}{\sum_{i=1}^{I} e_i} \sum_{i=1}^{I} \left\{ E \left[\sum_{j=1,i\neq j}^{I} \frac{\partial \mathbf{g}_{j.\theta}}{\partial \theta^T} \right]^{-1} (\mathbf{g}_{i.\theta} \mathbf{g}_{i.\theta}^T) E \left[\sum_{j=1,i\neq j}^{I} \frac{\partial \mathbf{g}_{j.\theta}^T}{\partial \theta} \right]^{-1} \right\} \Bigg|_{\theta=\tilde{\theta}}. \qquad (4.9)$$

Very little computational overhead is incurred in computing $\widetilde{\Sigma}_*$ using (4.9) since it is easily obtained from values calculated while solving for $\tilde{\theta}$. The fully iterated jackknife (4.8) is preferable in general; however, it is useful to compute this simple one-step approximation as a comparison since it should be fairly close to the fully iterated form. If this is not the case this might suggest that the solution $\tilde{\theta}$ is highly affected by a single individual or that the objective surface at $\tilde{\theta}$ is not 'elliptical', an indication that inference procedures may be suspect. It should also be noted that the jackknife estimates themselves, $\tilde{\theta}_{(-i)}$'s, are useful for diagnostics.

4.3.1 Identifiability

For the simpler case of finite mixtures of Poisson distributions with covariates, Wang et al. (1996) show that a sufficient condition for identifiability is that the design matrix, $\mathbf{X}$, is of full rank. Assessment of local identifiability is also useful here and is common in the study of latent class models. A numerical check is performed at convergence of the algorithm presented in Section 4.2 by computing the eigenvalues of $\partial \mathbf{g}_\theta / \partial \theta^T$ evaluated at $\tilde{\theta}$ and verifying that they are all non-zero. If this is the case then the model is locally identifiable (Bandeen-Roche et al., 1997). In addition, note that the algorithm protects against boundary cases by removing the gth component if $p_g \to 0$.

4.4 Illustrations

4.4.1 Cherry Bark Tortrix Moth Study

This experiment was designed to test the effectiveness of pheromones in disrupting the mating patterns of the cherry bark tortrix moth (*Enarmonia fomosana*). The pheromone in question was considered to be competitive with caged virgin females in luring males into traps. It was postulated that the release of the pheromone would confuse males from locating and mating with females in the trees. It was also postulated that the appearance of mate-seeking males over the summer would arise from at least two types of hatching,

representing hidden components. This is discussed later. Twenty cherry trees were outfitted with pheromone-baited traps, attached in similar locations in each tree. All of the trees were fitted with scent dispensers, but ten were selected at random (treatment trees) and their dispensers were filled with female pheromones; the remaining ten were used as controls. Approximately once a week, over a 19 week period, the traps were emptied and the number of moths caught were counted. The baits were refreshed at three-week intervals.

The model derived in the previous section was fit to this data resulting in the identification of a three-component model; the treatment group was assigned to the baseline. The estimates of the smoothing parameters for the spline covariate effects diverge to infinity resulting in the sub-group treatment splines being constants, $\gamma_{g1}(t) = \gamma_{g1}$, so that each component has a proportional intensity form. Table 4.1 presents parameter estimates for the treatment effects γ_{g1}, overdispersion parameters τ_g, group membership probabilities p_g and their associated estimated jackknife standard errors. The covariate estimates for all components are large and significant. Note that the estimate of the covariate effect for component one is about half that for the other two components indicating a differential treatment effect among the groups. The estimates of the within-group dispersion parameters τ_g are significant at the 5% level so all the subpopulations are exhibiting heterogeneity. Each group is well represented in the sample with estimated subpopulation membership probabilities between 30%-40%. Residual plots and jackknife diagnostics showed no glaring discrepancies in model fit.

Figure 4.1 a) and b) show the plots of the estimated overall baseline intensities, $\lambda_0(t) = \sum_{g=1}^{G} p_g \lambda_{g0}(t)$, and cumulative baseline intensities, $\Lambda_0(t) = \sum_{g=1}^{G} p_g \Lambda_{g0}(t)$, along with their associated 95% jackknife point-wise confidence bands. The form indicates that the cherry bark tortrix moths mate in two waves. The first wave in the overall intensity corresponds to maturity of the so-called "instar" moth larvae that overwinter in a developmental pupal state. These are ready to breed earlier than those that overwinter as eggs, which generally hatch in early summer, and contribute mostly to the second wave of breeding between 50 to 70 days into the summer. The second wave itself appears to be generated by two tight waves occurring around a fortnight apart. These observations are distinctly evident in the within-subpopulation baseline and cumulative baseline intensity estimates displayed in Figure 4.1 c) and d). Of particular interest is the intensity of group 1 that has a pronounced early first wave suggesting that this group represents predominantly the "instar" moths. Further, there is a substantial difference in the treatment effects for group 1 and the other two

		Estimate (s. e.)
One-Component Model		
Treatment	γ_{11}	3.321 (0.384)
Dispersion	τ_1	0.662 (0.205)
Three-component Model		
Treatment	γ_{11}	1.710 (0.771)
	γ_{21}	3.642 (0.669)
	γ_{31}	3.697 (0.474)
Dispersion	τ_1	0.681 (0.299)
	τ_2	0.676 (0.185)
	τ_3	0.559 (0.221)
Probability	p_1	0.314 (0.134)
	p_2	0.311 (0.144)
	p_3	0.375 (0.143)

Table 4.1: Estimates and estimated jackknife standard errors of the treatment effects γ_{g1}, overdispersion parameters τ_g and the probabilities of group membership p_g from the fit of the one- and three-component mixture to the cherry bark tortrix data.

groups suggesting that an attempt to eradicate the moth population may require differential treatments or, perhaps different doses, over the summer.

A one-component model was also fit to the data for comparison. As in the three-component case the fit results in a proportional intensity model as the smoothing parameter for the treatment effect diverges. Table 4.1 provides the estimates and their estimated jackknife standard errors for the treatment effect and overdispersion parameter; both of which are significant at 5%. The estimates themselves are similar to those of component two and three of the three-component mixture model. The baseline intensity and cumulative baseline intensity are displayed in Figure 4.2 and show the same trend as the overall baseline intensity and cumulative baseline intensity of the three-component model. For comparison, the step function estimates of Lawless and Zhan (1998) are provided as asterisks at the panel midpoints; both agree very well. The results obtained from both the one- and three-component models are in general accord. The three-component model, however, is helpful in distinguishing the differential effects for moths that overwinter in the "instar" phase, a question of interest in this study.
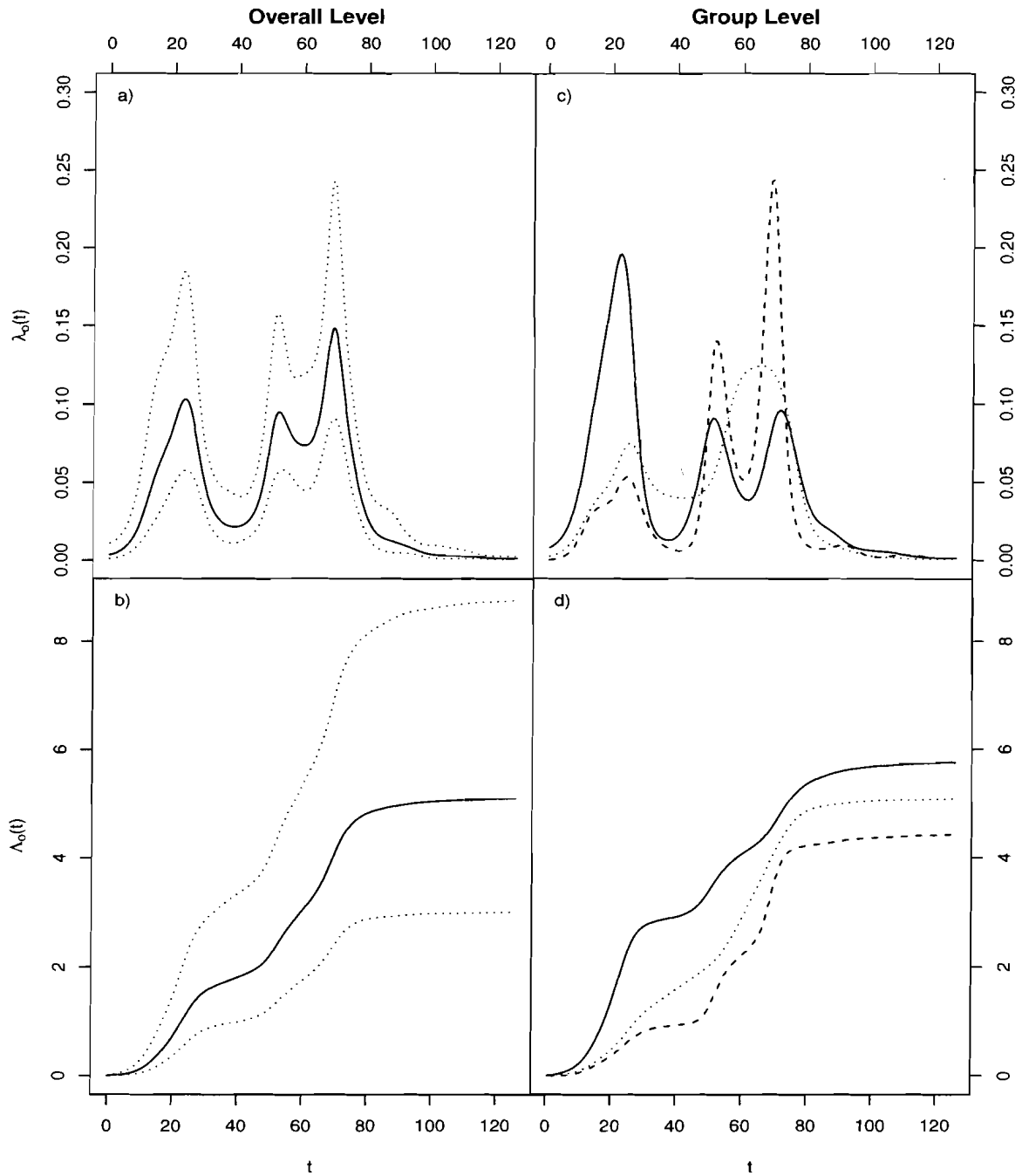
Figure 4.1: [a) and b)] Estimated overall baseline intensity $\lambda_0(t)$ and cumulative intensity $\Lambda_0(t)$ represented by a solid curve (——) along with their corresponding 95% point-wise confidence intervals ($\cdots\cdots$). [c) and d)] Estimated baseline intensities $\lambda_{g0}(t)$ and cumulative intensities $\Lambda_{g0}(t)$ of the three components. The solid curve (——) corresponds to group 1, the dashed curve (- - - -) to group 2 and the dotted curve ($\cdots\cdots$) to group 3.
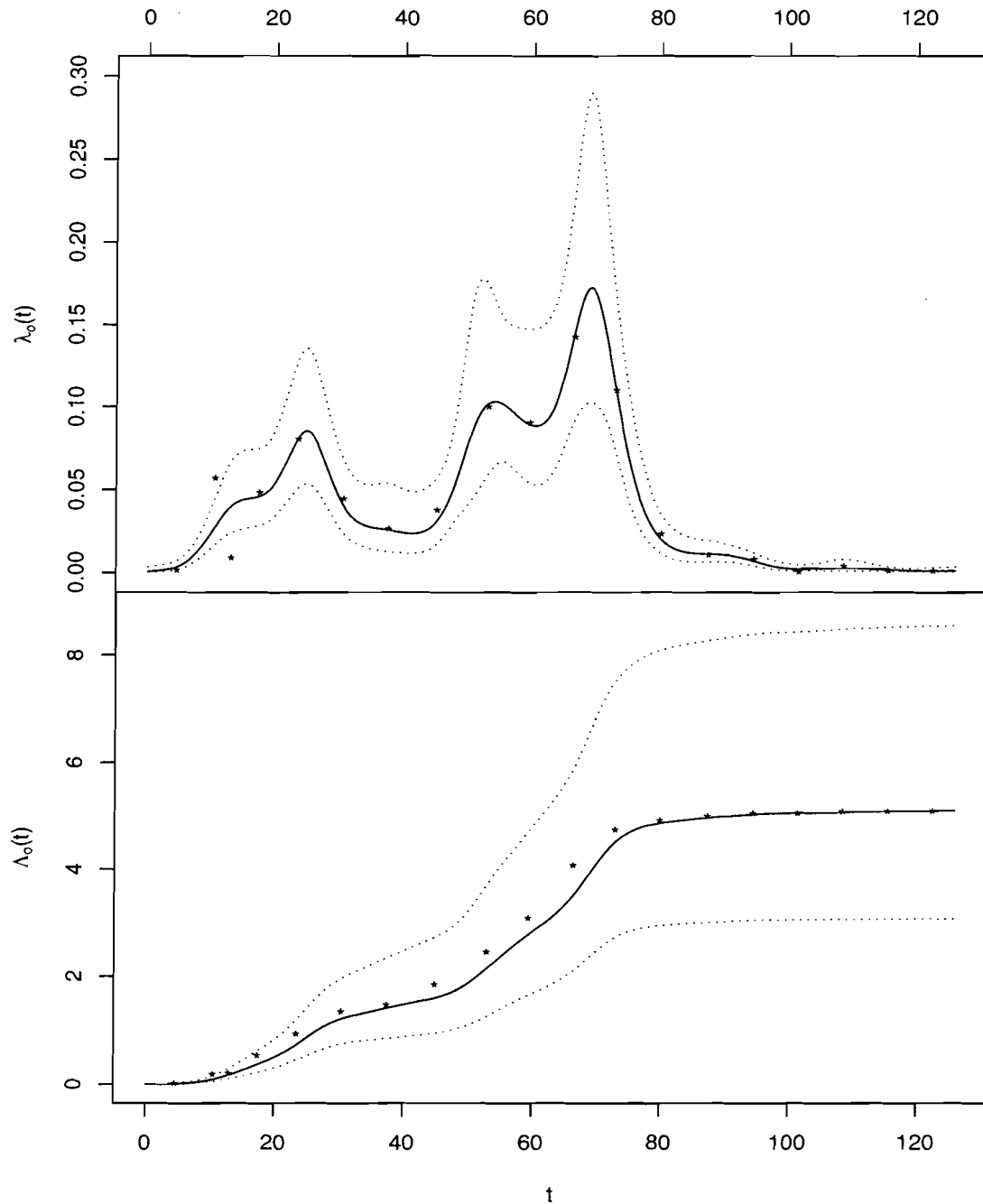
Figure 4.2: Estimated baseline intensity $\lambda_0(t)$ in the top panel and the estimated cumulative intensity $\Lambda_0(t)$ in the lower panel represented by solid curves (———) along with their corresponding 95% point-wise confidence intervals ($\cdots\cdots$). The two plots are overlaid with Lawless and Zhan's (1998) step-function estimates ($*$) placed at the panel midpoints.

4.4.2 Rat Carcinogenesis Experiment

This study was conducted by Thompson et al. (1978) to test whether the continued use of retinoids inhibits tumor recurrence. The experiment was carried out by injecting seventy-six rats with a carcinogen and subsequently administering retinyl acetate to prevent cancer for 60 days. After 60 days, the 48 rats that remained cancer-free were randomly assigned to a control group, or to a treatment group which carried on with retinoid prophylaxis. The rats were followed for 182 days after the initial injection and palpated for tumors every 2 to 5 days. Figure 4.3 gives a graphical summary of this data. Here $T_{io} = 0$ represents 60 days after injection with the carcinogen, resulting in a total follow-up of 122 days for each individual, with 25 individuals in the control group and 23 in the treatment group. The follow-up times were recreated as all unique times of occurrence of events since exact times of follow-up were not provided. This resulted in panels of average length of 3.70 days and standard deviation of 0.95 days with a maximum panel length of 6 days; this mimics closely the described 2-5 day gap between examinations as stated in the description of the study. The data was obtained from Table 1 in Gail et al. (1980).

The clustered overdispersed NHPP model developed in the previous section was fit to this data with a resultant fit of a single component model. The estimates of the smoothing parameters for both the baseline and treatment effect diverge to infinity, indicating that a proportional intensities assumption seems to hold and that the baseline is homogeneous. Figure 4.4 shows the estimated baseline and cumulative baseline intensity with 95% pointwise confidence intervals; empirical step-function estimates of the baseline and cumulative baseline intensity (Balshaw and Dean; 2002) are also provided for comparison. In the figure, the empirical step function estimates are replaced by point estimates at the panel midpoint rather than horizontal lines across the panels. Fitting the standard Weibull model (Lawless, 1987) with cumulative baseline $\Lambda_o(t) = t^\alpha$ yielded an estimate of α of 0.97 (s.e. 0.076), with no evidence against homogeneity ($\alpha = 1$). Table 4.2 provides estimates of the treatment effect and the overdispersion parameter. The three variance estimators are very similar and indicate that the treatment is effective in reducing tumor recurrence and that overdispersion is present. This example illustrates the utility of the chosen penalty (4.3) for selecting simple models in non-informative situations as well as the ability of the clustering procedure to isolate a simple one-component model where appropriate.
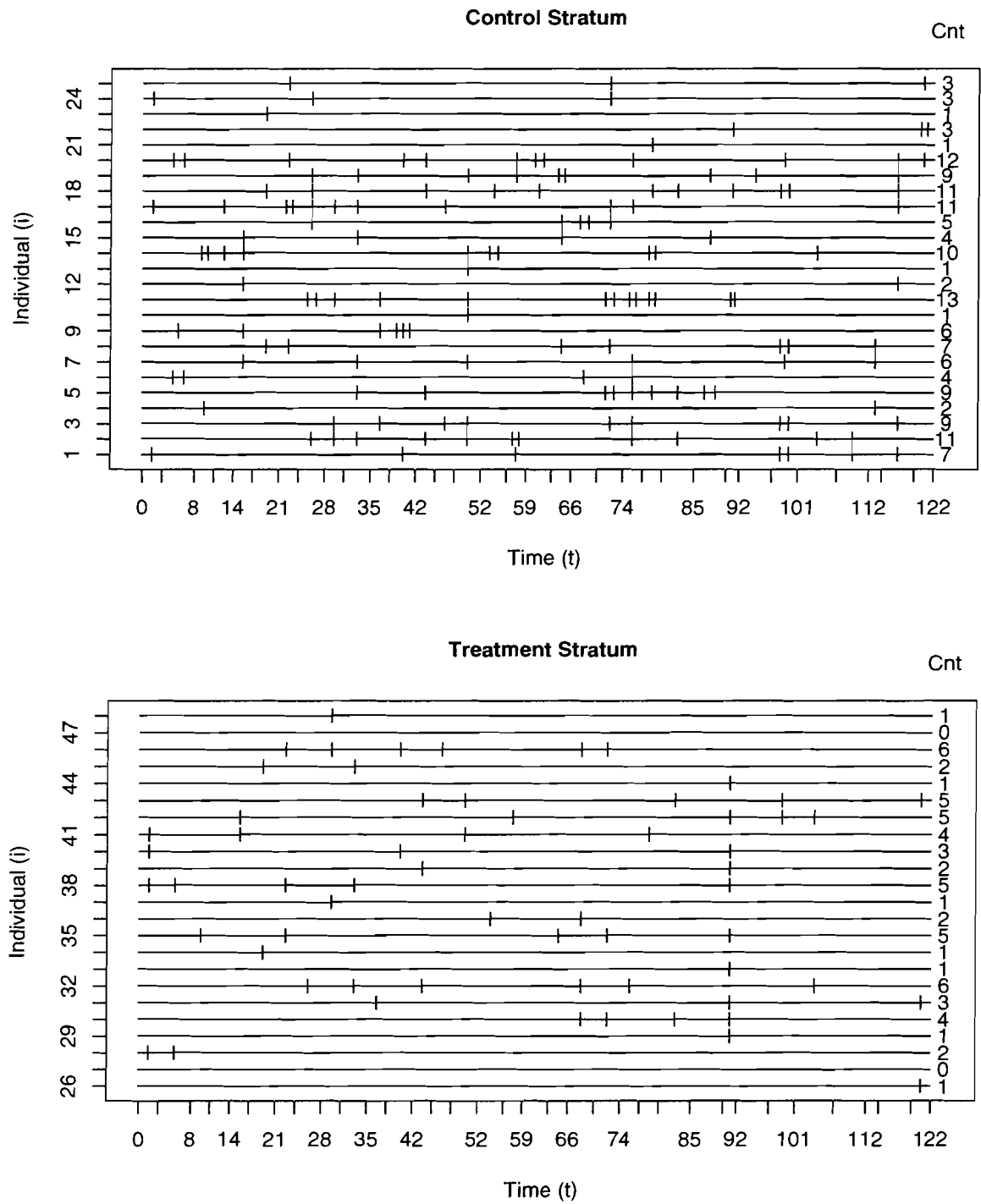
Figure 4.3: Data summary plot for the rat carcinogenesis data. The solid lines (———) represent time under observation, the light grey vertical lines are the panel follow-up times, the tick marks (|) indicate a count within the marked panel, spaced to be visually distinct, and the column labeled Cnt. represents the end of follow-up counts for each individual in the control and treatment groups.
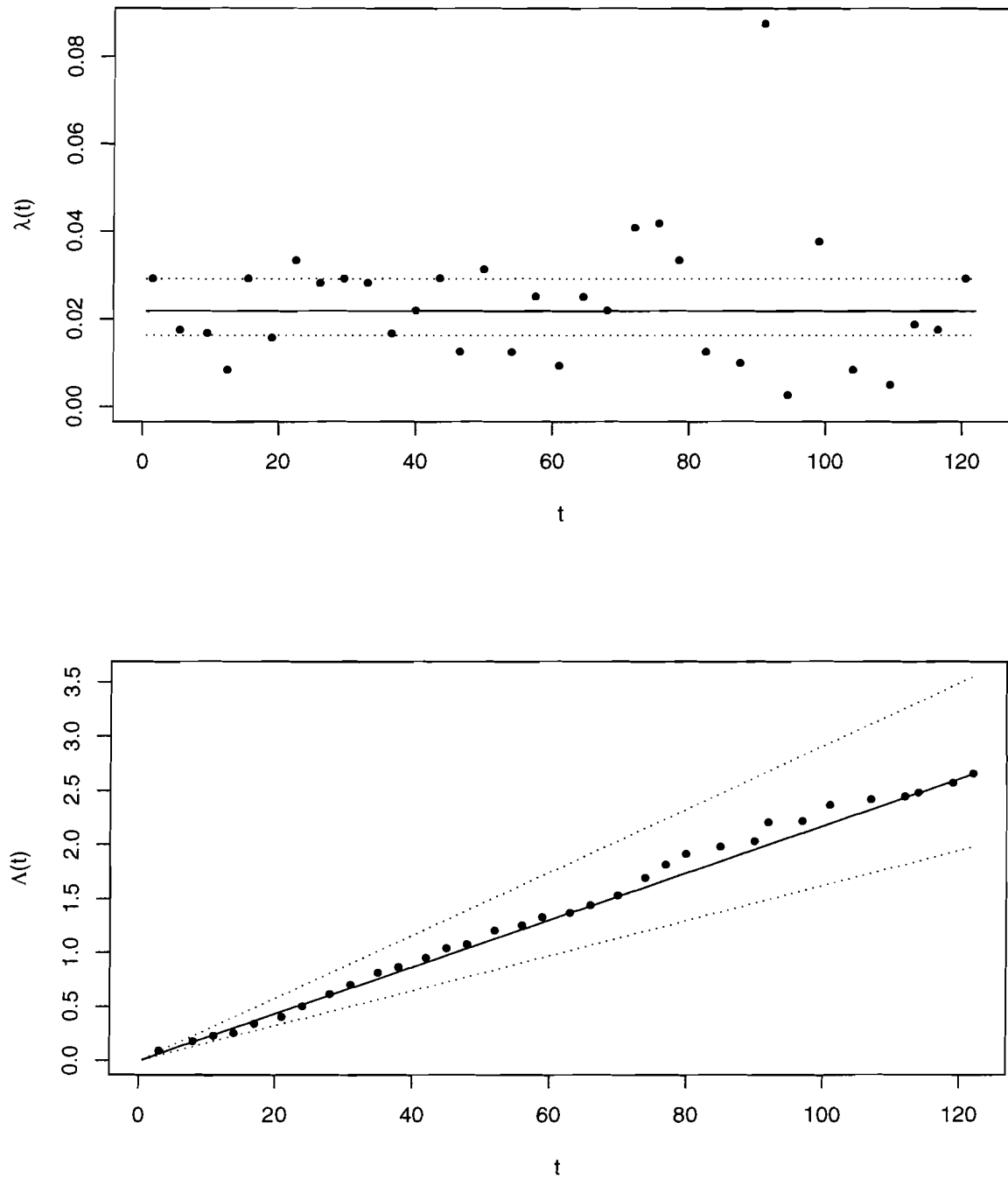
Figure 4.4: Estimated baseline intensity $\lambda_o(t)$ and cumulative intensity $\Lambda_o(t)$ represented by a solid line (——), their corresponding 95% point-wise confidence interval represented by dotted lines (· · · · ·) and overlaid empirical (step-function) model estimates (•) for the rat carcinogenesis data.

	Estimate	Robust SE	Jackknife SE	One-step SE
Baseline (γ_o)	-3.828	0.150	0.157	0.156
Treatment (γ_1)	0.823	0.197	0.206	0.205
Dispersion (τ)	0.286	0.051	0.066	0.052

Table 4.2: Parameter estimates and estimated robust, jackknife and one-step jackknife standard errors of the baseline log-intensity $\gamma_o(t)$, corresponding to the control group, the treatment effect $\gamma_1(t)$ and the overdispersion parameter τ from the fit of the one-component model to the carcinogenesis rat data.

4.4.3 Codling Moth SIR Study

The codling moth (*Cydia pomonella L.*) is a key pest of pear and apple trees globally. In particular, they have been a cause of lost revenue for the fruit farmers of the Okanagan Valley in the Canadian Province of British Columbia. The codling moths were controlled in the 50's and 60's by repeated use of a wide variety of insecticides; however advances in the use of anticidal control, pheromone trapping and mating disruption have changed the method of controlling this pest over the past 30 years. In particular, anticidal control or "sterile insect release" (SIR), where the mass release of sterilized male moths are used to over-flood the population in an attempt to make the number of viable matings rare, has been quite successful. The purpose of this study was to monitor the efficacy of anticidal control by pheromone trapping.

The data consist of counts of moths collected in baited traps installed on 506 trees scattered over the Okanagan valley. Each trap was counted and emptied once per week over the 20-week breeding period for several years. Clearly, the method of analysis presented in the previous sections is not appropriate for this data given its potential for spatial structure. In fact, it was of interest to see if there was any spatio-temporal difference in the breeding pattern of moths over the region. Here, as a preliminary analysis, one year of data using only unsterilized moths was analyzed using the methods presented in Section 4.2. Four components were identified. The estimated probabilities of group membership and their associated jackknife standard error estimates are 0.292 (0.290), 0.221 (0.166), 0.358 (0.404) and 0.129 (0.120) respectively. The estimates of the overdispersion parameters for all groups were quite large. Figure 4.5 presents the estimated baseline intensities for the four groups and shows that the moths breed in multiple waves. The four groups represent time-shifts of

this basic pattern with varying magnitudes of the breeding rate. The spatial position in the Okanagan valley of trees that strongly associate with each of the group-specific curves are displayed in Figure 4.6. An estimated group probability of $\tilde{p}_g > 0.6$ was chosen as a cutoff for determining that a tree is strongly associated with group g. This plot shows evidence of spatial clustering which merits further investigation.

4.5 Simulation Studies

To evaluate the estimators derived in Section 4.2, a simulation experiment was performed with two factors: the number of individuals with levels $I = 25, 50, 100$ and the number of panels with levels $e = e_i = 25, 50$, $i = 1, \ldots, I$. The panel follow-up times were equally spaced and concurrent for all individuals over the interval $t \in [0, 1)$. In a complete factorial design, ten thousand datasets, $\{N_{ij}, i = 1, \ldots, I, j = 1, \ldots, e\}$, were generated from the counting process model with intensity

$$\sum_{g=1}^{2} \nu_{gi} z_{gi} \lambda_{gi}(t) \, \mathbf{1} \{t \in [T_{j-1}, T_j)\}$$

where $z_{1i} \overset{i.i.d.}{\sim} Bernoulli(p_1 = 0.4)$, $z_{2i} = 1 - z_{1i}$, $\nu_{1i} \overset{i.i.d.}{\sim} \Gamma(0.2, 1/0.2)$ given $z_{1i} = 1, \nu_{2i} \overset{i.i.d.}{\sim} \Gamma(0.6, 1/0.6)$ given $z_{2i} = 1$, and $\lambda_{gi}(t) = \exp\{\gamma_{g0}(t) + \mathbf{1}\{i \in \text{Stratum } 1\} \gamma_{g1}(t)\}$. Figure 4.7 shows the selected functional forms for the intensities. These represent smoothly decreasing and bimodal recurrence rates, with associated treatment group effects as a constant or which decrease the rate of recurrence linearly. The specific goal considered first is to test the performance of the quasi-likelihood estimating equations. Here we focus on whether the penalty method performs well, whether the splines are accurately estimating the recurrence rates and how good coverage probabilities are in small samples. The probabilities of group membership were set at $p_1 = 0.4$ and $p_2 = 0.6$ so as not to have too few individuals per cluster, leaving control of sample size to I. The overdispersion parameters were chosen to represent a low level of dispersion ($\tau_1 = 0.2$) and a high level ($\tau_2 = 0.6$). The clustered overdispersed NHPP model was fit to each dataset.

Figure 4.8 presents the mean of the simulated estimates overlaying the true values for the baseline and treatment spline effects for both groups at all levels of number of individuals (I) and panels (e). The results presented in this plot suggest that the method is doing a good job of estimating the underlying functional forms. To see clearer how the estimates of
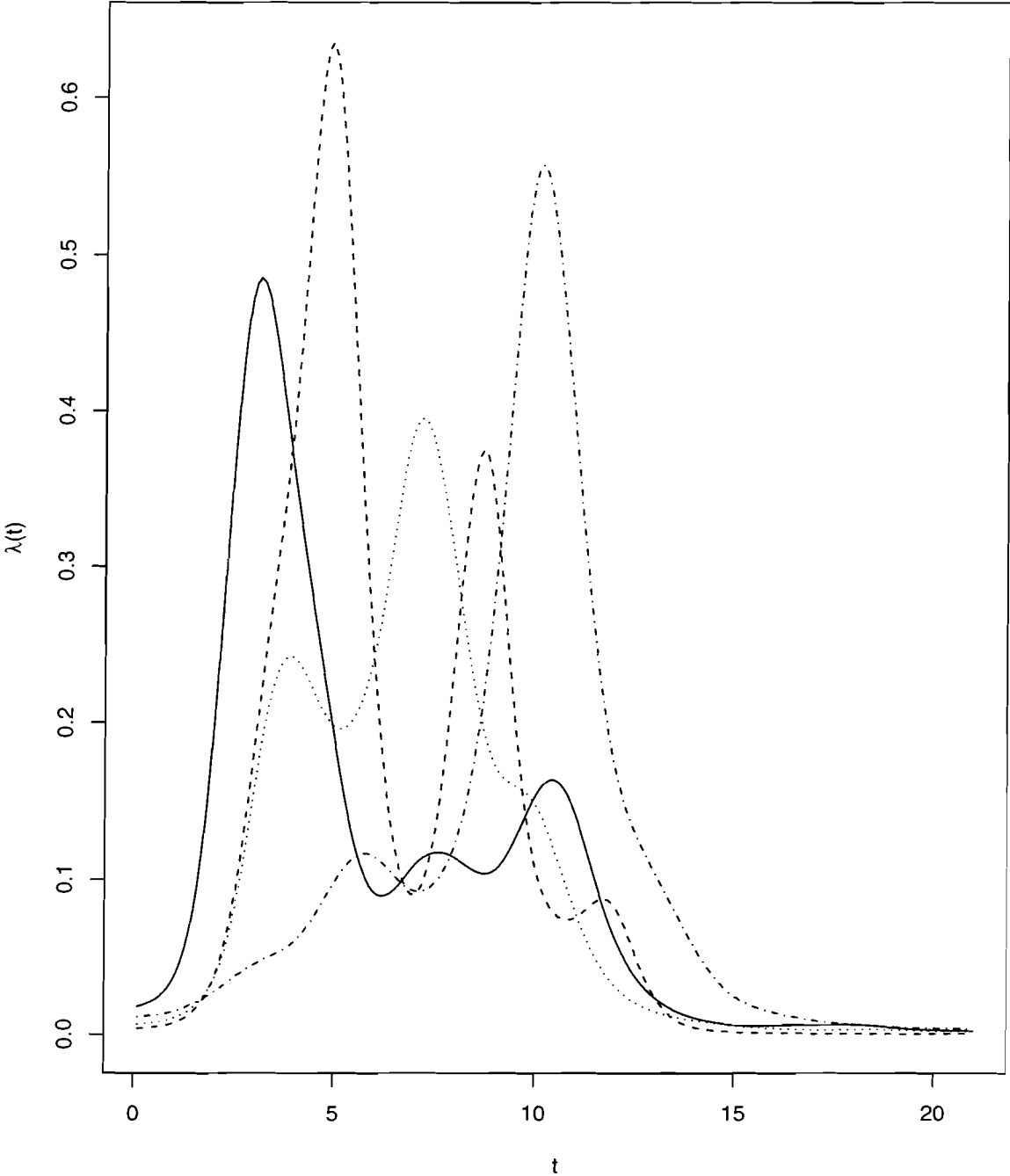
Figure 4.5: Estimated baseline intensities $\lambda_{go}(t)$ of the four components for the codling moth SIR data. The solid curve (———) corresponds to group 1, the dashed curve (- - - -) to group 2, the dotted curve ($\cdots\cdots$) to group 3 and the dash-dot curve (—·—·—) to group 4.

Figure 4.6: The spatial location of trees in the codling moth SIR data. Trees with estimated group membership probabilities indicating strong association are represented by g, $g = 1, 2, 3, 4$ and those that do not strongly associate displayed with a dot ($\cdot$).

Baseline Intensities

Treatment Effects

Intensity Function for Treatment Stratum
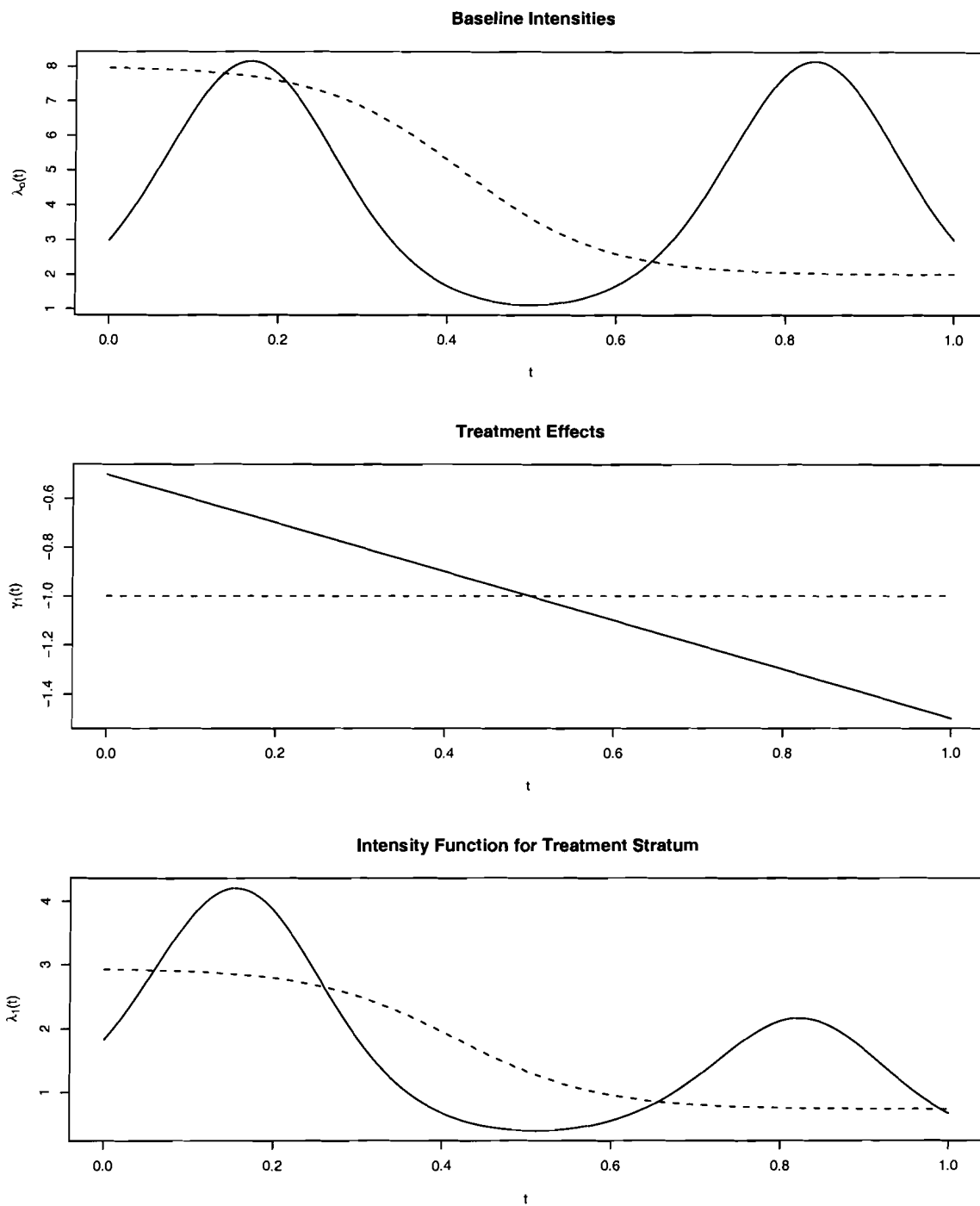
Figure 4.7: Functional forms for the baseline intensities $\lambda_{g0}(t) = \exp\{\gamma_{g0}(t)\}$, the treatment spline effects $\gamma_{g1}(t)$ and the intensities for the treatment stratum $\lambda_{g0}(t) = \exp\{\gamma_{g0}(t) + \gamma_{g1}(t)\}$; group 1 ($g = 1$) is represented by the solid curves (——) and group 2 ($g = 2$) is represented by the dashed curves (- - - -).

the spline effects, $\hat{\gamma}_{qg}(t)$, are affected by the factors of the experiment a similar plot of the biases is presented in Figure 4.9. This plot shows that the method performs better as the number of panels and individuals increases, as would be expected. However, it seems that increasing the number of panels reduces the bias at a faster rate than increasing the number of individuals. A comparison of the performance of the robust, jackknife and one-step jackknife standard error estimates of the spline effects for confidence interval construction is presented in Figure 4.10. To focus on differences amongst estimators, the coverage-axis ranges between 0.84 and 0.96. In general, coverage probabilities are on target when I is large. The performance seems to depend more strongly on the number of individuals than the number of panels. The jackknife and one-step jackknife estimators give similar coverage and perform better than the robust estimator.

Table 4.3 presents the mean values of estimates, simulated standard error estimates and mean estimated standard errors of the group probabilities p_g and overdispersion parameters τ_g, $g = 1, 2$. All estimates of the probabilities of group membership and their estimated standard errors are on target in all levels of the experiment. The same is not true for the estimates of the overdispersion parameters, however, which seem to improve as the number of individuals or number of panels increases, but more so for increasing I. The coverage probabilities of estimates of these parameters, also provided in the table, are poor for the case when there is a small sample size ($I = 25$) and improves only slowly with increasing I. That estimators of overdispersion parameters tend to behave poorly in small samples is not new; for example, recommendations that inference for these parameters be carried out by score tests have been made (Breslow, 1990; Dean and Lawless, 1989).

An extra run of this experiment was performed using inverse Gaussian distributed random effects instead of the the gamma random effects previously described. Importantly, there was no noticeable difference in point or interval estimates of the spline effects or the group probabilities, providing some evidence of robustness of the methods developed. The estimates of the overdispersion parameters, however, were affected, as would be expected since the Inverse Gaussian distribution has heavier tails than the gamma. Here we only present results concerning this parameter. These are displayed in Table 4.3 marked (IG), showing the tendency of some bias in these estimators.

An analogous simulation to investigate performance of the approach in the important one-component special case was also carried out using the true intensity as the intensity function of group 1 of the previous experiment, at various levels of overdispersion, $\tau =$
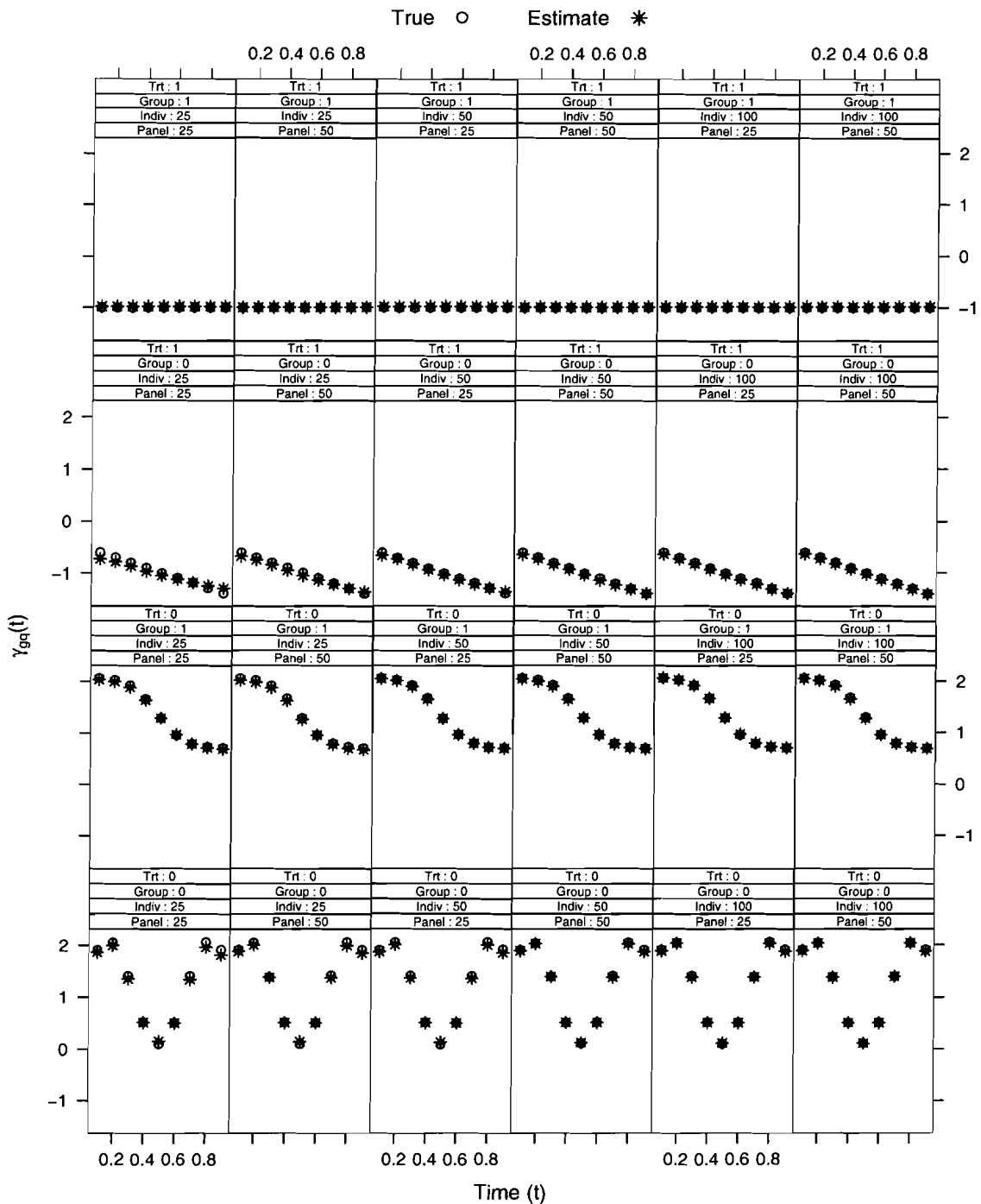
Figure 4.8: Plots of the simulated mean values (*) overlaying the true values (o) of the spline effects $\gamma_{gq}(t)$, $t = 0.1, \ldots, 0.9$ for groups $g = 1, 2$ and treatments $q = 0, 1$ for all levels of number of individuals $I = 25, 50, 100$ and numbers of panels $e = 25, 50$. The first row displays results for $\gamma_{21}(t)$, the second row for $\gamma_{11}(t)$, the third row for $\gamma_{20}(t)$, and the last row for $\gamma_{10}(t)$.
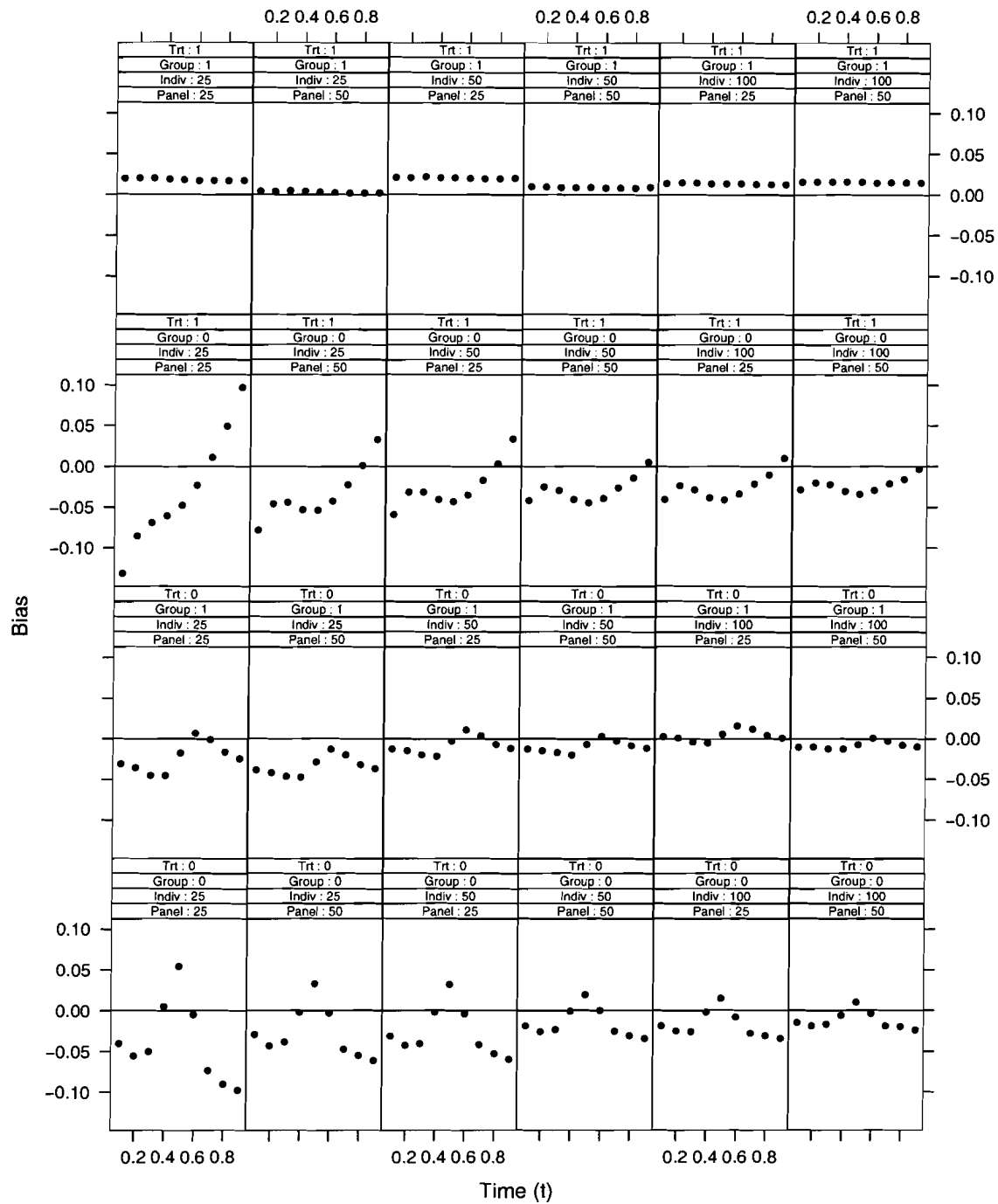
Figure 4.9: Plots of the simulated bias of the spline effects $\gamma_{gq}(t)$, $t = 0.1, \ldots, 0.9$ for groups $g = 1, 2$ and treatments $q = 0, 1$ for level of number of individuals $I = 25, 50, 100$ and number of panels $e = 25, 50$. The first row displays results for $\gamma_{21}(t)$, the second row for $\gamma_{11}(t)$, the third row for $\gamma_{20}(t)$, and the last row for $\gamma_{10}(t)$.
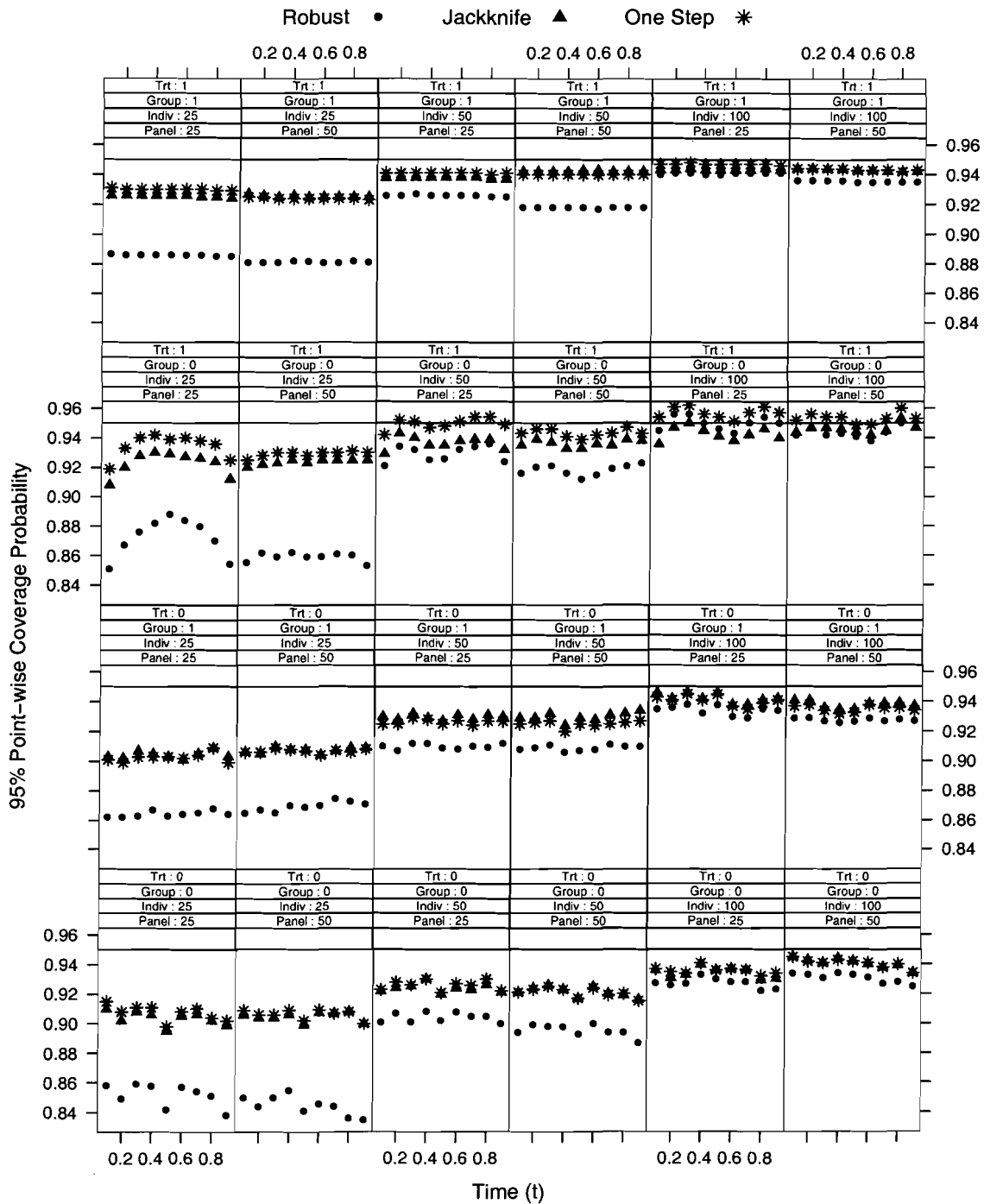
Figure 4.10: Point-wise coverage probabilities for 95% confidence intervals constructed from the robust (•), jackknife (▲) and one-step jackknife (∗) standard errors estimates of the spline effects $\gamma_{gq}(t), t = 0.1, \ldots, 0.9$ for groups $g = 1, 2$ and treatments $q = 0, 1$ from the simulated data analysis. All levels of numbers of individuals $I = 25, 50, 100$ and numbers of panels $e = 25, 50$ are displayed. The first row displays results for $\gamma_{21}(t)$, the second row for $\gamma_{11}(t)$, the third row for $\gamma_{20}(t)$, and the last row for $\gamma_{10}(t)$.

|	Panels	I	True	Mean	Sim. SE	SE (Rb.|Jk.|Os.)	95% (Rb.|Jk.|Os.)
p_1	25	25	0.4	0.418	0.103	(0.104 | 0.104 | 0.108)	(0.934 | 0.937 | 0.944)
		50	0.4	0.415	0.074	(0.074 | 0.073 | 0.074)	(0.942 | 0.941 | 0.944)
		100	0.4	0.411	0.050	(0.053 | 0.052 | 0.052)	(0.950 | 0.949 | 0.950)
	50	25	0.4	0.411	0.099	(0.098 | 0.101 | 0.103)	(0.937 | 0.940 | 0.943)
		50	0.4	0.407	0.071	(0.070 | 0.071 | 0.071)	(0.941 | 0.944 | 0.944)
		100	0.4	0.409	0.050	(0.050 | 0.050 | 0.050)	(0.948 | 0.950 | 0.950)
p_2	25	25	0.6	0.582	0.103	(0.104 | 0.104 | 0.108)	(0.934 | 0.937 | 0.944)
		50	0.6	0.585	0.074	(0.074 | 0.073 | 0.074)	(0.942 | 0.941 | 0.944)
		100	0.6	0.589	0.050	(0.053 | 0.052 | 0.052)	(0.950 | 0.949 | 0.950)
	50	25	0.6	0.589	0.099	(0.098 | 0.101 | 0.103)	(0.937 | 0.940 | 0.943)
		50	0.6	0.593	0.071	(0.070 | 0.071 | 0.071)	(0.941 | 0.944 | 0.944)
		100	0.6	0.591	0.050	(0.050 | 0.050 | 0.050)	(0.948 | 0.950 | 0.950)
τ_1	25	25	0.2	0.171	0.100	(0.096 | 0.107 | 0.123)	(0.730 | 0.791 | 0.758)
		50	0.2	0.195	0.080	(0.078 | 0.080 | 0.082)	(0.858 | 0.875 | 0.868)
		100	0.2	0.208	0.059	(0.059 | 0.057 | 0.060)	(0.925 | 0.925 | 0.929)
	50	25	0.2	0.165	0.093	(0.086 | 0.097 | 0.111)	(0.709 | 0.777 | 0.736)
		50	0.2	0.189	0.071	(0.067 | 0.072 | 0.070)	(0.842 | 0.864 | 0.853)
	(IG)	50	0.2	0.280	0.070	(0.066 | 0.069 | 0.069)	(0.906 | 0.919 | 0.930)
		100	0.2	0.203	0.055	(0.054 | 0.053 | 0.055)	(0.912 | 0.916 | 0.914)
τ_2	25	25	0.6	0.590	0.191	(0.181 | 0.199 | 0.193)	(0.819 | 0.869 | 0.827)
		50	0.6	0.581	0.166	(0.176 | 0.161 | 0.181)	(0.854 | 0.865 | 0.860)
		100	0.6	0.575	0.127	(0.147 | 0.121 | 0.148)	(0.893 | 0.871 | 0.895)
	50	25	0.6	0.601	0.195	(0.180 | 0.205 | 0.193)	(0.831 | 0.882 | 0.845)
		50	0.6	0.591	0.158	(0.171 | 0.161 | 0.176)	(0.870 | 0.885 | 0.877)
	(IG)	50	0.6	0.562	0.193	(0.191 | 0.172 | 0.197)	(0.804 | 0.788 | 0.812)
		100	0.6	0.585	0.122	(0.144 | 0.121 | 0.146)	(0.905 | 0.892 | 0.908)

Table 4.3: Mean values of estimates, simulated standard errors, (Robust|Jackknife|One-Step Jackknife) standard error estimates and corresponding coverage probabilities of 95% confidence intervals for the group probabilities p_g and overdispersion parameters τ_g, $g = 1, 2$. Results are included for all levels of numbers of individuals $I = 25, 50, 100$ and numbers of panels $e = 25, 50$ as well as the single run with data generated from the Inverse Gaussian overdispersion (IG) model.

0.2, 0.4, 0.6. Similar good performance of the methods was observed. In addition, the performance of the robust variance estimator improves, though the jackknife and one-step jackknife standard error estimators are still superior to the robust estimators for small samples.

To investigate the ability of the model to distinguish between similar component functional forms another simulation was conducted by generating data from shifted sinusoids for the group-specific intensities, $\lambda_{g0}(t) = \exp\{3\sin(3\pi t + s_g) + u_g\}$. Data was generated from a three-component model with 60 individuals and 25 panels in a manner analogous to the previous experiment with group one having no shift $u_1 = s_1 = 0$, the second group having only a horizontal shift $(s_2 \neq 0, u_2 = 0)$ and the third group having only a vertical shift $(s_3 = 0, u_3 \neq 0)$. Ten thousand replicates were generated per selected shift level. In this experiment the overdispersion parameters were fixed at $\tau_g = 0.2$ and $p_g = 1/3$ for all groups. These settings were selected to minimize the impact of these parameters since increasing the overdispersion simply increases the values of the shifts s_g and u_g at which the method performs poorly, and the use of identical group probabilities eliminates the need for considering subgroup sample size as a factor. Results indicate that the model is quite good at distinguishing the differing functional forms even for these similar shapes. The method breaks down as $s_2 \to 0$ and $u_3 \to 0$ as would be expected, yielding, in these cases, a fit of a one-component model. However, surprisingly good differentiation of groups occurs when the shifts are quite close. For example, Figure 4.11 displays plots of the true component intensities along with their estimated simulated mean values for two levels $(s_2 = \pi/2, u_3 = 2$; $s_2 = \pi/4, u_3 = 1)$.

We also considered how well the proposed algorithm (Section 4.2, Case 2) identifies the appropriate number of sub-components as $s_2 \to 0$ and $u_3 \to 0$. Table 4.4 shows the proportion of times a g-component model is estimated, for the cases with $(s_2 = \pi/2, u_3 = 2)$ and $(s_2 = \pi/4, u_3 = 1)$. There seems to be good differentiation of the three-subcomponent intensities; as these sub-group intensities converge, $s_2 \to 0$ and $u_3 \to 0$, the number of components selected also tends to decrease.
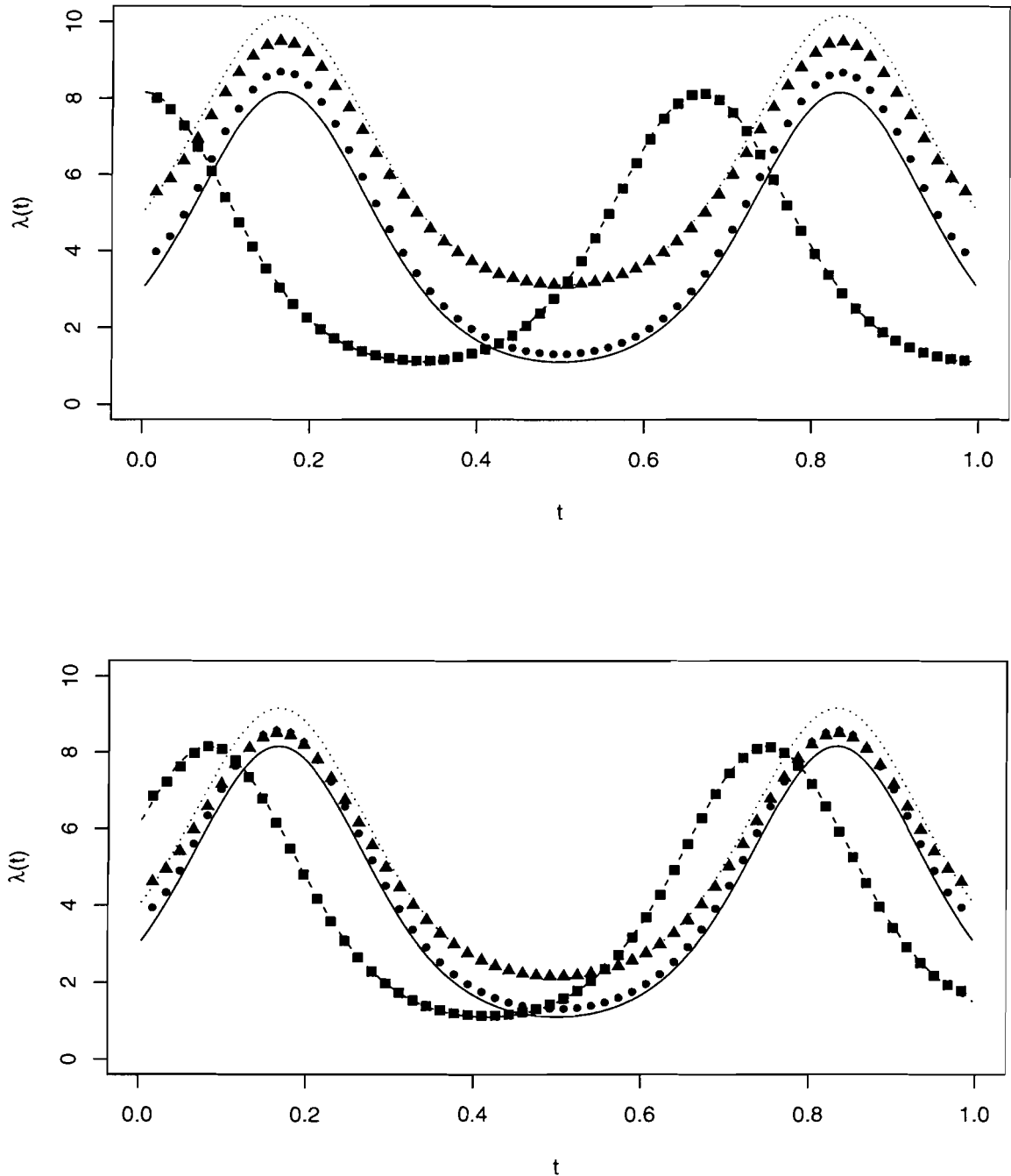
Figure 4.11:   Plots of the sinusoidal intensities for the three groups, $\lambda_{g0}(t)$ = $\exp\{3\sin(3\pi t + s_g) + u_g\}$, $g = 1, 2, 3$, represented by solid (——), dashed (- - - -) and dotted ($\cdots\cdots$) curves overlaid with their respective simulated mean estimates (•), (■) and (▲). In the upper panel s = $(0, \pi/2, 0)$ and u = $(0, 0, 2)$ while in the lower panel s = $(0, \pi/4, 0)$ and u = $(0, 0, 1)$.

# of Components (g)	Proportion of g-Component Models	
	(a)	(b)
2	0.159	0.451
3	0.490	0.347
4	0.266	0.168
5	0.076	0.030
6	0.008	0.004
7	0.001	0.000

Table 4.4: Proportion of times a g-component model was selected for a) $s_2 = \pi/2$, $u_3 = 2$ and b) $s_2 = \pi/4$, $u_3 = 1$.

4.6 Discussion

Discrete mixtures of overdispersed nonhomogeneous Poisson process spline models are a powerful tool for analyzing recurrent event panel data. They provide a means of incorporating the complex within- and between-individual correlation structures common to this type of data, extracting information on the underlying functional mechanisms that are driving the observed recurrence pattern and allow for the modeling of unobservable subpopulations that generate these observed recurrences. In addition, the single cluster model itself is a useful, flexible tool for the analysis of longitudinal panel count data. This chapter presents a method of estimation and inference for the broad framework of mixtures of such non-homogeneous counting process models requiring only low moment assumptions. The finite sample properties of the proposed estimating functions are also investigated.

Note that smooth forms are assumed for the underlying process governing the generation of events. Also, if the intensity oscillates more in the first half of the domain and is, say, linear in the second half then the method will tend to under-fit the function in the first regime and over-fit in the second. An adaptive penalty as discussed in Chapter 3 would be more suitable for this situation although such an extension would add further complexity to the model.

The mixture model presented can shed light on clusterings of individuals and differential behavior between unobserved sub-groups with regard to treatment effects. It is particularly useful as an exploratory tool or when scientific evidence provides evidence for the use of mixtures. More general models within the broad framework introduced here could also

be envisioned. For example, the authors have moth data akin to the cherry bark tortrix example discussed in Section 4.4.1 at several hundred sites over the Okanagan valley of the Canadian province of British Columbia. In this case it would be of interest to determine if sub-population behavior differs not only in time but also over space. Spatial correlation can also be incorporated into the current framework by considering spatially correlated random effects in the intensity. The discrete mixture methods developed here may also be adapted for survival analysis where the intensity (4.1) would represent group-specific hazard functions. These might be appropriate in cure-rate analysis where a proportion of individuals experience a cure or in the analysis of events pertaining to harsh treatments where a sub-population experiences early failure after treatment. These ideas are currently being investigated.

Chapter 5

Future Work

5.1 Extended Overdispersion Model

In much of the work presented in this thesis only first and second moment assumptions were made on the mixing distribution m. This is useful as it allows for the modeling of overdispersion without explicit distributional assumptions, and inference procedures discussed offer consistency in estimates of regression parameters and full efficiency in certain common cases for balanced designs (Dean and Balshaw, 1997). Here efficiency is with respect to maximum likelihood assuming full distributional assumptions for m. In Chapter 4 when estimating an individuals affinity to a particular sub-group, that is, estimation of the z_{gi}'s, relies on the correctness of a gamma assumption for the within-cluster mixing distributions m_g. Although it has been found that the gamma distribution is sufficiently flexible in many situations a method of assessing the validity of this assumption is desirable.

Let $N^{-1}(\tau, \omega, \varsigma)$ denote the generalized inverse Gaussian distribution with density,

$$m(y; \tau, \omega, \varsigma) = \frac{\varsigma^{-\tau}}{2K_\tau(\omega)} y^{\tau-1} \exp\left\{-\frac{1}{2}\omega(\varsigma y^{-1} + \varsigma^{-1}y)\right\} \mathbf{1}(y > 0) \quad \text{on } \Theta_\tau \qquad (5.1)$$

where $K_\tau(\cdot)$ is the modified Bessel function of the second kind (Abramowitz and Stegun, 1984) and

$$\Theta_\tau = \begin{cases} \{(\omega, \varsigma): \omega \geq 0, \varsigma \geq 0\}, & \tau > 0 \\ \{(\omega, \varsigma): \omega > 0, \varsigma > 0\}, & \tau = 0 \\ \{(\omega, \varsigma): \omega \geq 0, \varsigma > 0\}, & \tau < 0 \end{cases}.$$

Some useful identities for $K_\tau(\omega)$, $\tau \in \mathbb{R}$ and $\omega > 0$ are given by:

1) $K_\tau(\omega) = \frac{1}{2} \int_0^\infty t^{\tau-1} \exp\left\{-\frac{1}{2}\omega(t + t^{-1})\right\} dt$

2) $K_\tau(\omega) = K_{-\tau}(\omega)$

3) $K_{\tau+1}(\omega) = \frac{2\tau}{\omega} K_\tau(\omega) + K_{\tau-1}(\omega)$

4) $\frac{\partial}{\partial\omega} K_\tau(\omega) = \frac{\tau}{\omega} K_\tau(\omega) - K_{\tau+1}(\omega) = -\frac{\tau}{\omega} K_\tau(\omega) - K_{\tau-1}(\omega) = -\frac{1}{2} K_{\tau+1}(\omega) - \frac{1}{2} K_{\tau-1}(\omega)$.

Using these identities it is straight-forward to show by manipulating the moment generating function that the moments of $Y \sim N^{-1}(\tau, \omega, \varsigma)$ are given by

$$E[Y^k] = \frac{K_{\tau+k}(\omega)}{K_\tau(\omega)} \varsigma^k, \quad k = 1, \dots \, . \tag{5.2}$$

This is a very flexible class of distributions that includes as special cases: the gamma ($\omega \to 0$, $\tau > 0$), the reciprocal gamma ($\omega \to 0$, $\tau < 0$), the inverse Gaussian ($\tau = -1/2$), the reciprocal inverse Gaussian ($\tau = 1/2$) and the hyperbola distribution ($\tau = 0$). It should be noted that some care needs to be taken in the limiting cases when $\omega \to 0$. Similarly when $\tau = 0$ and either of ω or ς is small, (5.1) is numerically unstable and requires high precision arithmetic to be computed in practice. Another useful and little known property of the generalized inverse Gaussian distribution is that this distribution is conjugate in the Bayesian sense for the Poisson distribution. So if $N|Y = y \sim Pois(y\mu)$ and $Y \sim N^{-1}(\tau, \omega, \varsigma)$ then $Y|N = n \sim N^{-1}(\tau + n, \omega\varsigma, \omega/\varsigma + 2\mu)$ with mean

$$E[Y|N = n] = R_{\tau+n+1}\left(\sqrt{\omega^2 + 2\omega\varsigma\mu}\right)\left(\frac{\omega\varsigma}{\sqrt{\omega^2 + 2\omega\varsigma\mu}}\right)$$

where $R_\tau(\omega) = K_{\tau+1}(\omega)/K_\tau(\omega)$. Conjugacy here aids in computation both for the Bayesian approach because of the explicit posterior distribution and in the frequentist approach due to a closed form of the M-step required to implement EM type algorithms. For further information on the generalized inverse Gaussian distribution and its properties see Jørgensen (1982).

Assume that $N_i|V_i = \nu_i$, $i = 1, \dots, I$ are independently Poisson distributed with mean $\nu_i\mu_i$ and that $V_i \overset{i.i.d.}{\sim} N^{-1}(\tau, \omega, \varsigma = 1)$ with density given in (5.1); here the μ_i's will be assumed known for simplicity of presentation. The unconditional distribution of each N_i is then given by

$$\begin{aligned} P(N_i = n_i \mid \tau, \omega) &= \int_0^\infty \frac{(\nu_i\mu_i)^{n_i} \exp\left\{-\nu_i\mu_i\right\}}{n_i!} m(\nu_i; \tau, \omega, \varsigma = 1) \, d\nu_i \\ &= \frac{\mu_i^{n_i}}{n_i!} \left(\frac{\omega}{\sqrt{\omega^2 + 2\omega\mu_i}}\right)^{\tau+n_i} \frac{K_{\tau+n_i}\left(\sqrt{\omega^2 + 2\omega\mu_i}\right)}{K_\tau(\omega)} \end{aligned} \tag{5.3}$$

for $n_i = 0, 1, \ldots$. This is the so-called Sichel distribution (Sichel, 1974) with mean $E[N_i] = R_\tau(\omega)\mu_i$ and variance $V[N_i] = R_\tau(\omega)\mu_i + \left(K_{\tau+2}(\omega)/K_\tau(\omega) - R_\tau(\omega)^2\right)\mu_i^2$. Special cases of this mixed Poisson distribution of interest are the Negative Binomial

$$
\begin{aligned}
P_{nb}(N_i = n_i \mid \tau) &= \lim_{\omega \to 0} P(N_i = n_i \mid \tau, \omega) \\
&= \frac{\Gamma(n_i + \tau)}{n_i!\,\Gamma(\tau)} \left(\frac{\mu_i}{\tau + \mu_i}\right)^{n_i} \left(\frac{\tau}{\tau + \mu_i}\right)^\tau, \ \tau > 0
\end{aligned}
\tag{5.4}
$$

and Poisson inverse Gaussian distributions ($\tau = -1/2$). Denote $l(\tau, \omega) = \sum_{i=1}^{I} \log P(N_i = n_i \mid \tau, \omega)$ and $l_{nb}(\tau) = \sum_{i=1}^{I} \log P_{nb}(N_i = n_i \mid \tau)$ as the log-likelihood of the Sichel and negative binomial respectively. The hypothesis that the mixing distribution m is gamma distributed ($H_0 : \tau > 0, \ \omega \to 0$) can be assessed using the test statistic

$$
T = \frac{l(\hat{\tau}, \hat{\omega}) - l_{nb}(\hat{\tau}_{nb})}{l(\hat{\tau}, \hat{\omega})/(I - 2)}
\tag{5.5}
$$

where $\hat{\tau}$ and $\hat{\omega}$ are the solutions to the equations

$$
\frac{\partial}{\partial \omega} l(\tau, \omega) = \sum_{i=1}^{I} \left(\frac{n_i}{\omega} + R_\tau(\omega) - R_{\tau+n_i}\left(\sqrt{\omega^2 + 2\omega\mu_i}\right) \frac{\omega + \mu_i}{\sqrt{\omega^2 + 2\omega\mu_i}} \right) = 0,
\tag{5.6}
$$

$$
\frac{\partial}{\partial \tau} l(\tau, \omega) = \sum_{i-1}^{I} \left(\frac{\frac{\partial}{\partial \tau} K_{\tau+n_i}\left(\sqrt{\omega^2 + 2\omega\mu_i}\right)}{K_{\tau+n_i}\left(\sqrt{\omega^2 + 2\omega\mu_i}\right)} - \frac{\frac{\partial}{\partial \tau} K_\tau(\omega)}{K_\tau(\omega)} - \frac{\log(\omega + 2\omega\mu_i) - \log\omega}{2} \right) = 0,
$$

and $\hat{\tau}_{nb}$ the solution to

$$
\frac{\partial}{\partial \tau} l_{nb}(\tau) = \sum_{i=1}^{I} \left[\Psi(n_i + \tau) - \Psi(\tau) - \log\left(\frac{\tau}{\tau + n_i}\right) \right] = 0,
$$

with $\Psi(\tau)$ being the digamma function. The test statistic T (5.5) is asymptotically F-distributed with 1 numerator and $I - 2$ denominator degrees of freedom. There is no closed form expression for $\partial K_\tau(\omega)/\partial \tau$ but it can be computed accurately via a series expansion. Alternatively a finite difference approximation

$$
\frac{\partial}{\partial \tau} K_\tau(\omega) = \frac{-K_{\tau+2h}(\omega) + 8K_{\tau+h}(\omega) - 8K_{\tau-h}(\omega) + K_{\tau-2h}(\omega)}{12h} + O(h^4).
$$

may be used and this has been found to be sufficiently accurate in practice for a suitably small h (square root of machine precision). Similarly, a test for the hypothesis that m is inverse Gaussian can also be constructed by replacing $l_{nb}(\hat{\tau}_{nb})$ in (5.5) with $l(\tau = -1/2, \hat{\omega}_g)$ where $\hat{\omega}_g$ is the solution to (5.6) with τ fixed at $-1/2$.

The use of a broader family for the mixing distribution presented here could be extended to the more complex case considered in Chapter 4 and such an extension would be useful in model validation. Alternatively, the method could be used directly to obtain a suitable within-cluster mixing distribution contained within the Poisson generalized inverse Gaussian family.

5.2 Spatial Extensions

A current area of investigation is the extension of the functional counting process methods developed herein to the analysis of spatio-temporal data. Consider as a motivating example the codling moth sterile insect release (SIR) data. The goal of this study was to assess the effectiveness of anticidal control, the mass release of sterilized male moths (SIR) to over-flood the population in order to reduce viable matings. The data displayed in Figure 5.1 consist of counts of moths collected in baited traps installed on 506 trees scattered over the Okanagan valley in the Canadian province of British Columbia. The data at each tree site in the study is similar in structure to the cherry bark tortrix data and consists of trap counts collected weekly for twenty weeks over a summer. Clearly here it is not only of interest to model the rate of change of recurrences over time but also through space while simultaneously taking into account potential spatial variation likely to exist in data collected over a region. A method of incorporating such spatial correlation into the counting process framework discussed here would be through the use of multiplicative spatial random effects. In this case the intensity generating the spatio-temporal counting process would have the following form

$$\lambda(\mathbf{s}, t) = \exp\{g(\mathbf{s})\}\phi(\mathbf{s}, t)$$

where $\mathbf{s} = [x, y, z]^T$ a vector containing the spatial coordinates, $g(\mathbf{s})$ a mean zero stochastic spatial process with specified correlation structure and ϕ a smooth function over time and space. The space-time relationship ϕ could be modeled using multivariate smoothing techniques such as kernel methods like kriging, thin plate or tensor product splines (Schimek, 2000). Extensions of univariate smoothing to the multivariate setting required for such a modeling strategy are in general intuitive and straightforward. These methods however face practical computational challenges requiring careful and efficient implementation since large datasets are quite common in this area of study.
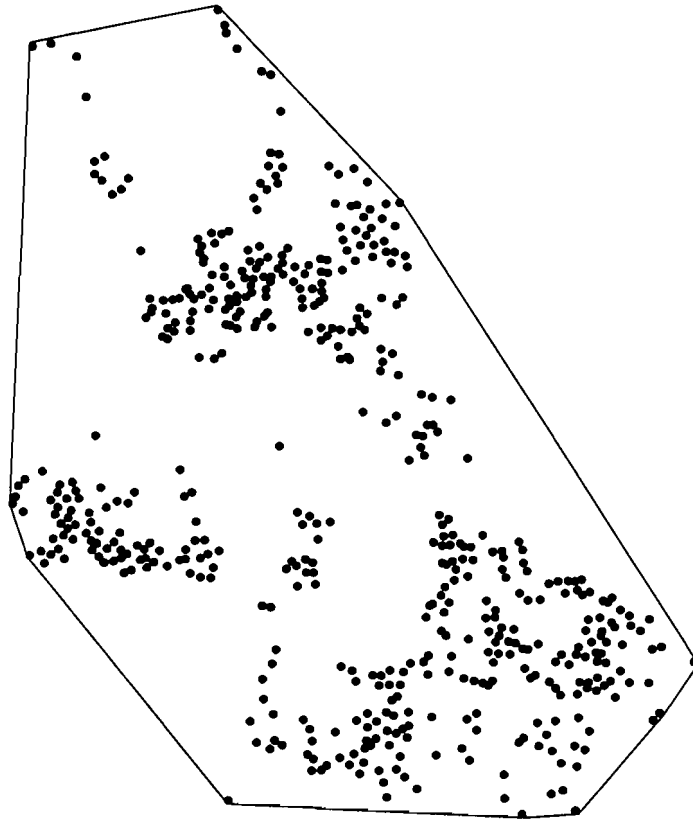
Figure 5.1: The spatial location of trees (•) in the codling moth SIR study.

5.3 Design of Recurrent Event Experiments

A topic of practical importance that has received very little attention in the literature is the design of recurrent event experiments. Dean and Balshaw (1997) explore the loss of efficiency in the analysis of end-of-follow-up counts versus event times in recurrent event studies and established conditions for full efficiency. An investigation of the loss of efficiency from the analysis of panel data panel data versus full event times was considered by Balshaw and Dean (2001). Extending such considerations to the models considered in this thesis would be useful. In addition, it would be helpful to develop strategies for optimal allocation of followup times and sample size at the various followup times. Some individuals may be observed less frequently than others to provide a scheme whereby the data is rich at specific times and less so where it need not be.

The functional approach developed here is particularly suited for determining efficient selection of followup times based on a priori information on the expected form of the recurrence rate. Such information may be available from pilot studies, for example. A more complicated scenario might involve adapting the study design after a fixed number of enrollments and based on preliminary analysis of the data adapt the followup schedule accordingly.

To illustrate suppose that a small exploratory study were conducted which indicated that the recurrence rate of individuals in this cohort were observed to follow a specific functional form $f(t)$. Further suppose that we were going to model f using a cubic spline so that f would be expressed as

$$f(t) = \sum_{h=1}^{k+4} \psi_h B_{h,\boldsymbol{\xi}}(t) \tag{5.7}$$

where $B_{h,\boldsymbol{\xi}}(t)$ denotes the cubic spline basis functions with defining knot sequence $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_k]^T$, k the number of knots and ψ_h the spline coefficients. The question of optimal number and position of follow-up times is now related to the "free-knot" spline smoothing approach discussed in Chapter 1. The idea being that one would want shorter followup when there is higher expected curvature in the rate function which is also where extra knots are warranted. Figure 5.2 gives a graphical demonstration of the postulated best knot positions for fitting a hypothesized recurrence pattern. The optimal number of knots may be determined by adding knots until a desired precision is obtainable. It seems reasonable that these knots could be used as a guideline for scheduling followup times. The exact relationship between knots and followup schedule and how this interaction may lead to

more efficient design is worthy of further theoretical investigation.
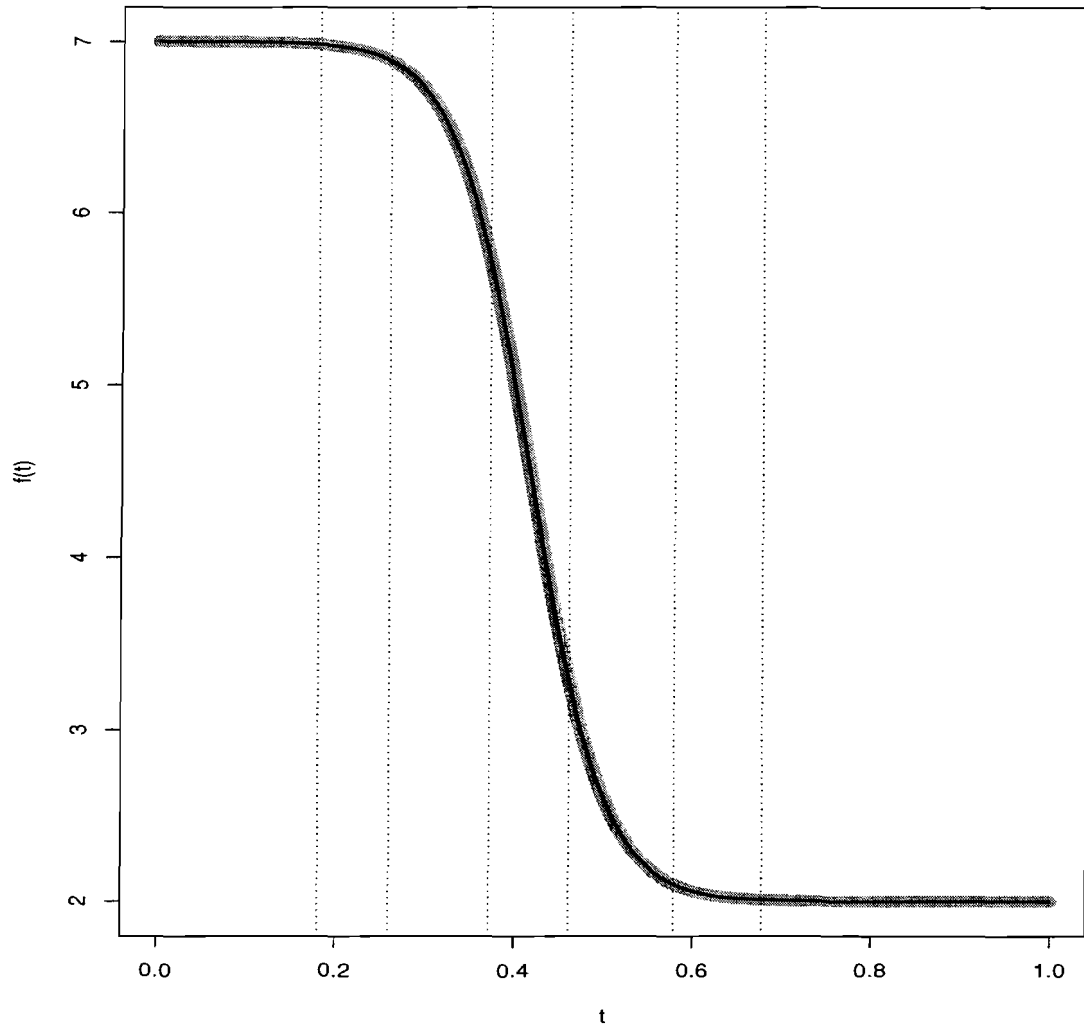
Figure 5.2: Plot of a hypothetical recurrence function displayed by the thick grey line along with the best fitting six knot B-spline approximation (———) with the knot positions indicated by the dotted vertical lines.

Bibliography

Abramowitz, M. and Stegun, I. A., editors (1984). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. John Wiley & Sons.

Abu-Libdeh, H., Turnbull, B. W. and Clark, L. C. (1990). Analysis of multi-type recurrent events in longitudinal studies: Application to a skin cancer prevention trial. *Biometrics* **46**, 1017–1034.

Andersen, P. K., Borgan, O., Borgan, O., Borgan, O., Gill, R. D. and Keiding, N. (1995). *Statistical Models Based on Counting Processes*. Springer-Verlag.

Balshaw, R. F. and Dean, C. B. (2001). Efficiency recovered by the analysis of panel count data rather than end-of-follow-up counts in studies of recurrent events. *Unpublished manuscript* .

Balshaw, R. F. and Dean, C. B. (2002). A semiparametric model for the analysis of recurrent-event panel data. *Biometrics* **58**, 324–331.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L. and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* **92**, 1375–1386.

Beliakov, G. (2004). Least squares splines with free knots: global optimization approach. *Applied Mathematics and Computation* **149**, 783–798.

Böhning, D. (2003). The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components. *Statistics and Computing* **13**, 257–265.

Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association* **85**, 565–571.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38–44.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *Annals of Statistics* **17**, 453–510.

Burnett, R., Bartlett, S., Krewskil, D., Robert, G. and Raad-Young, M. (1994). Air pollution effects on hospital admissions: a statistical analysis of parallel time series. *Environmental and Ecological Statistics* **1**, 325–332.

Chandrasekar, B. and Kale, B. K. (1984). Unbiased statistical estimation functions for parameters in presence of nuisance parameters. *Journal of Statistical Planning and Inference* **9**, 45–54.

Cook, R. J. and Lawless, J. F. (2002). Analysis of repeated events. *Statistical Methods in Medical Research* **11**, 141–166.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377–403.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association* **82**, 1079–1091.

de Boor, C. (1978). *A practical guide to splines.* Springer-Verlag.

Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* **84**, 467–472.

Dean, C. B. (1991). Estimating equations for mixed Poisson models. In V. P. Godambe, editor, *Estimating Functions.* Oxford University Press, pages 35–46.

Dean, C. B. (1998). Overdispersion. In P. Armitage and T. Colton, editors, *Encyclopedia of Biostatistics.* John Wiley & Sons, pages 467–472.

Dean, C. B. and Balshaw, R. (1997). Efficiency lost by analyzing counts rather than event times in Poisson and overdispersed Poisson regression models. *Journal of the American Statistical Association* **92**, 1387–1398.

Demidenko, E. (2004). *Mixed models.* John Wiley & Sons.

Efron, B. (1987). *The Jackknife, the Bootstrap, and Other Resampling Plans.* Society for Industrial & Applied Mathematics.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–102.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics* **31**, 1208–1211.

Golub, G. and Pereyra, V. (2003). Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems* **19**, R1–R26.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman & Hall.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer-Verlag.

Hansen, M. H. and Kooperberg, C. (2002). Spline adaptation in extended linear models. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics* **17**, 2–51.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–338.

Hausman, J., Hall, B. H. and Griliches, Z. (1984). Econometric models for count data with an application to the patents-r&d relationship. *Econometrica* **52**, 909–38.

Heckman, N. E. and Ramsay, J. O. (2000). Penalized regression with model-based penalties. *Canadian Journal of Statistics* **28**, 241–258.

Heyde, C. C. (1997). *Quasi-likelihood and its application*. Springer-Verlag.

Hougaard, P., Lee, M.-L. T. and Whitmore, G. A. (1997). Analysis of overdispersed count data by mixtures of Poisson variables and Poisson processes. *Biometrics* **53**, 1225–1238.

Hu, F. and Kalbfleisch, J. D. (2000). The estimating function bootstrap. *Canadian Journal of Statistics* **28**, 449–481.

Hu, X. J., Sun, J. and Wei, L.-J. (2003). Regression parameter estimation from panel counts. *Scandinavian Journal of Statistics* **30**, 25–43.

Jiang, W., Turnbull, B. W. and Clark, L. C. (1999). Semiparametric regression models for repeated events with random effects and measurement error. *Journal of the American Statistical Association* **94**, 111–124.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*, volume 9. Springer-Verlag.

Jørgensen, B. and Knudsen, S. J. (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics* **31**, 93–114.

Jupp, D. L. B. (1978). Approximation to data by splines with free knots. *SIAM Journal on Numerical Analysis* **15**, 328–343.

Karlin, S. and Taylor, H. M. (1975). *A first course in stochastic processes*. Academic Press, second edition.

Kendall, M. G. (1951). Regression, structure and functional relationship. I. *Biometrika* **38**, 11–25.

Kendall, M. G. (1952). Regression, structure and functional relationship. II. *Biometrika* **39**, 96–108.

Kooperberg, C., Bose, S. and Stone, C. J. (1997). Polychotomous regression. *Journal of the American Statistical Association* **92**, 117–127.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association* **90**, 78–94.

Lawless, J. F. (1987a). Regression methods for Poisson process data. *Journal of the American Statistical Association* **82**, 808–815.

Lawless, J. F. (1987b). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* **15**, 209–225.

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 Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canadian Journal of Statistics* **26**, 549–565.

Lele, S. R. (2003). Impact of bootstrap on the estimating functions. *Statistical Science* **18**, 185–190.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B* **62**, 711–730.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B* **61**, 381–400.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Annals of Statistics* **11**, 86–94.

Lindstrom, M. J. (1999). Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics* **8**, 333–352.

Lipsitz, S. R., Dear, K. B. G. and Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics* **50**, 842–846.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association* **92**, 107–116.

Mao, W. and Zhao, L. H. (2003). Free-knot polynomial splines with confidence intervals. *Journal of the Royal Statistical Society, Series B* **65**, 901–919.

McCullagh, P. and Nelder, J. A. (1999). *Generalized Linear Models*. Chapman & Hall.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley-Interscience.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley & Sons.

Molinari, N., Daures, J.-P., Durand, J.-F., Daurès, J.-P. and Durand, J.-F. (2001). Regression splines for threshold selection in survival data analysis. *Statistics in Medicine* **20**, 237–247.

Nielsen, J. D. and Dean, C. B. (2005). Regression splines in the quasi-likelihood analysis of recurrent event data. *Journal of Statistical Planning and Inference* **134**, 521–535.

Nocedal, J. and Wright, S. J. (1999). *Numerical optimization*. Springer-Verlag.

Pepe, M. S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88**, 811–820.

Pintore, A., Speckman, P. and Holmes, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika* **93**, 113–125.

Pittman, J. (2002). Adaptive splines and genetic algorithms. *Journal of Computational and Graphical Statistics* **11**, 615–638.

Rosen, O., Jiang, W. and Tanner, M. A. (2000). Mixtures of marginal models. *Biometrika* **87**, 391–404.

Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics* **51**, 874–887.

Ross, S. M. (1990). *A Course in Simulation*. Macmillan Publishing Co.

Ross, S. M. (1996). *Stochastic processes*. John Wiley & Sons, second edition.

Ruppert, D. and Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics* **42**, 205–223.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric regression*, volume 12. Cambridge University Press.

Schimek, M. G. (2000). *Smoothing and Regression: Approaches, Computation, and Application*. John Wiley & Sons.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. Wiley-Interscience.

Schwetlick, H. and Schütze, T. (1995). Least squares approximation by splines with free knots. *BIT* **35**, 361–384.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Annals of Statistics* **20**, 1768–1802.

Sichel, H. S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society, Series A* **137**, 25–34.

Sleeper, L. A. and Harrington, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association* **85**, 941–949.

Staniswalis, J. G., Thall, P. F. and Salch, J. (1997). Semiparametric regression analysis for recurrent event interval counts. *Biometrics* **53**, 1334–1353.

Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society, Series B* **62**, 293–302.

Thall, P. F. (1988). Mixed Poisson likelihood regression models for longitudinal interval count data (Corr: V45 p1039). *Biometrics* **44**, 197–209.

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657–671.

Tierney, L., Kass, R. E. and Kadane, J. B. (1989). Approximate marginal densities of nonlinear functions. *Biometrika* **76**, 425–433.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial & Applied Mathematics.

Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics* **15**, 443–462.

Wang, P., Puterman, M. L., Cockburn, I. and Le, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics* **52**, 381–400.

Wecker, W. E. and Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association* **78**, 81–89.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–26.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics* **14**, 1261–1295.