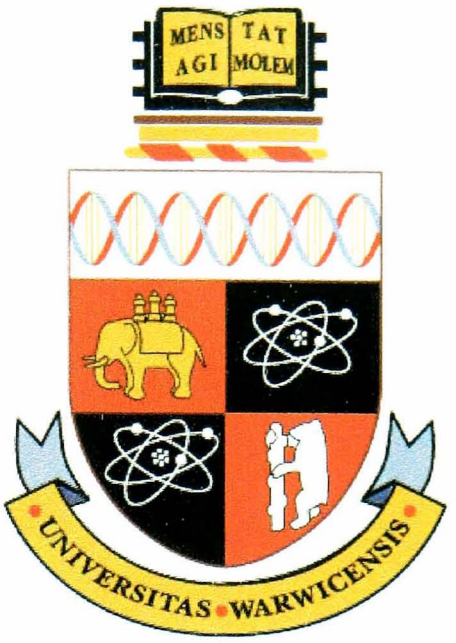


Aspects of competing risks survival analysis

by

Simon James Bond B.A., M.Sc.

Thesis

Submitted to The University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

March 2004

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would like to acknowledge the sponsorship provided by the Engineering and Physical Sciences Research Council and the grant provided by the Knowle Hill School Fund. I thank my supervisor, Ewart Shaw for his advice, along with the other members of staff at the Department of Statistics at the University of Warwick, particularly, John Copas, Jane Hutton and Jim Smith. The support of my parents, family and friends is gratefully acknowledged also. Finally I would like to dedicate this thesis to the memory of John Payne.

Declarations

I hereby declare that this is my own work, except where explicitly stated, and that it has not been submitted for a degree at the University of Warwick or any other university.

Simon James Bond

March 2004

Abstract

This thesis is focused on the topic of *competing risks survival analysis*. The first chapter provides an introduction and motivation with a brief literature review. Chapter 2 considers the fundamental functional of all competing risks data: the crude incidence function. This function is considered in the light of the counting process framework which provides powerful mathematics to calculate confidence bands in an analytical form, rather than bootstrapping or simulation.

Chapter 3 takes the Peterson bounds and considers what happens in the event of covariate information. Fortunately, these bounds do become tighter in some cases. Chapter 4 considers what can be inferred about the effect of covariates in the case of competing risks. The conclusion is that there exist bounds on any covariate-time transformation. These two preceding chapters are illustrated with a data set in chapter 5.

Chapter 6 considers the result of Heckman and Honoré (1989) and investigates the question of their generalisation. It reaches the conclusion that the simple assumption of a univariate covariate-time transformation is not enough to provide identifiability.

More practical questions of modeling dependent competing risks data through the use of frailty models to induce dependence is considered in chapter 7. A practical and implementable model is illustrated.

A diversion is taken into more abstract probability theory in chapter 8 which considers the Bayesian non-parametric tool: Pólya trees. The novel framework of this tool is explained and some results are obtained concerning the limiting random density function and the issues which arise when trying to integrate with a realised Pólya distribution as the integrating measure.

Chapter 9 applies the theory of chapters 7 and 8 to a competing risks data set of a prostate cancer clinical trial. This has several continuous baseline covariates and gives the opportunity to use a frailty model discussed in chapter 7 where the unknown frailty distribution is modeled using a Pólya tree which is considered in chapter 8.

An overview of the thesis is provided in chapter 10 and directions for future research are considered here.

Abbreviations

C.D.F.	cumulative distribution function
M.G.F.	moment generating function
M.C.M.C.	Markov chain Monte Carlo
A.I.C.	Akaike's information criterion
s.d.	standard deviation
s.e.	standard error
G.L.M.	generalised linear model
G.L.M.M.	generalised linear mixed model
Var	variance
Cov	covariance
I	indicator function
$\mathbb{R}$	real numbers
$\mathbb{R}_+$	positive real numbers
$\mathbb{R}^p$	p-fold product-space of real numbers
Γ	the gamma function
$\mathbb{P}$	probability measure

Ω	sample space
$\mathbb{E}$	expectation operator
Σ	covariation process or matrix
σ	standard deviation
n	sample size
L	likelihood
l	log-likelihood
$N(\cdot, \cdot)$	normally distributed random variable
Φ	C.D.F. of a standard normal random variable
Z, X	(unobserved) covariates
z, x	observed covariates
η	a linear predictor
β	(fixed effects) coefficients
b	random frailties
S	survival function
T	random failure time
$T^{\min}$	earliest failure time
λ	hazard function
Λ	cumulative hazard function
Q	crude incidence function
F	cause-specific survival function
s, t, u, v, w	fixed time points

$\mathbf{t}$	vector of fixed time points
$\mathbb{1}$	vector of 1s
i, j, k	integer indices
f, g	density functions
C	random cause of failure (excluding chapter 8)
N	count, or counting process
Y	'still at risk' indicator
δ	censoring indicator
$\mathcal{P}$	a Pólya tree or a realised random probability measure
Π	a partitioning
π	a specific partition
C	a random probability (chapter 8)
$\mathcal{A}$	a set of random probabilities
f_∞	limiting random density

Chapter 1

Introduction

Survival analysis is concerned with time-to-event studies. These commonly arise in medical trials where the interest is in how long patients survive under different conditions; they also have applications in reliability studies in the field of engineering where the question may be how long, or how much use can be obtained, from a component; another common use is in econometrics where the interest is in the duration of a person's employment or unemployment. The common theme is a random variable distributed on the positive reals, which may be observed exactly, or may be censored where the data only tells that the random variable is greater than an observed cut-off point. A natural extension to this framework is to observe the cause of failure as well as the time of failure. For example, in a medical study we observe that the patient may die from several possible diseases. One way of modeling such data is to assume that each individual has a vector of *latent failure times* $(T_1, \ldots, T_k)$ for each of the possible causes of failure, labeled $1, \ldots, k$, and that we observe the earliest of the T_i s . The meaning of the term *competing risks* is that once the earliest failure has occurred it is no longer possible to observe any of the other failure times. As a concrete example, once a patient has died

from lung cancer we have no opportunity of knowing when they would have died from heart failure.

This thesis is concerned with developing practical tools for the analysis of such data sets.

In any statistical analysis the initial steps should be to perform exploratory analysis, where we try to obtain a general picture of the data and spot any gross features. In competing risks this should manifest itself in the examination of the *crude incidence curves*. This is the estimation of the function of time, t, and cause, k,

$$Q_k(t) = \mathbb{P}\{\text{failure time } < t \cap \text{ cause } = k\},$$

this is clearly what the data tells, and makes no assumptions about the existence of latent failure times and any inter-dependence. In practice the tool most commonly, and erroneously, used is the Kaplan-Meier estimate which, to have any interpretation, relies on an assumption that the latent failure times are independent. It can be speculated that the reason for this is the lack of software to calculate the crude incidence as opposed to the abundance of software for the Kaplan-Meier estimate, along with a lack of variance estimates, confidence intervals, and confidence bands. The theory of counting processes is used in chapter 2 to derive a suitable estimator for the crude incidence function along with its asymptotic properties . To use martingale and stochastic process theory to describe a univariate random variable may seem excessive but the payoff is that it is easy to form variance estimates and confidence intervals. In addition it is possible to form confidence bands, in other words, probability statements about the entire incidence curve rather than statements about a single point, which was previously impossible without the use of counting process theory. The purpose of the chapter is to present together in one place the counting process theory applied to competing risks and as such it is not particularly novel.

Within the latent failure time framework it is natural to ask what dependence structure exists, and to consider what is the marginal distribution of the failure times, as this could lead to inference about the effect of an intervention which removed a cause of failure, such as the introduction of a vaccine. However this is the fundamental problem with competing risks data. As proven in Crowder (2001), there are infinitely many joint distributions which could give rise to a specified set of crude incidence functions; in particular given a joint distribution and its crude incidence functions it is always possible to construct another joint distribution that exhibits independence between the latent failure times and has identical crude incidence functions.

However this assumes our data form a homogeneous sample, whereas in reality data sets are rarely of this form and in the case of randomised control trials typically have covariates and treatments. In addition expert opinion and theory from the relevant discipline may provide reasons for assuming a particular dependence structure. With extra data and assumptions we can make headway against the problem of non-identifiability.

In chapter 3 we consider what can be said about the marginal distribution of the latent failure times in the context of two-armed randomised trial. With a homogeneous sample there exist an upper and lower bound between which a marginal distribution must lie. These bounds are invariant of sample size and, typically, are too far apart to be of practical use. They represent what the marginals are under the most extreme possible forms of dependence and were published in Peterson (1976). However if the difference between the two-arms of a trial can be described by a transformation on the time scale, and this transformation is known then we can improve these bounds in some regions. This is a new and valuable result.

Having considered what can be inferred if a time transformation is known, chapter 4 considers what can be said about an unknown time transformation in the case of

competing risks, in a two-armed trial. It turns out that there are also a set of bounds, which, given the examples considered, could be of practical use. This also is a new and valuable result.

In chapter 6 the question of identifiability is considered in depth. The starting point is the important result in Heckman and Honoré (1989) who show that the non-identifiability problem goes away if two assumptions are made: first, that the time transformation has a diagonal derivative matrix; second, that the influence of the covariates can be represented through the proportional hazards structure. Here we demonstrate that, assuming we know the underlying dependence structure such as the Copula representation, then we can identify a generalised covariate-time transformation without assuming a proportional hazards structure. A second theorem of chapter 6 shows that we can go in the other direction: given a general covariate-time transformation, we can identify the dependence structure. Unfortunately we cannot then tie these two theorems together and obtain identifiability with a generalised covariate-time transformation; an extra assumption, such as proportional hazards, is needed. These three theorems are new and shed light on what exactly is the boundary between identifiability and non-identifiability in terms of data and assumptions.

In chapter 7 we return to the world of practical applications of the identifiability result. In particular we focus on assuming any dependence is due to some unobserved covariate which would give conditional independence. This is effectively a random effects or frailty model where dependence is induced by having to 'integrate out' the unobserved frailty. In practice this integration is problematic since it can be over a dimension that is proportionally increasing with the size of the data set. Hence there are a large number of possible methods that approximate the integral in a manner that is amenable to its subsequent maximisation. These are surveyed in this chapter 7. A minor novel

development is proposed in the algorithm for maximising the likelihood that uses the interval bisection algorithm rather than the Newton-Raphson algorithm.

Throughout chapter 7 it is assumed that the unobserved frailty variables are distributed according to a prespecified parametric family of distributions such a log-normal or gamma. However, the effect of this assumptions needs to be considered. Chapter 8 considers a novel Bayesian infinite-parametric distribution: a Pólya tree. This can be used in place of the parametric frailty distribution. The new results presented here concern how to perform integration with respect to a Pólya tree and how to set the parameters so as to give a clear interpretation of the strength of the prior.

In chapter 9 we use the tools considered in chapters 7 and 8 to analyse a randomised control trial with several competing risk end-points and a sufficient number of covariates to permit identifiability. Both classical parametric and Bayesian infinite-parametric analyses are presented and compared.

The last chapter provides an overview of the thesis and considers possible areas for future research.

Chapter 2

Asymptotic bounds on the crude incidence function

2.1 Introduction and Motivation

Given a data set, any competent statistician initially tries to look at the data in an exploratory fashion making as few assumptions as is possible. The aim being to detect gross features of the data, to spot outliers and data-entry errors, and to formulate a general problem. In competing-risks survival analysis the basic tool that should be used is the crude incidence function, also known as the *sub-distribution* function (Crowder 2001) or *occurrence probability*. This is defined as the probability of observing failure from a specified cause before a fixed time. As such it is clearly estimable from a competing risks data set and has a meaning that is easily interpretable by non-statisticians, by statements such as, "Given a hundred patients we expect x of them to die of prostate cancer before 2 years."

However at present, the crude incidence function is competing, in the scientific

literature, with the Kaplan-Meier estimate (Kaplan and Meier 1958) where all causes other than one of interested are treated as censored observations. The problem with this is that it implicitly makes untestable and, *a priori*, unlikely assumptions of independence between latent failure times. Perhaps more clearly, the use of the Kaplan-Meier estimate assumes that the censoring mechanism acts independently of the failure time, but this is a strong assumption in the competing risks setting. If independence does hold then the Kaplan-Meier estimate can be interpreted as the marginal survival distribution of the multivariate latent failure time distribution. This is less simple to convert into plain English statements for non-statisticians.

I would speculate that the major reason for the current prevalence of incorrectly using the Kaplan-Meier estimate is its widespread availability in most statistical software packages. Three years ago in 2000 the author was unaware of any software which calculated estimates for the cumulative incidence. Only since 2002 has there been software available in R (Ihaka and Gentleman 1996) that computes such estimates.

The major point of this chapter is to consider the mathematical properties of the conventional estimator and to put confidence intervals and confidence bands on these estimates. Most of the theoretical mathematical work on counting processes should be credited to Aalen (1978), with important summary monographs in Andersen, Borgan, Gill and Keiding (1993) , Fleming and Harrington (1991), and Jacobsen (1982). The major aim is to present the proofs and results concerning the crude incidence together in a self-contained unit rather than the partial results spread across the statistical literature. The work produced here on confidence bounds is an improvement on the current practice of simulation or bootstrapping as it is easier to calculate. Also, there is a danger that if a non-statistician is presented with a set of pointwise confidence intervals for an estimated function, then he or she will incorrectly interpret it as a confidence band, since the

(Bayesian) question "What is the probability that the curve lies in this region?" is far more natural than trying to consider each time-point in turn and pretending to ignore the remaining time points.

The mathematical properties of the standard estimator of the crude incidence function (Prentice and Kalbfleisch 1978) are consistency and weak convergence to a Gaussian process. These mathematical results can be used to derive confidence intervals for the value of the crude incidence at individual time points and confidence bands for time intervals, both of which are valuable additions to the statistical tool under consideration, and a practical example is presented at the end of the chapter.

The mathematical part of the chapter uses the theory of counting processes, as exemplified in Andersen et al. (1993) and Fleming and Harrington (1991), to represent the competing risks data set. Then martingale theory is used to prove consistency in conjunction with Lenglart's inequality. To prove weak convergence, the martingale central limit theory is used, but the real benefit of the counting process/martingale approach is the ease with which we can derive variance/covariance matrices for estimators when they are in the form of stochastic integrals.

2.2 Definitions

To start with, we will not be concerned with covariates and shall assume that the data form a homogeneous sample from a multivariate density on the positive reals, $(T_{i1}, \ldots, T_{ik}), i = 1, \ldots, n$, which represent the the latent failure times from k causes of failure on a sample of n individuals. The data are observed in the form $(T_i^{\min}, C_i, \delta_i)$, where $T_i^{\min} = \min(T_{i1}, \ldots, T_{ik}, U_i)$ and $C_i = \arg\min(T_{i1}, \ldots, T_{ik}) \times \delta_i$, $\delta_i = I(U_i = T_i^{\min})$, where U_i is a censoring variable. From this we wish to define a set of counting processes $(N_{i1}(t), \ldots, N_{ik}(t))$, and $(Y_{i1}(t), \ldots, Y_{ik}(t))$, where $N_{ik}(t) = I(T_{ik} \leq$

$t, C_i = k)$ and $Y_{ik}(t) = I(t \leq T_i^{\min})$. Because we are restricting attention to a competing risks setting, the value of $Y_{ij}(t)$ is identical for all j, so henceforth we will just refer to $Y_i(t)$. Intuitively, $N_{ij}(t)$ starts at zero and jumps to one at time $T_{ij} = T_i^{\min}$ if the cause of failure $C = j$, if not it remains at zero for all time; the process $Y_i(t)$ indicates if the individual remains observed at time t. We are interested in estimating the function $F_j(t)$ which is defined as $\mathbb{P}\{T^{\min} < t, C = j\}$.

Next we define the process $M_{ij} = N_{ij} - \int_0^t Y_i d\Lambda_j$, where $\Lambda_j(t)$ represents the cumulative cause-specific hazard, $\int_0^t \lambda_j(s)ds, \lambda_j(t) = \lim_{h\to 0} \mathbb{P}(t < T_i^{\min} < t+h, C_i = j | T_i^{\min} > t)/h$, for all i. It will be shown that M_{ij} are square integrable orthogonal martingales, and much use of this will be made to study the asymptotic properties of the conventional estimates from the competing risks literature. To finish this section of definitions we will define $N_{\cdot j} = \sum_{i=1}^n N_{ij}, Y_{\cdot} = \sum_{i=1}^n Y_i$ the number of patients still at risk, and $M_{\cdot j} = \sum_{i=1}^n M_{ij}$, similarly $M_{\cdot\cdot}$, and $N_{\cdot\cdot}$ are defined as summation over causes, $j = 1, \ldots, k$ for $M_{\cdot j}$, and $N_{\cdot j}$ respectively.

2.3 Martingale Properties

In this section we will prove that the process $M_{ij}(t)$, as defined above, is a square integrable martingale and we will also derive an expression for the covariance process. These properties are essential in deriving the asymptotic properties of the crude incidence function. A comprehensive and concise review of the martingale theory used is given in section II.3, pp. 64, of Andersen et al. (1993)

Proposition 2.3.1. *Under the assumption that $\Lambda_j(t) < \infty$, it follows that*

$$M_{ij}(t) = N_{ij}(t) - \int_0^t Y_i(u) d\Lambda_j(u)$$

is a martingale with respect to the filtration

$$\mathcal{F}_t = \sigma\{N_{ij}(s), Y_i(s); 0 \leq s \leq t, i = 1, \ldots, n, j = 1, \ldots, k\}.$$

Proof. To show that $\mathbb{E}|M_{ij}(t)| < \infty$ is trivial since both N_{ij} and Y_i only take values of 0 or 1, and by assumption $\Lambda_j(t) < \infty$. Since Λ_j is a deterministic function, it is clear that $M_{ij}(t)$ is measurable with respect to $\mathcal{F}_t$.

It remains to show that $\mathbb{E}(M_{ij}(t)|\mathcal{F}_s) = M_{ij}(s)$ for all $0 < s < t$. To do this, we must think in terms of the original definition of the N_{ij} and Y_j in terms of $T_i^{\min} = \min(T_{i1}, \ldots, T_{ik})$ and $C_i = \arg\min(T_{i1}, \ldots, T_{ik})$, where for the purposes of this proof we simply treat censoring as just another cause. Now, conditional on $\mathcal{F}_s$, we can identify the occurrence of two complementary events: $\{s > T_i^{\min}\}, \{s < T_i^{\min}\}$. If the first event has occurred—the failure has happened—then we also have observed the value of C_i, whereas if the second event has occurred, then we do not know the value of C_i. The need to know the value of C_i in the case of the first event is why we need the larger filtration generated by the N_{ik} for *all* possible causes, k, rather than just the filtration generated by N_{ij} for the j of interest. Before continuing with the proof, I will restate the definition of the cause-specific hazard,

$$\lambda_j(t) = \lim_{\delta t \to 0} \mathbb{P}(t < T^{\min} < t + \delta t \cap C = j | T^{\min} > t)/\delta t.$$

Consider the two events:

1. $\{T_i^{\min} < s\}$,
   in this case $M_{ij}(s) = M_{ij}(t) = I\{C_i = j\} - \int_0^{T_i^{\min}} \lambda_j(u)du$.

2. $\{s < T_i^{\min}\}$,

here, $M_{ij}(s) = 0 - \int_0^s \lambda_j(u)du$, whereas,

$$\begin{aligned}
\mathbb{E}(M_{ij}(t)|\mathcal{F}_s) &= \mathbb{P}(T_i^{\min} < t \cap C_i = j | s < T_i^{\min}) - \int_0^s \lambda_j(u)du \\
&\quad - \int_s^t \mathbb{P}(T_i^{\min} > u | T_i^{\min} > s)\lambda_j(u)du, \\
&= 0 - \int_0^s \lambda_j(u)du + \frac{\mathbb{P}(s < T_i^{\min} < t \cap C_i = j)}{\mathbb{P}(s < T_i^{\min})} \\
&\quad - \int_s^t \frac{\mathbb{P}(T_i^{\min} > u)}{\mathbb{P}(T_i^{\min} > s)} \frac{f_j(u)}{\mathbb{P}(T_i^{\min} > u)}du,
\end{aligned}$$

where f_j is the cause-specific density defined as

$$\lim_{\delta t \to 0} \mathbb{P}(t < T_i^{\min} < t + \delta t \cap C_i = j)/\delta t.$$

So the first two terms equate to $M_{ij}(s)$; in the last term, the integrand simplifies through cancellation and the denominator is a constant, so, integrating f_j from s to t gives us $\mathbb{P}(s < T_i^{\min} < t \cap C_i = j)$. Hence the last two terms cancel leaving just $M_{ij}(s)$ as desired.

□

Finally, we will calculate an expression for $\langle \mathbf{M} \rangle(t)$, the $k \times k$ matrix of predictable variation processes $\langle M_{ip}, M_{iq} \rangle(t),\ p, q \in \{1, 2, \ldots, k\}$.

Proposition 2.3.2. *Assuming that $T_{ip} \neq T_{iq}$ almost surely, and that Λ_j are continuous, the predictable variation process for $p, q \in \{1, 2, \ldots, k\}$ is,*

$$\langle M_{ip}, M_{iq} \rangle(t) = \begin{cases} \int_0^t Y_i(u)d\Lambda_j(u) & p = q = j \\ 0 & p \neq q. \end{cases}$$

Proof. Using the standard "integration by parts" result for right continuous, bounded variation processes, and suppressing subscript i,

$$
\begin{aligned}
M_p(t)M_q(t) &= M_p(t)M_q(t) - M_p(0)M_q(0) \\
&= \int_0^t [M_p(u-)dM_q(u) + M_q(u)dM_p(u)] \\
&= \int_0^t [M_p(u-)dM_q(u) + M_q(u-)dM_p(u)] + \sum_{u \le t} \Delta M_p(u)\Delta M_q(u),
\end{aligned}
$$

where ΔM_j denotes the jumps sizes at discontinuities of M_j. Now, clearly $\Delta M_j(t) = \Delta N_j(t)$ and since $\mathbb{P}(T_{ip} = T_{iq}) = 0$, in the case of $p \neq q$, this reduces to

$$
M_p(t)M_q(t) - 0 = \int_0^t [M_p(u-)dM_q(u) + M_q(u-)dM_p(u)] .
$$

Since the right hand side is clearly a mean-zero martingale the proof is complete for $p \neq q$.

For $p = q = j$, since $N_j(t)$ is a counting process, we have that

$$
\sum_{u \le t} (\Delta M_j(u))^2 = \sum_{u \le t} (\Delta N_j(u))^2 = \sum_{u \le t} \Delta N_j(u) = N_j(t).
$$

Hence subtracting off $\int_0^t Y(u)d\Lambda_j(u)$ we get

$$
M_j^2(t) - \int_0^t Y(u)d\Lambda_j(u) = 2\int_0^t M_j(u-)dM_j(u) + M_j(t).
$$

□

2.4 Estimators

An estimator of the cumulative cause-specific hazard is,

$$
\widehat{\Lambda}_j(t) = \int_0^t J(s)\frac{dN_{\cdot j}(s)}{Y_{\cdot}(s)},
$$

where we define $J(t) = I(Y_{\cdot}(t) > 0)$ and 0/0=0. It can easily be seen that, since $N_{\cdot j}$ has jumps of size 1 at the failure times, this estimate is equal to the more recognisable

$\sum_{T_i^{\min} < t} I(C_i = j)/Y_\cdot(T_i^{\min}-)$. From this expression it is clear that this estimator of the cause-specific hazard is identical to the conventional Nelson-Aalen estimator (Aalen 1978, Nelson 1969) if we were to treat all failures from causes, other than the one of interest, as cases of censoring. From this observation it follows that the asymptotic properties of the estimator include consistency and weak convergence to a Gaussian processes where the variance of the estimator can be estimated in a number of ways (Klein 1991).

A consistent estimate of the overall survival function $S(t)$ is the Kaplan-Meier estimator, $\widehat{S}(t)$. This is defined to be

$$\widehat{S}(t) = \prod_{i:\, T_i^{\min} < t} \left\{1 - d\widehat{\Lambda}_\cdot(T_i^{\min})\right\},$$

where $d\widehat{\Lambda}_\cdot$ denotes the increments that occur at (non-censored) observed times in the Nelson-Aalen estimate of the overall cumulative hazard function. Its properties, including consistency and weak convergence to a Gaussian process, are well documented (Andersen et al. 1993, Fleming and Harrington 1991).

The crude incidence function is defined as

$$Q_j(t) = \mathbb{P}(T^{\min} < t, C = j) = \int_0^t S(s) d\Lambda_j(s). \tag{2.1}$$

Given the final expression, a proposed estimator is the process

$$\widehat{Q}_j(t) = \int_0^t \widehat{S}(s) J(s) d\widehat{\Lambda}_j(s).$$

The merit of these two estimators, the Kaplan-Meier and the crude incidence function, is that they take into account the assumption of independent censoring. If it cannot be assumed that the censoring is independent then it should be considered as another competing risk. Indeed without any censoring present, or if censoring is treated

as a dependent competing risk, the Kaplan-Meier estimator reduces to the empirical distribution of $T^{\min}$, and the crude incidence estimate is $n^{-1}\sum_i I\{T_i^{\min} < t, C_i = j\}$. However for practical purposes, independent censoring is a very common assumption and the increased precision must be utilised. An example of independent censoring is a clinical trial that has to be analysed and written up for publication; if there are patients who have not experienced the event of interest then they have to be censored and it may be fair to assume that the date chosen to close the trial is independent of the individual patients' outcomes.

2.5 Asymptotic Properties

This section provides the fundamental mathematical results in which we are interested. In 2.5.2 we prove that our estimate of the crude incidence function tends to the true value as the size of the data increases, and in 2.5.4 we show that the error, scaled by $\sqrt{n}$, tends to a Brownian process. Neither the mathematics involved, nor the results themselves, are innovative. The mathematics closely follows the route laid out in Andersen et al. (1993) which proves similar results about the Nelson-Aalen estimator and the Kaplan-Meier estimator, although proposition 2.5.1 is completely the work of the author. The results of consistency and of weak convergence have been stated in the literature, for example Pepe and Mori (1993), but the author is unaware of any self-contained and thorough proof. The main boon of the counting process representation is the ease with which variance estimators of the crude incidence functions can be derived.

2.5.1 Basic Theorems

In this subsection we will present two basic theorems used in the proof of the main result, theorem 2.5.3. The first theorem, a version of Rebolledo's (Rebolledo 1980)

martingale central limit theorem, is taken verbatim from Andersen et al. (1993), the second theorem is known as Lenglart's inequality (Lenglart 1977), and is a useful result for the purposes of proving limits in probability.

Martingale Central Limit Theorem

For each $n = 1, 2, \ldots$, let $\mathbf{M}^{(n)}(t) = (M_1^{(n)}(t), \ldots, M_k^{(n)}(t))$ be a vector of k local square-integrable martingales, defined for $t \in \mathcal{T} = [0, \tau)$ (τ is a fixed, 'termination' time). Also, for each $\epsilon > 0$, let $\mathbf{M}_\epsilon^{(n)}$ be a vector of local square-integrable martingales, containing all the jumps of $\mathbf{M}^{(n)}$ larger in absolute values than ϵ. Write $\langle \mathbf{M}^{(n)} \rangle$ for the $k \times k$ matrix of of predictable variation processes, $\langle M_i^{(n)}, M_j^{(n)} \rangle$.

Now, define $\mathbf{U}$ to be a continuous Gaussian vector martingale with $\langle \mathbf{U} \rangle = [\mathbf{U}] = \mathbf{V}$, a continuous deterministic $k \times k$ positive semi-definite matrix on $\mathcal{T}$. So, $\mathbf{U}(t) - \mathbf{U}(s) \sim N(\mathbf{0}, \mathbf{V}(t) - \mathbf{V}(s))$, and is independent of $(\mathbf{U}(u); u \leq s)$ for all $0 \leq s \leq t$.

For completeness we will define the optional variation process $[\mathbf{M}^{(n)}](t) = \sum_{s \leq t} |\Delta \mathbf{M}^{(n)}(s)|^2$, where $\Delta \mathbf{M}^{(n)}(s) = \lim_{\delta s \to 0} \mathbf{M}^{(n)}(s + \delta s) - \mathbf{M}^{(n)}(s)$ (the discontinuities in $\mathbf{M}^{(n)}$). This can be though of as the empirical, or observed version of the predictable variation process. By $(D(\mathcal{T}))^k$, we mean the space of $\mathbb{R}^k$-valued functions which are right continuous with left-hand limits, defined on $\mathcal{T}$ and endowed with the Skorohod topology (Billingsley 1999, pp. 123-124 and chapter 3).

Theorem 2.5.1 (Rebolledo's theorem). *Let $\mathcal{T}_0 \subseteq \mathcal{T}$ and consider the conditions*

$$\langle \mathbf{M}^{(n)} \rangle(t) \xrightarrow{P} \mathbf{V}(t) \text{ for all } t \in \mathcal{T}_0 \text{ as } n \to \infty, \tag{2.2}$$

$$[\mathbf{M}^{(n)}](t) \xrightarrow{P} \mathbf{V}(t) \text{ for all } t \in \mathcal{T}_0 \text{ as } n \to \infty, \tag{2.3}$$

$$\langle M_{\epsilon i}^{(n)} \rangle(t) \xrightarrow{P} 0 \text{ for all } t \in \mathcal{T}_0, i, \epsilon > 0 \text{ as } n \to \infty, . \tag{2.4}$$

Then either of (2.2), (2.3), together with (2.4), imply

$$(\mathbf{M}^{(n)}(t_1), \ldots, \mathbf{M}^{(n)}(t_m)) \xrightarrow{D} (\mathbf{U}(t_1), \ldots, \mathbf{U}(t_m)) \text{ as } n \to \infty \tag{2.5}$$

for all $t_1, \ldots, t_m \in \mathcal{T}_0$*; moreover, both (2.2),(2.3) then hold. Furthermore, if* $\mathcal{T}_0$ *is dense in* $\mathcal{T}$*, then the same conditions imply*

$$\mathbf{M}^{(n)} \xrightarrow{D} \mathbf{U} \text{ in } (D(\mathcal{T}))^k \text{ as } n \to \infty,$$

and $\langle \mathbf{M}^{(n)} \rangle$ *and* $[\mathbf{M}^{(n)}]$ *converge uniformly in probability to* $\mathbf{V}$ *on compact subsets of* $\mathcal{T}$*.*

Lenglart's Inequality

If $(X(t); 0 \leq t \leq \tau)$ is a local submartingale on $\mathcal{T}$ with a strictly positive compensator $Y(t)$, then for any $\eta > 0$ and $\delta > 0$,

$$\mathbb{P}(\sup_{t \in \mathcal{T}} X(t) > \eta) \leq \delta/\eta + \mathbb{P}(Y(\tau) > \delta). \tag{2.6}$$

This applies, in particular, to $X = \int H dN$ where H is a predictable, non-negative process and N is a counting process; it also applies to M^2 where M is a local martingale, in which case the inequality is,

$$\mathbb{P}(\sup_{t \in \mathcal{T}} |M(t)| > \eta) \leq \delta/\eta^2 + \mathbb{P}(\langle M \rangle(\tau) > \delta).$$

2.5.2 Consistency

In this subsection I will prove that the estimate of the crude incidence function is consistent.

Theorem 2.5.2. *Let* $t \in [0, \tau)$*, where* $\tau = \inf(t : S(t) = 0)$*, and assume that, as* $n \to \infty$*,*

$$\int_0^t \frac{J^{(n)}(u)}{Y_\cdot^{(n)}(u)} d\Lambda_j(u) \xrightarrow{P} 0, \quad j = 1, 2, \ldots, k \tag{2.7}$$

and

$$\int_0^t (1 - J^{(n)}(u)) d\Lambda_\cdot(u) \xrightarrow{P} 0. \tag{2.8}$$

Then, as $n \to \infty$,

$$\sup_{s \in [0,t]} |\widehat{Q}_j^{(n)}(s) - Q_j(s)| \xrightarrow{P} 0.$$

Proof. Begin by observing,

$$\widehat{Q}_j(t) - Q_j(t) = \int_0^t \left[\widehat{S}(s) J(s) d\widehat{\Lambda}_j(s) - S(s) d\Lambda_j(s) \right] \tag{2.9}$$

$$= \int_0^t \frac{\widehat{S}(s)J(s)}{Y_\cdot(s)} dM_{\cdot j}(s) + \int_0^t (\widehat{S}(s) - S(s)) d\Lambda_j(t) - \int_0^t \widehat{S}(s)(1 - J(s)) d\Lambda_j(s). \tag{2.10}$$

Now, considering the expressions on the right hand side in turn, for the first one we can use the martingale version of Lenglart's inequality, 2.5.1, so

$$\begin{aligned} \mathbb{P}\left(\sup_{t \in [0,\tau]} \left| \int_0^t \frac{\widehat{S}(s)J(S)}{Y_\cdot(s)} dM_{\cdot j}(s) \right| > \eta \right) &\leq \delta/\eta^2 + \mathbb{P}\left(\int_0^t \left\{ \frac{\widehat{S}(s)J(s)}{Y_\cdot(s)} \right\}^2 Y_\cdot(s) d\Lambda_j(s) > \delta \right), \\ &= \delta/\eta^2 + \mathbb{P}\left(\int_0^t \frac{\widehat{S}^2(s)J(s)}{Y_\cdot(s)} d\Lambda_j(s) > \delta \right), \end{aligned}$$

hence by condition (2.7), and since $\widehat{S}^2 < 1$,

$$\sup_{t \in [0,\tau]} \left| \int_0^t \frac{\widehat{S}(s)J(S)}{Y_\cdot(s)} dM_{\cdot j}(s) \right| \xrightarrow{P} 0.$$

For the second expression, noting that summing condition (2.7) across j gives

$$\int_0^t \frac{J^{(n)}(u)}{Y_\cdot^{(n)}(u)} d\Lambda_\cdot(u) \xrightarrow{P} 0$$

and hence the conditions for theorem IV.3.1 in Andersen et al. (1993, p. 261) are satisfied, hence $\sup_{s \in [0,\tau]} |\widehat{S}(s) - S(s)| \xrightarrow{P} 0$. Given that $|\widehat{S}(s) - S(s)| \leq 1$, using the dominated convergence theorem, we obtain

$$\int_0^t (\widehat{S}(s) - S(s)) d\Lambda_j(t) \xrightarrow{P} 0.$$

For the final expression,

$$0 \leq \int_0^t \widehat{S}(s)(1 - J(s))d\Lambda_j(s)$$
$$\leq \int_0^t (1 - J(s))d\Lambda_j(s) \xrightarrow{P} 0,$$

by condition (2.8) □

2.5.3 Technical Lemma

This section draws heavily on the theory of product-integration and the functional delta method as outlined in Andersen et al. (1993). The aim is to obtain the result detailed in proposition 2.5.1. This is then used in the proof of theorem 2.5.3, which gives an expression for the weak convergence of the crude incidence function estimator.

The delta method, in its simplest form, assumes a random sequence, X_n, that satisfies

$$a_n(X_n - x) \xrightarrow{D} N,$$

where N is typically a standard normal distribution, x is a fixed point, and a_n is an increasing sequence of constants (typically applied with $a_n = \sqrt{n}$). The result is that

$$a_n(g(X_n) - g(x)) \xrightarrow{D} g'(x)N,$$

where g, g' are a smooth function and its derivative.

In Andersen et al. (1993, page 111, theorem II.8.1) and Gill (1989, theorem 3) this is extended and formalised for X_n that are processes through time and for more general functions that are *compactly differentiable* so as to include $\widehat{S}(\widehat{\Lambda})$, the Kaplan-Meier estimator.

Now consider the mapping $S(\cdot)$ defined as

$$S(\Lambda(t)) = \lim_{\max|u_{i+1}-u_i|\to 0} \prod_{u_0=0}^{u_n=t} (1 - \{\Lambda(u_{i+1}) - \Lambda(u_i)\}) \tag{2.11}$$

The existence and uniqueness of this limit is proven in Gill and Johansen (1990), but if we consider a sequence of $\{u_{n\,i}\}$ such that $\Lambda(u_{n\,i+1}) - \Lambda(u_{n\,i}) = \{\Lambda(t) - \Lambda(0)\}/n$ for all $i, 1 \leq i \leq n$, then the right hand of (2.11) equals $\lim_{n\to\infty}(1 - \{\Lambda(t) - \Lambda(0)\}/n)^n$. If Λ is a continuous function then such a sequence $\{u_{n\,i}\}$ exists and hence $S(\Lambda(t))$ is equal to $\exp(-\{\Lambda(t) - \Lambda(0)\})$. Hence the mapping S gives the conventional definition of the survival function when applied to the true cumulative hazard function, which is assumed to be continuous and satisfies $\Lambda(0) = 0$. Moreover, when S is applied to the Nelson-Aalen estimate of the cumulative hazard, $\widehat{\Lambda}$, which is a step function, then $S(\widehat{\Lambda})$ clearly equates to the Kaplan-Meier estimator.

Now using the standard result with an appropriate sequence of constants, a_n,

$$a_n\{\widehat{\Lambda}_n(t) - \Lambda(t)\} \xrightarrow{D} B(\sigma(t)),$$

where $B(\cdot)$ represents Brownian motion and applying the generalised delta method we have the result that

$$a_n\{S(\widehat{\Lambda}_n(t)) - S(\Lambda(t))\} \xrightarrow{D} dS(\Lambda).B(\sigma(t)),$$

and

$$\left[a_n\{S(\widehat{\Lambda}_n(t)) - S(\Lambda(t))\}\right] - \left[dS(\Lambda).a_n\{\widehat{\Lambda}_n(t) - \Lambda(t)\}\right] \xrightarrow{P} 0.$$

The meaning of dS is defined on the same page in (Andersen et al. 1993) in definition II.8.1 . If we had defined $S(\Lambda)$ as $\exp(-\Lambda)$ then one could derive dS as the derivative, with respect to Λ, of this function and obtain $dS(\Lambda(t)) = -S(\Lambda(t))$. It turns out that this result is true, but as we need to define our mapping S to hold for both the Kaplan-Meier estimate *and* the underlying survival function, the complex definition in (2.11) is required.

To calculate dS use proposition II.8.7 in (Andersen et al. 1993, page 114) in conjunction with the chain rule, to give

$$\begin{aligned}(dS(\Lambda).h)(t) &= -\int_0^t S(s-)\frac{S(t)}{S(s)}dh(s) \\ &= -S(t)\int_0^t dh(s) = -S(t)h(t),\end{aligned}$$

where the continuity of the underlying true survival function allows the cancellation.

All this comes together to provide a lemma that is used in the proof of theorem 2.5.3.

Proposition 2.5.1. *Given a time interval $t \in [0,\tau]$ where $S(t) > 0$ and the assumptions of theorem 2.5.2, then*

$$\left[a_n\{\widehat{S}(t) - S(t)\}\right] - \left[-a_n S(t)\{\widehat{\Lambda}(t) - \Lambda(t)\}\right] \xrightarrow{P} 0; \text{ as } n \to \infty.$$

where a_n is an increasing sequence as will be defined in theorem 2.5.3.

2.5.4 Weak Convergence

In this subsection we will consider the weak convergence of the estimator of the crude incidence function. The proof uses Rebolledo's Central Limit theorem, but what it really shows is a result about the limiting distribution of

$$\widehat{Q}_j^*(t) = \int_0^t J(u)\widehat{S}(u)d\widehat{\Lambda}_j(u),$$

where $J(t) = I\{Y.(t) > 0\}$. To obtain a result of any practical use we need an additional condition which implies that the difference between $\widehat{Q}_j^*$ and Q_j will converge in probability to zero and the 'nuisance' factor, J, can be ignored. To apply the central limit theorem we need to assume that the estimate of the covariance matrix process

converges, and that these estimates converge to a continuous function. These assumptions may appear to be rather convoluted and implausible, but we will provide simple sufficient conditions for their validity at a later point.

Theorem 2.5.3. *Assume that there exists a sequence of positive constants* $\{a_n\}$, *increasing to infinity as* $n \to \infty$, *and a function* $y(t) > 0$ *such that the following exist for all* $t \in [0, \tau]$, *where* $\tau = \inf\{t : S(t) = 0\}$:

$$\begin{aligned}
\Sigma_{11}(t,j) &= \int_0^t S^2(u)/y(u)d\Lambda_j(u) + \int_0^t Q_j^2(u)/y(u)d\Lambda_\cdot(u) \\
&\quad +2\int_0^t S(u)Q_j(u)/y(u)d\Lambda_j(u) \qquad (2.12)\\
\Sigma_{12}(t,j) &= \int_0^t S(u)/y(u)d\Lambda_j(u) + \int_0^t Q_j(u)/y(u)d\Lambda_\cdot(u) \qquad (2.13)\\
\Sigma_{22}(t,j) &= \int_0^t 1/y(u)d\Lambda_\cdot(u), \qquad (2.14)
\end{aligned}$$

where $j = 1, 2, \ldots, k$, *and assume that*

(A)

$$\begin{aligned}
&a_n^2\left\{\int_0^t \widehat{S}^2(u)J(u)/Y_\cdot(u)d\Lambda_j(u) + \int_0^t Q_j^2(u)/Y_\cdot(u)d\Lambda_\cdot(u)\right.\\
&\quad \left.+2\int_0^t \widehat{S}(u)J(u)Q_j(u)/Y_\cdot(u)d\Lambda_j(u)\right\} \xrightarrow{P} \Sigma_{11}(t,j), \qquad (2.15)\\
&a_n^2\left\{\int_0^t \widehat{S}(u)J(u)/Y_\cdot(u)d\Lambda_j(u) + \int_0^t Q_j(u)/Y_\cdot(u)d\Lambda_\cdot(u)\right\} \xrightarrow{P} \Sigma_{12}(t,j), \qquad (2.16)\\
&a_n^2\left\{\int_0^t 1/Y_\cdot(u)d\Lambda_\cdot(u)\right\} \xrightarrow{P} \Sigma_{22}(t,j), \qquad (2.17)
\end{aligned}$$

as $n \to \infty$ *for all* $t \in [0, \tau]$;

(B) For all $\epsilon > 0, i, j = 1, 2, \ldots, k$,

$$a_n^2 \int_0^t \frac{\left[\widehat{S}(u)J(u) + Q_j(u)\right]^2}{Y.(u)} I\left(a_n \frac{\widehat{S}(u)J(u) + Q_j(u)}{Y.(u)} > \epsilon\right) d\Lambda_j(u)$$

$$+a_n^2 \sum_{i \neq j} \int_0^t \frac{Q_j^2(u)}{Y.(u)} I\left(a_n \frac{Q_j(u)}{Y.(u)} > \epsilon\right) d\Lambda_i(u) \xrightarrow{P} 0, \tag{2.18}$$

$$a_n^2 \int_0^t \frac{1}{Y.(u)} I\left(\frac{a_n}{Y.(u)} > \epsilon\right) d\Lambda.(u) \xrightarrow{P} 0, \tag{2.19}$$

as $n \to \infty$ *for all* $t \in [0, \tau]$;

(C) For $j = 1, 2, \ldots, k$, *and all* $t \in [0, \tau]$,

$$a_n \int_0^t (1 - J(u)) d\Lambda_j(u) \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{2.20}$$

Then

$$a_n(\widehat{Q}_j(t) - Q_j(t)) \xrightarrow{D} U_1(t, j) - Q_j(t)U_2(t, j), \text{ as } n \to \infty,$$

where $U_1(t, j), U_2(t, j)$ *is a Gaussian martingale with* $cov(U_1(t_1, j), U_2(t_2, j)) = \Sigma(t_1 \wedge t_2, j)$, *a* 2×2 *matrix. Also, for* $p, q = 1, 2$, *and* $j = 1, 2, \ldots, k$,

$$\sup_{t \in [0,\tau]} \left| a_n^2 \widehat{\Sigma}_{pq}(t, j) - \Sigma_{pq}(t, j) \right| \xrightarrow{P} 0 \text{ as } n \to \infty,$$

where expressions for $\widehat{\Sigma}_{pq}(t, j)$ *can be obtained from the expressions in condition (A) by replacing* Q_j *with* $\widehat{Q}_j$, *and* Λ_j *with* $\widehat{\Lambda}_j$.

Proof. By definition,

$$a_n(\widehat{Q}_j(t) - Q_j(t)) = a_n \left\{ \int_0^t \widehat{S}(u)J(u)d\widehat{\Lambda}_j(u) - \int_0^t S(u)d\Lambda_j(u) \right\}$$

$$= a_n \left\{ \int_0^t \widehat{S}(u)/Y.(u)dM_{.j}(u) + \int_0^t (\widehat{S}(u) - S(u))d\Lambda_j(u) - \int_0^t \widehat{S}(u)(1 - J(u))d\Lambda_j(u) \right\}.$$

By condition (C) we know that the third term tends in probability to zero hence we can ignore this term henceforth. In addition, given the range of S and $\widehat{S}$, we observe $|\widehat{S}(t) - S(t)| \leq 1$ and using the dominated convergence theorem with proposition 2.5.1,

$$\begin{aligned}
&\xrightarrow{P} a_n \left\{ \int_0^t \widehat{S}(u) J(u) / Y.(u) dM._j(u) - \int_0^t (\widehat{\Lambda}.(u) - \Lambda.(u)) S(u) d\Lambda_j(u) \right\} \\
&= a_n \left\{ \int_0^t \widehat{S}(u) J(u) / Y.(u) dM._j(u) - \int_0^t (\widehat{\Lambda}.(u) - \Lambda.(u)) dQ_j(u) \right\}
\end{aligned}$$

through integration by parts

$$\begin{aligned}
&= a_n \left\{ \int_0^t \widehat{S}(u) J(u) / Y.(u) dM._j(u) - Q_j(t)(\widehat{\Lambda}.(t) - \Lambda.(t)) \right. \\
&\qquad \left. + \int_0^t Q_j(u) d(\widehat{\Lambda}.(u) - \Lambda.(u)) \right\} \\
&= a_n \left\{ \int_0^t \frac{\widehat{S}(u) J(u)}{Y.(u)} dM._j(u) + \int_0^t Q_j(u) \frac{dM.(u)}{Y.(u)} - Q_j(t) \int_0^t \frac{dM.(u)}{Y.(u)} \right\}.
\end{aligned}$$

Note that a stochastic integral with respect to a martingale is also a martingale, hence the expression above is a linear combination of martingales. We now use standard stochastic integral theory (Revuz and Yor 1999, chapter IV, section 2, pp.137-145, theorem 2.2) which gives that

$$\langle \int H_1 dB_1, \int H_2 dB_2 \rangle = \int H_1 H_2 d\langle B_1, B_2 \rangle,$$

for predictable processes, H_1, H_2 and martingales B_1, B_2. Noting that the martingales, $M._j$, are orthogonal and that $d\langle M._j \rangle(t) = Y.(t) d\Lambda_j(t)$, we observe that the predictable variation for the stochastic processes,

$$\begin{aligned}
B_1^{(n)}(t) &= a_n \left\{ \int_0^t \frac{\widehat{S}(u) J(u)}{Y.(u)} dM._j(u) + \int_0^t Q_j(u) \frac{dM.(u)}{Y.(u)} \right\} \\
B_2^{(n)}(t) &= a_n \int_0^t \frac{dM.(u)}{Y.(u)},
\end{aligned}$$

equates to the left hand side of condition (A) which, by assumption, converges to $\Sigma(t, j)$, and thus satisfies (2.2) in the central limit theorem.

To check that the continuity condition (2.4) is satisfied, observe that because $M_{\cdot j}$ is associated with a counting process, it has jump sizes of 1. So if we break $M_{\cdot\cdot}$ into its components, $M_{\cdot 1}, \ldots, M_{\cdot k}$, we obtain,

$$B_{1\epsilon}^{(n)}(t) = a_n \left\{ \int_0^t \frac{\left[\widehat{S}(u)J(u) + Q_j(u)\right]}{Y_{\cdot}(u)} I\left(a_n \frac{\widehat{S}(u)J(u) + Q_j(u)}{Y_{\cdot}(u)} > \epsilon \right) dM_{\cdot j}(u) \right\}$$

$$+ a_n \sum_{i \neq j} \int_0^t \frac{Q_j(u)}{Y_{\cdot}(u)} I\left(a_n \frac{Q_j(u)}{Y_{\cdot}(u)} > \epsilon \right) dM_{\cdot i}(u)$$

$$B_{2\epsilon}^{(n)}(t) = a_n \int_0^t \frac{1}{Y_{\cdot}(u)} I\left(\frac{a_n}{Y_{\cdot}(u)} > \epsilon \right) dM_{\cdot\cdot}(u),$$

so using standard stochastic integral theory we see that the predictable variation processes, $\langle B_{1\epsilon}^{(n)} \rangle, \langle B_{2\epsilon}^{(n)} \rangle$, equate to the left hand side of condition (B) which converge in probability to zero, thus satisfying (2.4) of the central limit theorem. Hence we can apply the central limit theorem which proves the part about the weak convergence of the estimator.

The optional variation processes, $\left[B_i^{(n)}, B_j^{(n)} \right](t)$, are given by substituting $\widehat{\Lambda}_j$ for Λ_j in the expressions for $\langle B_i^{(n)}, B_j^{(n)} \rangle(t)$, and since, by theorem 2.5.2, $\widehat{Q}_j \xrightarrow{P} Q_j$ uniformly, we can apply the second part of the central limit theorem to see that

$$\sup_{t \in [0,\tau]} \left| a_n^2 \widehat{\Sigma}_{pq}(t,j) - \Sigma_{pq}(t,j) \right| \xrightarrow{P} 0 \text{ as } n \to \infty.$$

□

2.5.5 Sufficient conditions

The conditions used in deriving theorems 2.5.2 and 2.5.3 were chosen to be the weakest possible. The problem with this is that they are difficult to interpret. Typically we require that the censoring is not too heavy in the time period which we consider. This allows the use of the Glivenko-Cantelli theorem (Billingsley 1995, pp. 268–269) with

$a_n = \sqrt{n}$, which says that the empirical distribution tends to the underlying cumulative distribution function. If we have a censoring mechanism which censors all individuals after a fixed time point, τ, then the resulting $y(t) = 0$, for $t > \tau$. This will mean that the expressions in, say, (2.12) will be the integral of $1/0$, fortunately on a statistical, rather than mathematical, level this makes perfect sense as the censoring means we have no information after time τ. Similarly, if there exist a fixed time τ such that $\Lambda(t) = \infty$, for $t > \tau$, then this implies that every individual will fail before τ with probability 1, hence it is pointless to consider times after τ.

Here will be demonstrated a simple condition for the assumptions of theorems 2.5.2 and 2.5.3.

Proposition 2.5.2. *Assuming there exists a sequence, $a_n \to \infty$ as $n \to \infty$, such that*

$$\frac{Y_{\cdot}^{(n)}(t)}{a_n^2} \xrightarrow{P} y(t), \tag{2.21}$$

where

$$\inf_{s \in [0,t]} y(s) > 0 \tag{2.22}$$

where $y(t)$ is defined in theorem 2.5.3, and that $\Lambda_j(t) < \infty$ for all t, then conditions (A),(B),(C) of theorem 2.5.3 will be satisfied.

Proof. For condition (A), observe that the dominated convergence theorem for the sequence of random variables, $Y_{\cdot}^{(n)}(t)/a_n^2$, gives the result immediately.

For condition (B), for bounded functions $H(t)$, all the integrals are of the form,

$$\int_0^t \frac{H^2(u)a_n^2}{Y_{\cdot}^{(n)}(u)} I\left(\frac{H(u)a_n}{Y_{\cdot}^{(n)}} > \epsilon\right) d\Lambda_j(u) \tag{2.23}$$

$$\xrightarrow{P} \int_0^t \frac{H^2}{y(u)}(u) I\left(\frac{H(u)}{y(u)a_n} > \epsilon\right) d\Lambda_j(u) \tag{2.24}$$

$$\xrightarrow{P} \int_0^t 0 d\Lambda_j(u) = 0 \tag{2.25}$$

since $a_n \to \infty$ as $n \to \infty$.

For condition (C) observe that

$$\mathbb{P}\left\{a_n \int_0^t I(Y_\cdot^{(n)}(u) = 0) d\Lambda(u) > \epsilon\right\} \tag{2.26}$$

$$= \mathbb{P}[\omega : Y_\cdot^{(n)}(u, \omega) = 0] \to 0 \tag{2.27}$$

which gives the desired result. □

Weaker conditions are required to obtain consistency, so a weaker sufficient condition will suffice.

Proposition 2.5.3. *If $Y_\cdot^{(n)} \xrightarrow{P} \infty$ as $n \to \infty$ then conditions (2.7) and (2.8) of theorem 2.5.2 will be satisfied.*

Proof. Since $Y_\cdot^{(n)} \xrightarrow{P} \infty$, which implies that

$$\frac{J^{(n)}(u)}{Y_\cdot^{(n)}(u)} \xrightarrow{P} 0,$$

and hence by the dominated convergence theorem (2.7) is satisfied.

For (2.8) observe that,

$$\mathbb{P}\left\{\int_0^t (1 - J^{(n)}(u)) d\Lambda_\cdot(u) > \epsilon\right\} \tag{2.28}$$

$$= \mathbb{P}[\omega : Y_\cdot^{(n)}(u, \omega) = 0] \to 0, \tag{2.29}$$

thus obtaining the desired result. □

2.6 Applications

Having found a parameter-free distribution to base any inference around, and having found expressions for the covariation process, the standard route to calculating confidence intervals and bands is to perform simulations (Lin 1997). This is because although, pointwise, the estimator has an asymptotic Gaussian error process, the error

process does not have independent increments. This means that well understood results for Brownian motion cannot be applied, and at present, there are no known procedures to derive analytic confidence bands. However I will consider below the circumstances in which Brownian motion coincides or approximates the error process, and will illustrate the mechanics of producing confidence bands and intervals.

Using theorem 2.5.3 we observe that $(U_1(t,j) - Q_j(t)U_2(t,j))/a_n = D_j(t)$ is also a Gaussian random variable with mean zero but with covariance function (with $t < s$)

$$\text{Cov}(D_j(t), D_j(s)) = \Sigma_{11}(t,j) - (Q_j(t) + Q_j(s))\Sigma_{12}(t,j) + Q_j(t)Q_j(s)\Sigma_{22}(t,j),$$

It would be very convenient if all the terms involving s were to disappear, as then it could be inferred that the process has independent increments and hence is Brownian motion. So, collecting together the terms that have $Q_j(s)$ as a factor, we have

$$-\Sigma_{12}(t,j) + Q_j(t)\Sigma_{22}(t,j)$$

$$= -\int_0^t S(u)/y(u)d\Lambda_j(u) - \int_0^t Q_j(u)/y(u)d\Lambda_\cdot(u) + Q_j(t)\int_0^t 1/y(u)d\Lambda_\cdot(u)$$

using definition (2.1) of $Q_j(u)$ and reordering the terms

$$= Q_j(t)\int_0^t d\Lambda_\cdot(u)/y(u) - \int_0^t Q_j(u)d\Lambda_\cdot(u)/y(u) - \int_0^t dQ_j(u)/y(u)$$

using integration by parts

$$= \int_0^t \left[\int_0^u d\Lambda_\cdot(v)/y(v)\right] dQ_j(u) - \int_0^t dQ_j(u)/y(u) \tag{2.30}$$

At this point it is helpful to consider what $y(u)$ represents. It is a type of survival function where an event is defined as any observed failure time, including all causes of failure *and* censorings and, under the prevailing assumption of independent censoring,

it can be written as $y(u) = \exp(-\Lambda.(u) - \kappa(u))$, where κ is the hazard function for the censoring distribution. Hence

$$\int_0^u d\Lambda.(v)/y(v) = \int_0^u \exp(\Lambda.(u) + \kappa(u)) d\Lambda.(v)$$

$$= \exp(\Lambda.(u) + \kappa(u)) - \int_0^u \exp(\Lambda.(u) + \kappa(u)) d\kappa(v) = 1/y(u) - \int_0^u d\kappa(v)/y(v),$$

so substituting this expression into (2.30) some cancellation occurs and the resulting expression is

$$\int_0^t \left[\int_0^u d\kappa(v)/y(v) \right] dQ_j(u). \tag{2.31}$$

Now if there is no censoring, or equivalently if censoring is just considered as another cause of failure and not considered to be independent, then κ is zero and hence $D_j(t)$ has independent increments. If there is light censoring, or an initial period with no censoring, then we can say that there are approximately independent increments in the sense that (2.31) is approximately zero.

2.6.1 Asymptotic Pointwise Confidence Intervals

Theorem 2.5.3 provides a means to produce pointwise confidence intervals. And if we are only considering one point in time, there is no need to consider the covariation process across time. Standard theory of Z-statistics applies.

So for a fixed t, defining the interval,

$$I(t, \alpha) = (\widehat{Q}_j(t) - z_{\alpha/2}\widehat{\sigma}(t), \widehat{Q}_j(t) + z_{\alpha/2}\widehat{\sigma}(t)),$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution, we obtain that asymptotically,

$$\mathbb{P}(Q_j(t) \in I(t, j)) = \alpha.$$

However, it is possible that these confidence intervals will not lie entirely within the range $[0,1]$ in which Q_j lies. To avoid this problem use of the delta method can be made. The basic idea is that for a sufficiently well-behaved function $g(\cdot)$, and a random sequence $\{X_n\}$ such that $(X_n - x)/s.d.(X_n) \xrightarrow{D} N(0,1)$, for a constant x, it can be shown that

$$\frac{g(X_n) - g(x)}{|g'(x)|s.d.(X_n)} \xrightarrow{D} N(0,1).$$

Now if the transformation, g is chosen to map $[0,1]$ to $\mathbb{R}$ then we can obtain a confidence interval, on the g-scale, and then map it back to $[0,1]$ using g^{-1} and the resulting confidence interval will lie within the range of Q_j.

2.6.2 Simultaneous Confidence Bands

By a simultaneous confidence band on a one-dimensional function $H(t)$ we mean a 2-dimensional region I, such that

$$\mathbb{P}\{(t, H(t)) \in I; \forall t\} = 1 - \alpha/100,$$

for a given percentile α. Given the two-dimensional nature of I, it is clear that there are infinitely many candidates for I and, unlike the 1-dimensional analogy the pointwise confidence interval, there is no well defined notion of choosing the region I with smallest width. For the remainder of this section it will be assumed that the error process has approximately independent increments. With this assumption the literature offers three main choices: the Hall-Wellner Band (Hall and Wellner 1980), the Equal Precision band (Nair 1984), and the Gill bands (Aalen 1976, Gill 1980). An excellent discussion of the derivation of these bands is given in Andersen et al. (1993, pp. 208–213).

The Hall-Wellner band corresponds to the Kolmogorov-Smirnov band in the case of one cause and no censoring—the crude incidence function is the empirical distribution

function. The Equal Precision band has the property that the width of the band, or equivalently the distance of the extrema of the bands from $\widehat{Q}_j(t)$, at any point t remains proportional to the width of the pointwise confidence bands. However both the equal precision and the Hall-Wellner bands require the evaluation of quantiles of rather complicated distributions, which depend on the choice of t_1, t_2, when the region of interest is $\{t : t_1 \leq t \leq t_2\}$; the Hall-Wellner also rely explicitly on the choice of sequence, a_n, although, given that the overwhelmingly popular choice is for $a_n = \sqrt{n}$, thus invoking the strong law of large numbers to satisfy proposition 2.5.2, this is not a great complication. An advantage they offer is that when the assumptions of theorem 2.5.3 break down, which typically occur when there is a cut-off time T before which all individuals will either be censored or have failed, the bands do not explode to infinity, or become incalculable.

The Gill bands are simple to evaluate, do not depend upon the sequence a_n, and have the property that their width remains constant over time. However, for this ease of evaluation the price is that they tend to be larger in size, in some sense, apart from the tail of the distribution, which is a region where there is least information provided by the data and that is commonly not of great practical interest.

Derivations

To give a sketch of the derivations we need to note a few brief facts about Brownian Motion and the related process a Brownian Bridge. A Brownian Motion $B(t)$ is defined to be a Gaussian, mean zero random process, with $B(0) = 0$ almost surely, and with covariance process $\text{Cov}(B(t), B(s)) = s \wedge t$. The related process, a Brownian Bridge is $W(t) = B(t) - tB(1)$, is also a mean zero Gaussian process, but the covariance process is $s(1-t)$ for $0 < s \leq t < 1$, it can be seen that almost surely $W(1) = 0$, hence

the name. Now consider the processes defined as $B(t)/(1+t)$ and $W(t/(1+t))$ for $0 < t < \infty$. It is clear that they are both zero-mean Gaussian processes, have value zero at $t = 0$, and both their covariance processes for time points s, t are $(s \wedge t)/(1+s)(1+t)$, hence they have the same distribution.

To obtain a general class of confidence bands, we choose an arbitrary continuous, non-negative function q, and note that for a mean-zero Gaussian process $U(t)$, which starts at zero almost surely, with variance function $\sigma^2(t)$, and independent increments (which corresponds to $B(\sigma^2(t))$), we can observe that

$$\begin{aligned}\frac{U(\sigma^2(t))}{1+\sigma^2(t)} q\left(\frac{\sigma^2(t)}{1+\sigma^2(t)}\right) &\sim W\left(\frac{\sigma^2(t)}{1+\sigma^2(t)}\right) q\left(\frac{\sigma^2(t)}{1+\sigma^2(t)}\right) \\ &= W(x)q(x), \quad x \in [0,1]\end{aligned}$$

the point being that the right hand side follows a parameter-free distribution, and does not depend upon the variation process, $\sigma^2(\cdot)$.

It follows that

$$\sup_{t\in[t_1,t_2]} \left| \frac{a_n(\widehat{Q}_j(t) - Q_j(t))}{1+a_n^2\widehat{\sigma}^2(t)} q\left(\frac{a_n^2\widehat{\sigma}^2(t)}{1+a_n^2\widehat{\sigma}^2(t)}\right)\right| \xrightarrow{D} \sup_{x\in[c_1,c_2]} |q(x)W(x)|,$$

where $c_i = \sigma^2(t_i)/(1+\sigma^2(t_i))$, which can be estimated with $\widehat{c}_i = a_n^2\widehat{\sigma}^2(t)/(1+a_n^2\widehat{\sigma}^2(t))$. This can be inverted to give a $100(1-\alpha)\%$ band on $[t_1, t_2]$,

$$\widehat{Q}_j(t) \pm a_n^{-1} K_{q,\alpha}(\widehat{c}_1, \widehat{c}_2)(1+a_n^2\widehat{\sigma}^2(t))/q\left(\frac{a_n^2\widehat{\sigma}^2(t)}{1+a_n^2\widehat{\sigma}^2(t)}\right),$$

where $K_{q,\alpha}(c_1, c_2)$ is the upper α quantile of the distribution of

$$\sup_{x\in[c_1,c_2]} |q(x)W(x)|.$$

Equal Precision Bands

The equal precision bands are obtained by using $q(x) = 1/\sqrt{x(1-x)}$. In this case the a_n terms disappear and the band simplifies to

$$\widehat{Q}_j(t) \pm d_\alpha(\widehat{c}_1, \widehat{c}_2)\widehat{\sigma}(t),$$

where a $d_\alpha(c_1, c_2)$ can be found using the formula from Miller and Siegmund (1982)

$$\mathbb{P}\left\{ \sup_{x\in[c_1,c_2]} \left| W(x)/\sqrt{x(1-x)} \right| \geq d \right\}$$
$$= \frac{4\phi(d)}{d} + \phi(d)(d - 1/d)\log\left(\frac{c_2(1-c_1)}{c_1(1-c_2)}\right) + o(\phi(d)/d),$$

where ϕ is the standard normal density function.

Hall-Wellner Bands

The Hall-Wellner bands are defined by choosing the function $q(x) = 1$. Here the sequence a_n does enter into the calculation, but the quantile $K_{q,\alpha}(c_1, c_2)$ is possibly easier to calculate as it is the upper α quantile of the distribution of

$$\sup_{x\in[c_1,c_2]} |W(x)|.$$

Software to evaluate such quantiles is available at

`http://www.nrcan.gc.ca/gsc/mrd/sdalweb/wiener/index.html` and a summary of the mathematics behind the software is in Chung (1987). A conservative estimate can be obtained by taking $[c_1, c_2] = [0, 1]$; the appropriate quantiles are tabulated below.

$K_{1,\alpha}(0,1)$	1.2238	1.3581	1.4802	1.6276	1.9495	2.2246
α	0.1	0.05	0.025	0.01	0.001	0.0001

Table 2.1: Upper quantiles of the supremum of the modulus of a Brownian Bridge on the unit interval

Gill Bands

These bands are only applicable when the region of interest is $0 < t < t_2$. To derive them, note that the choice of a_n is fairly arbitrary, indeed if we have a sequence which satisfies the conditions for theorem 2.5.3, and another (possibly random) sequence $b_n \to b < \infty$, then we could apply theorem 2.5.3 using $a_n b_n$ as the sequence. If we define $\gamma_n = a_n^2 b_n^2 \widehat{\sigma}^2(t_2)$, then the Hall-Wellner band is

$$\widehat{Q}_j(t) \pm \gamma_n^{-1/2} e_\alpha(0, \gamma_n/(1+\gamma_n)) \widehat{\sigma}(t_2)(1 + \gamma_n \widehat{\sigma}^2(t)/\widehat{\sigma}^2(t_2)), \tag{2.32}$$

where $e_\alpha(c_1, c_2)$ is the upper α quantile of the distribution of $\sup_{x \in [c_1, c_2]} |W(x)|$.

Now we can *define* the sequence b_n to give a constant $\gamma_n = \gamma$, and consider what happens as $\gamma \to 0$. Now consider the distribution $\gamma^{-1/2} W(\gamma t/(1+\gamma))$, this is a zero-mean Gaussian variable and its covariance process is

$$\frac{s}{1+\gamma}\left(1 - \frac{\gamma t}{1+\gamma}\right) \to s, \quad 0 < s \leq t < 1, \quad \text{as } \gamma \to 0.$$

So $\gamma^{-1/2} W(\gamma t/(1+\gamma))$ converges in distribution to a standard Brownian motion. Hence (2.32) becomes,

$$\widehat{Q}_j(t) \pm u_\alpha \widehat{\sigma}(t_2),$$

where u_α is the upper α quantile of the distribution of

$$\sup_{x \in [0,1]} |B(x)|,$$

which only depends upon α and not c_2. The values u_α can be calculated using the result,

$$\mathbb{P}\left\{ \sup_{x \in [0,1]} |B(x)| > u \right\} = 1 - \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\{-\pi^2 (2k+1)^2 / 8u^2\},$$

which is derived in Billingsley (1999, chapter 2, section 9, pp. 94–101). Some values of u_α are given in table 2.2.

u_α	1.960	2.241	2.498	2.807	3.481	4.056
α	0.1	0.05	0.025	0.01	0.001	0.0001

Table 2.2: Upper quantiles of the supremum of the modulus of Brownian motion on the unit interval

2.7 Example

To finish this chapter, I will illustrate the computation of the various confidence intervals and bands. The data are the survival times (in months) of 121 breast cancer patients from the clinical records of one hospital over the period 1929 to 1938. The causes of death are 'Cancer' (78 patients) and 'Other' (18 patients), there is also mild censoring (25 patients) of which the earliest is at 111 months. The data are in Boag (1949) and are included in appendix A.

To calculate the estimates of the Crude Incidence function, Q_j, and the associated variation process, σ_j^2, the S-plus code listed in appendix B was used. Code already exists within the R software (Ihaka and Gentleman 1996): the `cmprsk` package, which calculates and plots the crude incidence function. However, this code is more geared towards hypothesis testing between groups as outlined in Gray (1988), and it does not produce confidence intervals or bands. To use this we need to estimate $[c_1, c_2]$, where $\widehat{c}_i = n\widehat{\sigma}^2(t_i)/(1+n\widehat{\sigma}^2(t_i))$, which in this case were $[0.008063, 0.4093]$ where the t_i were chosen to coincide with the first and last failure times. Using these parameters the 95% critical values of the distributions used for the Equal Precision, Hall-Wellner, and Gill bands on this interval were calculated to be 3.058, 1.206, 2.241, respectively; for the

pointwise 95% confidence interval 1.96 was used. The results are plotted below.

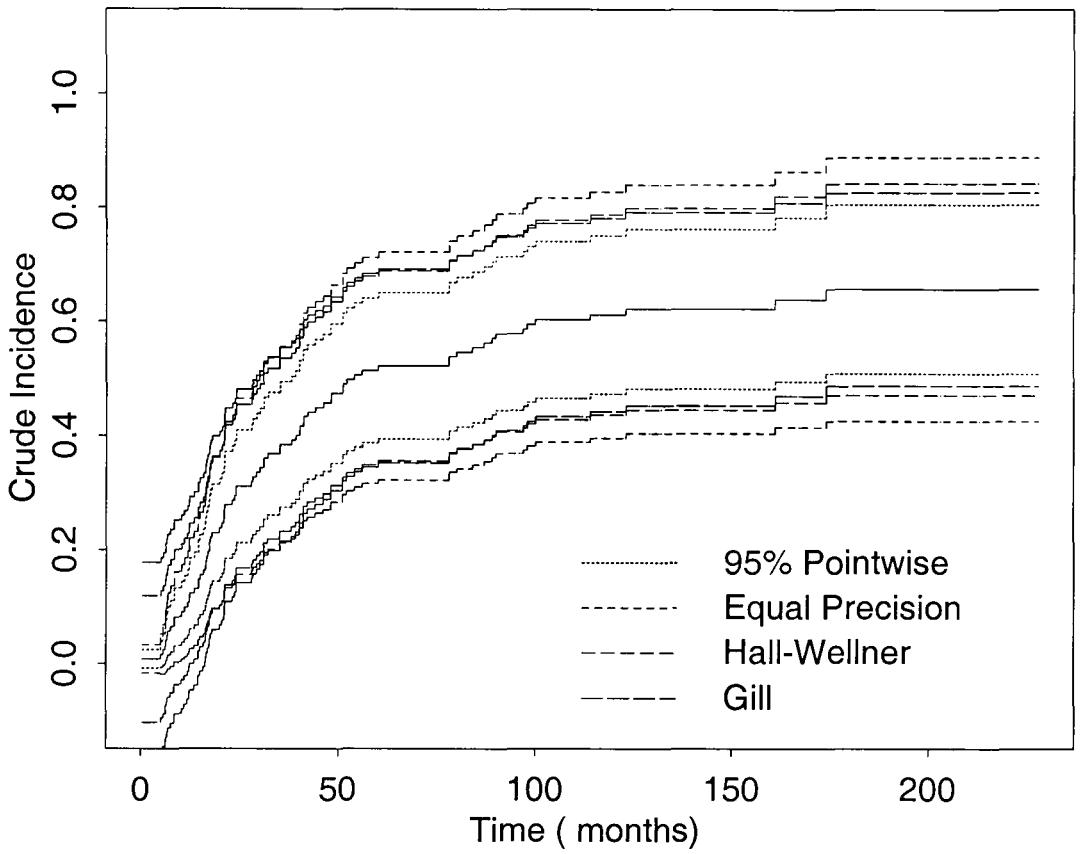


Figure 2.1: Crude Incidence function for Cancer with confidence bands

As would be anticipated, the 95% pointwise confidence interval lies closer to the point estimate than all the simultaneous confidence bands. At the start of the time interval the Equal Precision is the narrowest band, followed by the Hall-Wellner band, and the Gill band is the widest; by the end of the time interval this ordering has been reversed. The Equal Precision and Hall-Wellner bands intersect at about 20 months, the Gill band intersects with the Equal Precision at about 35 months, and the Gill band intersects with the Hall-Wellner band much later at 80 months. However, even at the very end of the time interval, the Gill band only offers an improvement, in terms of

width, of $2 \times (0.1854 - 0.1696)100\% = 3.1\%$, suggesting that the Hall-Wellner bands are a sensible compromise in this case.

The equivalent crude incidence function and associated confidence intervals/bands is shown for 'Other' causes in figure 2.2. Bootstrapping was performed, with 1000 replicates, and 96.5 % coverage was obtained for the Hall-Wellner bands, 48.9 % coverage for the Equal Precision bands and 97.1 % coverage for the Gill bands. The poor performance of the Equal precision bands, in this case, is due to the time interval starting as early as possible, and most of the occurrences of the bands being breached occur near the start. This is a consequence of the fact that the Equal Precision bands are of the form $\widehat{Q} + K \cdot \widehat{\sigma}$, whereas the other two bands include an additive constant thus avoiding the problem of a very small $\widehat{\sigma}$. If the time interval is slightly changed from $[0.3, 228]$, to $[17.3, 228]$, the coverage of the Equal Precision bands becomes 95.6 %.

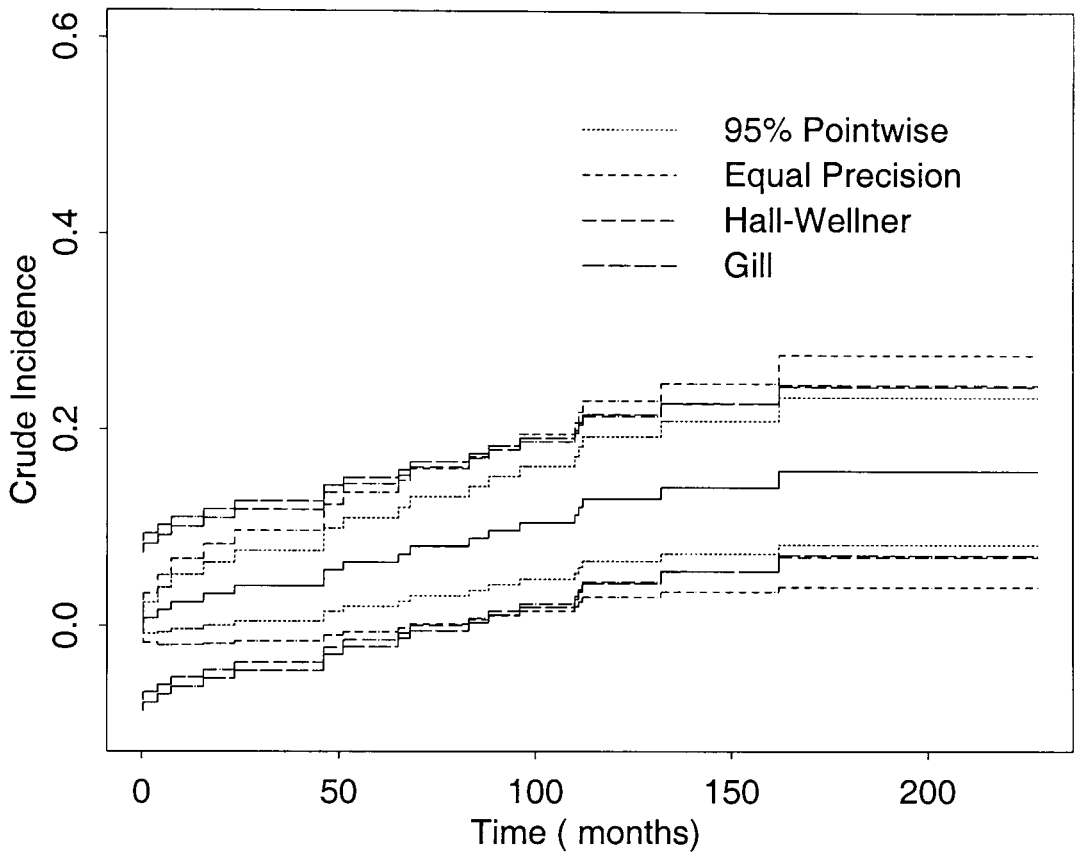


Figure 2.2: Crude Incidence function for Other Causes with confidence bands

2.8 Summary

In this chapter we have presented together the mathematical properties of the conventional estimator of the crude incidence function. With these we have developed methods for calculating confidence intervals and bands for the estimates and have illustrated their important potential for use in exploratory analysis. The limited software available does not calculate any such confidence bands and rather than giving a visual tool which communicates the uncertainty of the estimates, concentrates instead on hypothesis testing between sub-groups.

The next chapter moves on from the exploratory phase of the statistical process and considers what can be inferred about the dependence between the latent failure times and what sort of comparisons can be made between sub-groups of individuals.

Chapter 3

Improved bounds for the joint survival in the case of a two-armed trial

3.1 Introduction

A common way to represent competing risks data is through the use of latent failure times. This assumes that, when dealing with k causes, there exists a vector, $(T_1, \ldots, T_k)$, of random failure times, each associated with a particular cause, but rather than observing all of these values we only observe one them, the earliest, and have only the knowledge that the other times must be larger than the observed time. A criticism of this, which is explored in Prentice and Kalbfleisch (1978), is that Nature is not some idiot scientist who can be criticised on the grounds:

> To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to

say what the experiment died of.

—Ronald Fisher, Indian Statistical Congress, Sankhya, ca 1938.

Rather, these latent times do not exist, and to perform counter-factual inference is a dangerous thing to do. Nevertheless there are questions typically asked about such latent times: what is the dependency between the latent times; what happens if we remove a particular cause of failure; what are the effects of treatments or covariates on a particular latent failure time. These are perfectly valid questions, and they rely on the latent failure times to be considered. If data alone cannot provide a specific answer we must consider what range, or set, of answers can be inferred, and how extra assumptions will affect the problem.

In this chapter I consider the marginal distribution of the latent failure times. Consider two scenarios: conditionally on the earliest latent time, all the remaining latent times are immediately afterwards; conditionally on the earliest latent time, all the remaining latent times are at infinity. The data cannot distinguish between these scenarios since we only observe the earliest time, but the dependence structures are very different and clearly the marginal survival in the first scenario will decay at a faster rate than the second scenario.

These two extremes are considered formally in Peterson (1976) and they produce a set of bounds on the marginal survival which are known as Peterson's bounds. They will be presented formally here also. However they are only applicable to a homogeneous data set. I will consider the simplest increment into heterogeneity: the two-armed trial. For this chapter we are making the assumption that the effect of a treatment can be represented by a time transformation, whereby to calculate the joint survival function function of the latent failure times at a point $\mathbf{t}$ in the experimental arm, say, we could perform a transformation $\mathbf{t} \mapsto \phi(\mathbf{t})$, and evaluate the joint survival of the control arm at

this new point. Now we do not know the joint survival in the control or the experimental arms, but if we did know this time-transformation, ϕ, then Peterson's bounds can be improved. It is a strong assumption in knowing ϕ and I will consider this question more fully in chapter 4. However, based on the assumption that we know ϕ, this chapter explores what extra information we have on the joint survival.

3.2 Definition of the covariate-time transformation

The model assumptions shortly to be defined are motivated as a generalisation of accelerated failure time models. The accelerated failure time model assumes the effect of covariates is to speed up, or slow down, time by a factor determined by the covariates, Z. Formally, in a univariate example

$$S[t|Z] = S[f(Z)t|z_0],$$

where z_0 is a reference value of Z such that $f(z_0) = 1$. Hence if $f(z_1) > 1$ then the probability of a failure time larger than t, conditional on $Z = z_1$ is smaller than the probability, conditional on $Z = z_0$. It is a useful, practical alternative to the proportional hazards assumption that is both parsimonious and easy to interpret.

Another viewpoint is that the transformed variable $f(Z)T$ has a distribution that does not depend upon the value of Z, and is identical to the distribution conditional on $Z = z_0$, since

$$\mathbb{P}[f(Z)T > t|Z] = \mathbb{P}[T > t/f(Z)|Z] = \mathbb{P}[T > f(Z)\{t/f(Z)\}|z_0] = \mathbb{P}[T > t|z_0]. \tag{3.1}$$

A simple generalisation is to replace $f(z)t$ with an arbitrary function $\phi(t, z)$ that is zero at $t = 0$, is increasing in t, and satisfies $\phi(t, z_0) = t$ for some reference value, z_0.

Extending this into the latent failure time framework, the covariate-time transformation is defined to be the mapping,

$$\boldsymbol{\phi} : [0,\infty)^k \times \Omega_Z \mapsto [0,\infty)^k,$$

where Ω_Z is the sample space of the variable Z, such that,

$$S[\mathbf{t}|Z=z] = S[\boldsymbol{\phi}(\mathbf{t},z)|Z=z_0], \tag{3.2}$$

$$\boldsymbol{\phi}(\mathbf{0},z) = \mathbf{0}, \tag{3.3}$$

$$\boldsymbol{\phi}(t_1,\ldots,u,\ldots,t_k,z) \geq \boldsymbol{\phi}(t_1,\ldots,v,\ldots,t_k,z), \text{ for } u \geq v. \tag{3.4}$$

This is a completely general framework, and contains any joint distribution of $(T_1,\ldots,T_k,Z)$, since the role of $\boldsymbol{\phi}$ is to map between the contours of the survival function for the different values of Z. As such it is too general. The rest of the chapter will consider the special case where the ith element of $\boldsymbol{\phi}(\mathbf{t},z)$ can be simplified to $\phi_i(t_i,z)$, a function on (t_i,z) rather than $(t_1,\ldots,t_k,z)$; henceforth referred to as the *rectangular* assumption. It is also assumed that the ϕ_is are continuous in t, and hence with (3.4) implies that $\phi_i(\cdot,z)$ has an inverse, $\psi_i(\cdot,z)$. By a similar argument to (3.1) it can be shown that $(\phi_1(T_1,Z),\ldots,\phi_k(T_k,Z))$, conditional on Z, has the same distribution as $(T_1,\ldots,T_k)$ conditional on $Z=z_0$:

$$\begin{aligned}
& \mathbb{P}[\phi_1(T_1,Z) > t_1,\ldots,\phi_k(T_k,Z) > t_k|Z] \\
= \; & \mathbb{P}[T_1 > \psi_1(t_1,Z),\ldots,T_k > \psi_k(t_k,Z)|Z] \\
= \; & \mathbb{P}[T_1 > \phi_1\{\psi_1(t,Z),Z\},\ldots,T_k > \phi_k\{\psi_k(t,Z),Z\}|Z=z_0] \\
= \; & \mathbb{P}[T_1 > t_1,\ldots,T_k > t_k|Z=z_0].
\end{aligned}$$

So if a value of Z is observed such that $\phi_i(t_i,z) > t_i$ then this would be interpreted as accelerating the latent failure time T_i. Specific examples are shown in

section 4.2, including both the accelerated failure time model and the proportional hazards model. Chapter 4 considers the estimation of the covariate-time transformation, but this chapter will assume it is known *a priori* and will consider what can be inferred about the joint distribution.

3.3 A Geometric Introduction

First we will introduce Peterson's bounds in the case of two causes of failure. Formally, assume there is a pair of variables $(T_1, T_2) \in \mathbb{R}^2_+$ and we observe $T^{\min} = \min(T_1, T_2)$ and the cause of failure C. Now if we wish to calculate $S(t_1, t_2) = \mathbb{P}(T_1 > t_1 \cap T_2 > t_2)$, referring to figure 3.1, this is equivalent to integrating over the infinite rectangle with a 'lower left' vertex at point $X = (t_1, t_2)$ with measure corresponding to the joint density of (T_1, T_2). However, given the observed data we can only estimate the crude incidence function for each cause, $\mathbb{P}(T^{\min} < t, C = i) = Q_i(t)$. For the purposes of this chapter we will work with a closely related function: the cause-specific survival function,

$$F_j(t) = \mathbb{P}(T^{\min} > t, C = i) = Q_i(\infty) - Q_i(t),$$

the quantity $Q_i(\infty) = F_i(0) = \mathbb{P}(C = i)$. Examining figure 3.1 it is clear that, for example, $F_1(t_1)$ corresponds to integrating over the infinite triangle with sides formed by the diagonal, and the vertical line withe lower end-point $B = (t_1, t_1)$, similarly $F_2(t_1)$ is the integral over the lower triangle at B. Clearly, the rectangle we wish to integrate lies within the union of the upper triangle at B and the lower triangle at A, which therefore provides an upper bound, $F_1(t_1) + F_2(t_2)$. Whereas the union of the upper and lower triangles at B lies within the rectangle thus providing a lower bound, $F_1(t_1) + F_2(t_1)$.

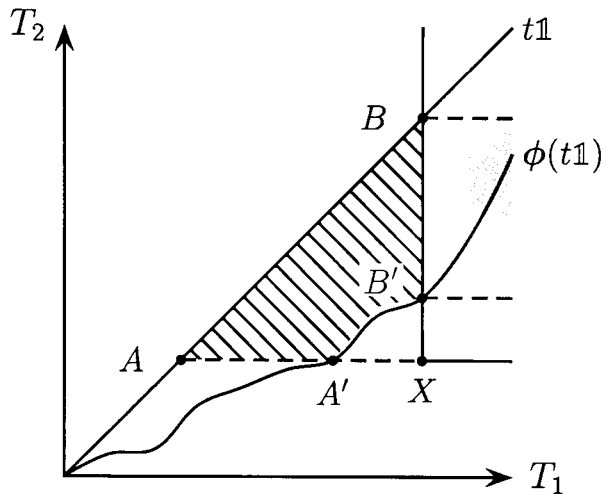


Figure 3.1: Illustration of the 2-d case

This is formalised, in the finite-dimensional case, as

$$\sum_i F_i\left(\max(\mathbf{t})\right) \leq S(\mathbf{t}) \leq \sum_i F_i(t_i).$$

Now consider the implications of (3.2) along with the rectangular assumption, in the case where $Z \in \{0, 1\}$ represents the two arms of a trial, with $z_0 = 0$. To simplify notation, $\phi(\mathbf{t}, Z = 1)$ is condensed into $\phi(\mathbf{t})$. In figure 3.1 the curved line represents the image of the mapping of the diagonal under the covariate-time transformation : $(t, t) \mapsto (\phi_1[t, 1], \phi_2[t, 1])$. Now under (3.2) the joint survival function at any point on this line, conditional on $Z = 0$, is equal to the survival function, conditional on $Z = 1$, for a specific point *on the diagonal*. This is useful because the survival function of a point on the diagonal, is simply the univariate survival function of $T^{\min}$, and this is clearly estimable from the data. Under the rectangular assumption, given a point $(t_1, \ldots, t_k)$ on this image, the relevant point on the diagonal is $(\psi_i(t_i), \ldots, \psi_i(t_i))$ (where the value of i is irrelevant since $\psi_i(t_i) = \psi_j(t_j)$, given that $(t_1, \ldots, t_k)$ lies on the curved line).

Further, the rectangular assumption also means that, for example, defining the 'lower pseudo-triangle' $\mathcal{A}$ as the region to the right/below the curved line, and above

the horizontal line through X, then $\mathbb{P}\{(T_1, T_2) \in \mathcal{A} | Z = 0\} = F_2\{u(\mathcal{A}) | Z = 1\}$, where $u(\mathcal{A}) = \psi_2(t_2)$ in this case since the curved line lies below the diagonal. These relationships will be formally derived later, but the important thing to note is: first, that the rectangle with 'lower-left vertex' at the point X is contained within the region defined as the union of the lower pseudo-triangle at A' and the upper pseudo-triangle at B'; second, the same rectangle, in turn, contains the union of the upper and lower pseudo-triangles at B'. The probabilities of these regions, conditional on $Z = 0$, can be evaluated and they will provide tighter bounds for the joint survival at X than the Peterson bounds.

3.4 Extension to finite dimensions

Next we will formalise these ideas in a generalised context of there being $p > 2$ causes of failure. First we will consider the p–dimensional version of Peterson's bounds.

Theorem 3.4.1 (Finite-dimension worst-case bounds).

$$\sum_i^p F_i(u_i) \geq S(u_1, u_2, \ldots, u_p) \geq \sum_i^p F_i(\max\{u_k\})$$

Proof. It is clear that $\bigcap_i^p \{T_i > u_i\} \supset \bigcap_i^p \{T_i > \max\{u_k\}\}$. The next step is to partition the smaller subset by the events $\{T_j = \min\{T_k\}\}$, but these partitions can be simplified to $\bigcap_i^p \{T_i > \max\{u_k\}\} \cap \{T_j = \min\{T_k\}\} = \{T_j > \max\{u_k\}\} \cap \{T_j = \min\{T_k\}\}$. So taking the probabilities of these events we obtain the lower bound,

$$\begin{aligned} \mathbb{P}(\bigcap_i^p \{T_i > u_i\}) &\geq \sum_i^p \mathbb{P}(\{T_i > \max\{u_k\}\} \cap \{T_i = \min\{T_k\}\}) \\ &= \sum_i^p F_i(\max\{u_k\}). \end{aligned}$$

To obtain the upper bound the event of interest is partitioned, again, by the

events $\{T_j = \min\{T_k\}\}$, so that

$$\bigcap_i^p \{T_i > u_i\} = \bigcup_j^p \left\{ \bigcap_i^p \{T_i > u_i\} \cap \{T_j = \min\{T_k\}\} \right\},$$

but clearly $\bigcap_i^p \{T_i > u_i\} \cap \{T_j = \min\{T_k\}\} \subset \{T_j > u_j\} \cap \{T_j = \min\{T_k\}\}$. So taking the probability of these events, the upper bound is obtained,

$$\mathbb{P}(\bigcap_i^p \{T_i > u_i\}) \leq \sum_i^p \mathbb{P}(\{T_i > u_i\} \cap \{T_i = \min\{T_k\}\} = \sum_i^p F_i(u_i).$$

□

The next stage is to generalise the effect of a binary covariate from two causes to $p > 2$ causes. In the ensuing derivations the following lemma will be used

Lemma 3.4.1. *Psuedo-triangular regions*

$$\begin{aligned} F_j^1(t) &= \mathbb{P}(T_j > t \cap T_j = \min\{T_k\} | Z = 1) \\ &= \mathbb{P}(\psi_j(T_j) > t \cap \psi_j(T_j) = \min\{\psi_k(T_k)\} | Z = 0) \end{aligned} \tag{3.5}$$

Proof. By the definition of the covariate-time transformation (using the original, expanded notation for ϕ and ψ), it is assumed that

$$\begin{aligned} \mathbb{P}(\mathbf{T} \in \mathcal{A} | Z) &= \mathbb{P}(\boldsymbol{\phi}\{\mathbf{T}, Z\} \in \boldsymbol{\phi}\{\mathcal{A}, Z\} | Z) \\ &= \mathbb{P}(\mathbf{T} \in \boldsymbol{\phi}\{\mathcal{A}, Z\} | Z = z_0), \end{aligned}$$

where $\mathcal{A}$ is any measurable region in the sample space of the latent failure times. Now consider $\mathcal{A} = \{\mathbf{t} : t_j > t, t_i \geq t_j, i \neq j\}$, we wish to find $\boldsymbol{\phi}\{\mathcal{A}, Z\} = \{\mathbf{t} : t_i = \phi_i(u_i, Z), \mathbf{u} \in \mathcal{A}\}$. By the rectangular assumptions the ϕ_is can be inverted, hence $\boldsymbol{\phi}\{\mathcal{A}, Z\} = \{\mathbf{t} : \psi_j(t_j, Z) > t, \psi_i(t_i, Z) > \psi_j(t_j, Z), i \neq j\}$. Hence the event $\{\mathbf{T} \in \boldsymbol{\phi}[\mathcal{A}, Z]\}$ can be written as $\{\psi_j(T_j) > t \cap \psi_j(T_j) = \min[\psi_k(T_k)]\}$, reverting back to the condensed notation for the two-armed trial. □

Now it is possible to generalise the two dimensional case.

Theorem 3.4.2 (Alternative bounds).

$$\sum_{j=1}^{p} F_j^1(\psi_j(u_j)) \geq S(u_1, u_2, \ldots, u_p | Z = 0) \geq \sum_{j=1}^{p} F_j^1(\max\{\psi_k(u_k)\}).$$

Proof. For the lower bound, observe that, by definition, $\max\{\psi_k(u_k)\} \geq \psi_i(u_i)$ and hence since ϕ_i is non-decreasing, $\{T_i > u_i\} \supseteq \{T_i > \phi_i[\max\{\psi_k(u_k)\}]\}$. So

$$\mathbb{P}\left(\bigcap_i^p \{T_i > u_i\}\right) \geq \mathbb{P}\left(\bigcap_i^p \{T_i > \phi_i[\max\{\psi_k(u_k)\}]\}\right),$$

but if the event on the the right hand side is partitioned by the sets $\{\psi_j(T_j) = \min\{\psi_k(T_k)\}\} = \mathcal{C}_j$, some simplifications occur,

$$\begin{aligned}
&\bigcup_j^p \left\{\bigcap_i^p \{T_i > \phi_i[\max\{\psi_k(u_k)\}]\} \cap \mathcal{C}_j\right\} \\
&= \bigcup_j^p \left\{[\psi_j(T_j) > \max\{\psi_k(u_k)\}] \cap \mathcal{C}_j\right\},
\end{aligned}$$

hence taking the probability of these events, and using equation 3.5 the lower bound is obtained,

$$S(u_1, u_2, \ldots, u_p | Z = 0) \geq \sum_{j=1}^{p} F_j^1(\max\{\psi_k(u_k)\}).$$

For the upper bound observe that $\bigcap_i^p [T_i > u_i] \cap \mathcal{C}_j \subset [T_j > u_j] \cap \mathcal{C}_j$, so taking the union over j the following is obtained,

$$\begin{aligned}
\bigcap_i^p [T_i > u_i] &= \bigcup_j^p \left\{\bigcap_i^p [T_i > u_i] \cap \mathcal{C}_j\right\} \\
&\subset \bigcup_j^p \left\{[T_j > u_j] \cap \mathcal{C}_j\right\} \\
&= \bigcup_j^p \left\{[\psi_j(T_j) > \psi_j(u_j)] \cap \mathcal{C}_j\right\}
\end{aligned}$$

Hence evaluating the probability of these events, conditional on $Z = 0$, and using equation 3.5, the upper bound is obtained,

$$S(u_1, u_2, \ldots, u_p | Z = 0) \leq \sum_j^p F_j^1(\psi_j(u_j)).$$

□

3.4.1 Marginals

As a corollary of theorems 3.4.1 and 3.4.2 we can evaluate the marginal survival, $S_j(t)$, for latent time T_j. This corresponds to evaluating the joint survival function for at a point $(0, \ldots, 0, t_j, 0, \ldots, 0)$. Hence there are two possible bounds.

$$F_j^0(t) + \sum_{i \neq j} Q_j^0(0) \geq S_j(t|Z = 0) \geq \sum_i F_i^0(t),$$

which comes from theorem 3.4.1, and

$$F_j^1(\psi_j(t)) + \sum_{i \neq j} F_i^1(0) \geq S_j(t|Z = 0) \geq \sum_i F_i^1(\psi_j(t)).$$

3.5 Which bounds are tighter?

It is convenient to be able to decide, in advance, which of the two sets of bounds is tighter. Intuitively if we have a point which is 'nearer', in some sense, to the line $(\phi_1(t), \ldots, \phi_k(t))$ than to $(t, \ldots, t)$ then the alternative bounds will be tighter. However, short of simply evaluating both sets of bounds, it is hard to say what being nearer means in this case.

A limited result is presented below in which it is assumed that it is possible to permute the labelling of the events is such that $\psi_1(t) \leq \psi_2(t) \leq \ldots \leq \psi_p(t)$, for all t, or equivalently $\phi_1(t) \geq \phi_2(t) \geq \ldots \geq \phi_p(t)$. This is not true in general for all t. However,

if this is the case then the three regions, $A = \{\psi_1(u_1) > \psi_2(u_2) > \ldots > \psi_p(u_p)\}$, $C = \{u_1 < u_2 < \ldots < u_p\}$, and $B = \mathbb{R}_+^p/A/C$, tell us where either set of bounds may be optimal. In region A, theorem (3.4.2) provides the tighter bounds; in region C, theorem (3.4.1) provides the tighter bounds; in region B it is not possible to tell which set of bounds will be tighter without their evaluation.

3.5.1 Region A

For the first case, in region A, the assumption that $\psi_i(u) \geq \psi_1(u)$ implies that $u_1 \geq \phi_i(\psi_1(u_1))$, hence

$$\bigcap_i^p \{T_i > u_1\} \subset \bigcap_i^p \{T_i > \phi_i(\psi_1(u_1))\}. \tag{3.6}$$

However, the right hand side, $\bigcap_i^p \{\psi_i(T_i) > \psi_1(u_1)\}$, equals $\bigcap_i^p \{\psi_i(T_i) > \max(\psi_k(u_k))\}$, if and only if $\psi_1(u_1) = \max(\psi_k(u_k))$, which is satisfied by the definition of region A. However theorem (3.4.2) shows that

$$\mathbb{P}\left(\bigcap_i^p \{\psi_i(T_i) > \max(\psi_i(u_i))|Z = 0\right) = \sum_i^p F_i^1(\max(\psi_k(u_k))).$$

Also, given that $\psi_1(u_1) \geq \psi_i(u_i)$, applying the function ϕ_1 gives $u_1 \geq \phi_1(\psi_i(u_i)) \geq \phi_1(\psi_1(u_i)) = u_i$. So,we have that in region A, $u_1 = \max(u_i)$ and theorem (3.4.1) shows that $\mathbb{P}(\bigcap_i^p \{T_i > u_1 = \max(u_i)\}) = \sum_i^p F_i^0(u_1)$. So we can infer that

$$\begin{aligned} \sum_i^p F_i^0(\max(u_k)) &\leq \sum_i^p F_i^1(\max(\psi_k(u_k))) \\ &\leq S(u_1, u_2, \ldots, u_p|Z = 0), \end{aligned}$$

hence theorem (3.4.2) provides the tighter lower bound in region A. In fact the proof shows that all we need is that $\psi_i(u_i) \leq \psi_k(u_k), \forall k$, which is a subset of region A. The upper bounds are where we need the stricter conditions of region A.

To obtain the inequality between the two upper bounds we need to show that, in Region A,

$$\bigcup_j^p \left\{ [T_j > u_j] \cap [\psi_j(T_j) = \min\{\psi_k(T_k)\}] \right\} \subset \bigcup_j^p \left\{ [T_j > u_j] \cap [T_j = \min\{T_k\}] \right\}. \tag{3.7}$$

To do this partition the left hand side by the events $[T_i = \min\{T_k\}]$, and observe that

$$[\psi_j(T_j) = \min\{\psi_k(T_k)\}] \cap [T_i = \min\{T_k\}] = \emptyset, \tag{3.8}$$

for $j > i$, because, taking the event on the left to be true,

$$T_j < \phi_j(\psi_i(T_i)) < \phi_j(\psi_j(T_i)) = T_i.$$

Hence we have that

$$\bigcup_j^p \left\{ [T_j > u_j] \cap [\psi_j(T_j) = \min\{\psi_k(T_k)\}] \right\} =$$

$$\bigcup_j \bigcup_{i \geq j} \left\{ [T_j > u_j] \cap [\psi_j(T_j) = \min\{\psi_k(T_k)\}] \cap [T_i = \min\{T_k\}] \right\}.$$

Given that $[\psi_j(T_j) = \min\{\psi_k(T_k)\}]$ and that $\psi_j(u_j) \geq \psi_i(u_i)$ for $j \leq i$, we can write $\psi_i(T_i) \geq \psi_j(T_j) \geq \psi_j(u_j) \geq \psi_i(u_i)$, hence it is implied that

$$[T_j > u_j] \cap [\psi_j(T_j) = \min\{\psi_k(T_k)\}]$$

$$\subset [T_i > u_i] \cap [\psi_j(T_j) = \min\{\psi_k(T_k)], \ i \geq j.$$

Hence

$$\bigcup_j^p \left\{ [T_j > u_j] \cap [\psi_j(T_j) = \min\{\psi_k(T_k)\}] \right\}$$

$$\subset \bigcup_j \bigcup_{i \geq j} \left\{ [T_i > u_i] \cap [\psi_j(T_j) = \min\{\psi_k(T_k)\}] \cap [T_i = \min\{T_k\}] \right\}.$$

However, due to equation (3.8), we can take the union over all i, change the order of the union-operators, and observe that

$$\bigcup_i \bigcup_j \left\{ [T_i > u_i] \cap [T_i = \min\{T_k\}] \cap [\psi_j(T_j) = \min\{\psi_k(T_k)\}] \right\} = \bigcup_i \left\{ [T_i > u_i] \cap [T_i = \min\{T_k\}] \right\}.$$

So, trivially changing the (now) dummy variable i to j, equation (3.7) is proven. □

In the case of the marginal distributions, the alternative bounds provide tighter bounds for latent time 1.

3.5.2 Region C

The proofs are very similar to region A. For the lower bound, observe that $\psi_i(u) \geq \psi_p(u)$, so $T_i > \phi_i(\psi_p(u_p)) > \phi_i(\psi_i(u_p)) = u_p$, hence we can reverse equation (3.6) to get

$$\bigcap_i [\psi_i(T_i) > \psi_p(u_p)] \subset \bigcap_i [T_i > u_p].$$

In region A it was necessary to show that $\psi_i(u_1) = \max\{\psi_k(u_k)\} \Rightarrow u_1 = \max\{u_k\}$, whereas in region C it is known that $u_p = \max\{u_k\}$, but given that $\psi_p(t) > \psi_i(t)$ it is clear that $\psi_p(u_p) > \psi_i(u_p) > \psi_i(u_i)$, and hence $\psi_p(u_p) = \max\{\psi_k(u_k)\}$. Given these two conditions, the arguments for region A can be followed to convert the probabilities of the relevant events into the sums of crude incidence functions, and the ordering of the lower bounds is reversed.

For the upper bounds, equation (3.7) needs to be reversed to

$$\bigcup_i \left\{ [T_i > u_i] \cap [T_i = \min\{T_k\}] \right\} \subset \bigcup_i \left\{ [T_i > u_i] \cap [\psi_i(T_i) = \min\{\psi_k(T_k)\}] \right\}.$$

To do this we partition the left hand side by $[\psi_j(T_j) = \min\{\psi_k(T_k)\}]$, and apply

equation (3.8) to equate this to

$$\bigcup_j \bigcup_{i \geq j} \left\{ [T_i > u_i] \cap [T_i = \min\{T_k\}] \cap [\psi_j(T_j) = \min\{\psi_k(T_k)\}] \right\}.$$

Now by the definition of region C it is observed that if $T_i = \min\{T_k\}$ and $T_i > u_i$ then for $j \leq i$, $T_j \geq T_i \geq u_i \geq u_j$ hence

$$[T_i > u_i] \cap [T_i = \min\{T_k\}] \subset [T_j > u_j] \cap [T_i = \min\{T_k\}],\ j \leq i.$$

From here the argument is identical to region A, except the roles of i and j are reversed.

Here, in the case of the marginals the conventional Peterson bounds provide tighter bounds for latent time p.

3.5.3 Region B

In region $B = \mathbb{R}_+^p / A / C$, neither of these two arguments apply so both sets of bounds must be evaluated and the tighter values used.

3.6 Example

To illustrate these bounds, they were calculated using a simulated data set. The joint distribution was chosen so the marginals of the two causes, C, conditional of the covariate, Z, were exponential with hazards as set out in table 3.1.

	$Z=0$	$Z=1$
$C=1$	1	1.5
$C=2$	2	2.5

Table 3.1: Marginal hazards

A dependency was induced between the two causes, by assuming there was a Gamma frailty with mean, 1, and variance 2. This was achieved using the algorithm derived in Genest and MacKay (1986). Hence the time transformation is defined as

$$\phi(\mathbf{t}) = \begin{pmatrix} 1.5t_1 \\ 2.5/2t_2 \end{pmatrix}.$$

The 10,000 realised values are plotted in figure 3.2.

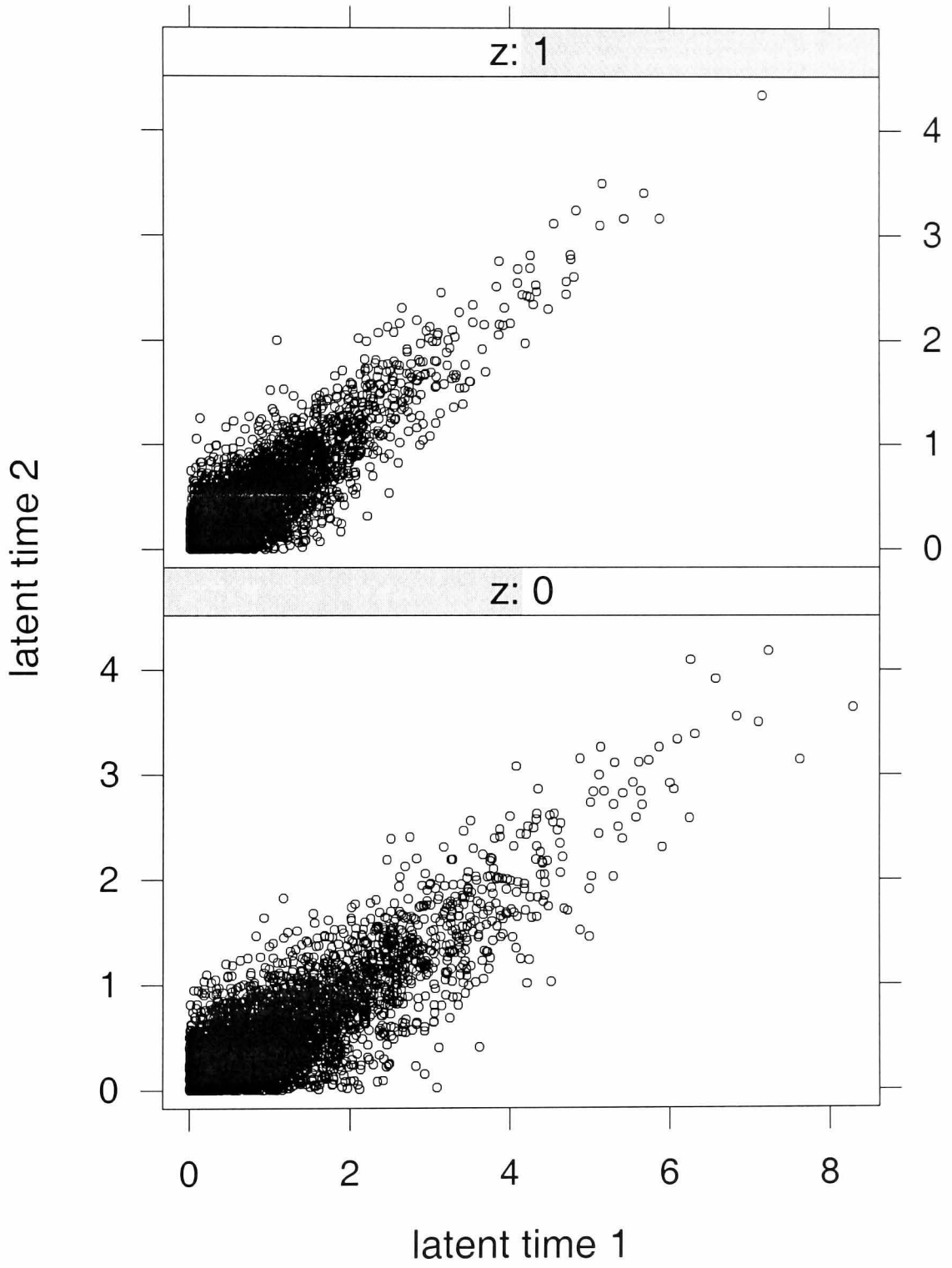


Figure 3.2: Realised values

From this the minimum of the two times was taken, and estimators of the crude incidence functions, conditional on Z, were calculated using the method in Prentice and Kalbfleisch (1978). With these the special case when $u_2 = 0$, i.e. the marginal distribution of T_1, was considered, and $Z = 0$, the improved bounds were calculated. No attempt was made to estimate the time transformation, ϕ as as this is known in advance, although in reality they would have to be estimated from the data. The results are shown in figure 3.3.

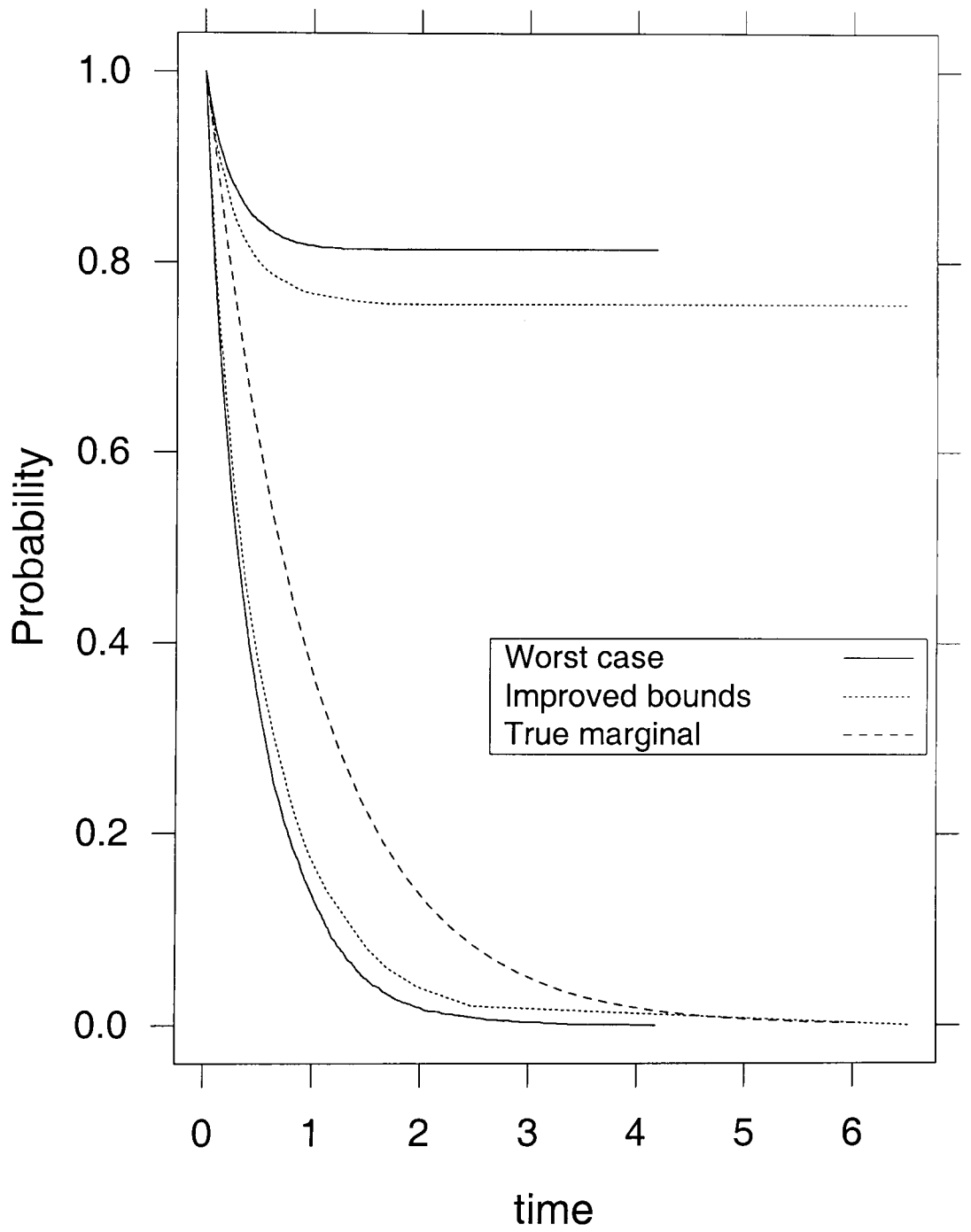


Figure 3.3: Improved bounds

Given the sample size of 10,000, the margin of error associated with these estimates is negligible. However, when a subset of size 100, with an equal split between the values of Z, was taken and the values of the hazard ratios, α_i for cause i, were estimated, using Cox proportional hazards (Cox 1972), the results are not so positive.

The estimate of α_1 was very crude as it assumes independence between the latent failure times; the data was regressed on Z, and cases other than $C = 1$ were treated as censored. The estimate $\widehat{\alpha}_1 = 1.77$ was obtained. The bounds obtained for the marginal distribution of T_1 conditional on $Z = 0$, are shown in figure 3.4,

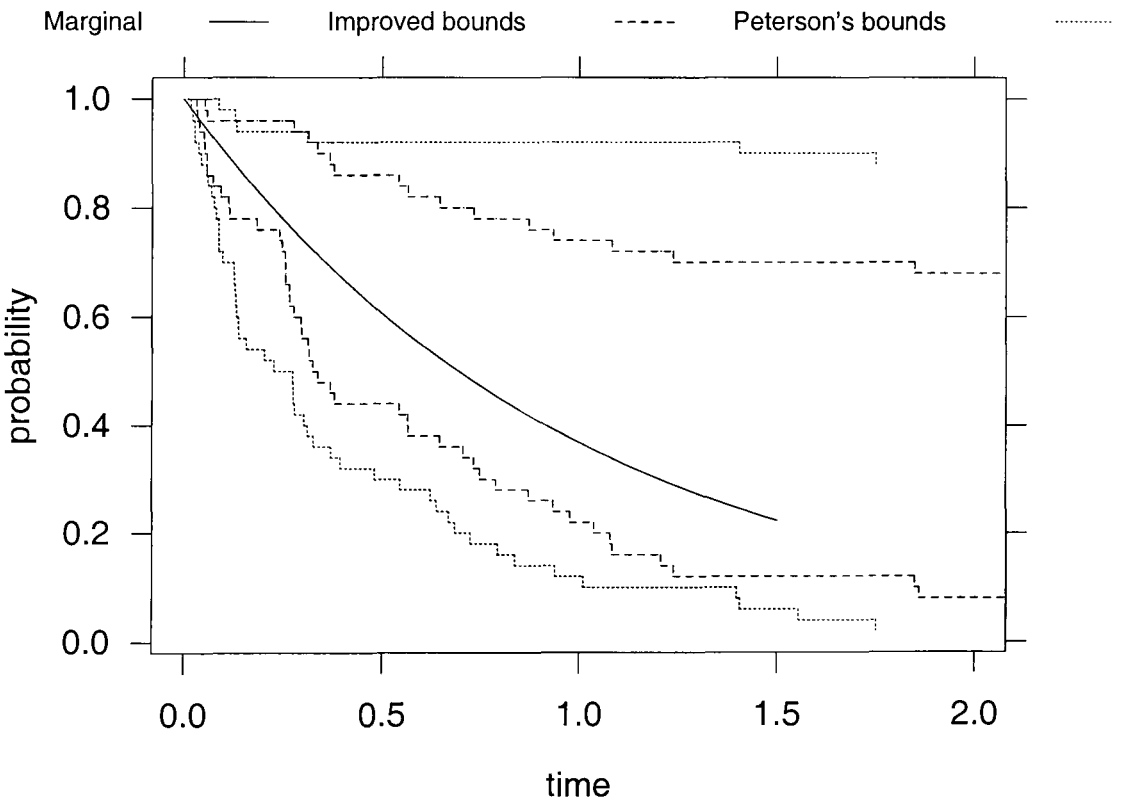


Figure 3.4: Bounds on a data set, $n = 100$

where it is observed that the upper bounds are briefly in the wrong order between $t = 0$ and $t = 0.5$. If we skip ahead to chapter 4 where we obtain bounds on the

covariate-time transformation and use this specific data set as an example, examining figure 4.2 we see that the estimated covariate-time transformation of $\phi_1(t) = 1.77t$ lies above its upper bound for a brief period early on. This is a consequence of the modelling assumptions not exactly agreeing with the data. The effect is that it forces the bounds on the marginal survival to be in the wrong order for a brief period early on.

3.7 Confidence Bands

From the derivations of the asymptotic form of $Q_i(t) = F_i(0) - F_i(t)$, in chapter 2 we see that the estimates of the crude incidence function can be represented in terms of orthogonal martingales $\{M_{\cdot 1}(t), \ldots, M_{\cdot k}(t)\}$ as:

$$\widehat{Q}_i(t) - Q_i(t) = \int_0^t \frac{\widehat{S}(u)J(u)}{Y_\cdot(u)} dM_{\cdot j}(u) + \int_0^t Q_j(u)\frac{dM_\cdot(u)}{Y_\cdot(u)} - Q_j(t)\int_0^t \frac{dM_\cdot(u)}{Y_\cdot(u)}.$$

Hence the covariation process for *cross-terms* is

$$\begin{aligned}
&\mathrm{Cov}(\widehat{Q}_i(s) - Q_i(s), \widehat{Q}_j(t) - Q_j(t)) \\
&= \int_0^{s\wedge t} \frac{Q_i(u)Q_j(u)d\Lambda_\cdot(u)}{Y_\cdot(u)} + Q_i(s)Q_j(t)\int_0^{s\wedge t} \frac{d\Lambda_\cdot(u)}{Y_\cdot(u)} \\
&- Q_i(s)\int_0^{s\wedge t} \frac{Q_j(u)d\Lambda_\cdot(u)}{Y_\cdot(u)} - Q_j(t)\int_0^{s\wedge t} \frac{Q_i(u)d\Lambda_\cdot(u)}{Y_\cdot(u)} \\
&+ \int_0^{s\wedge t} \frac{S(u)J(u)Q_j(u)}{Y_\cdot(u)} d\Lambda_i(u) - Q_j(t)\int_0^{s\wedge t} \frac{S(u)J(u)}{Y_\cdot(u)} d\Lambda_i(u) \\
&+ \int_0^{s\wedge t} \frac{S(u)J(u)Q_i(u)}{Y_\cdot(u)} d\Lambda_j(u) - Q_i(t)\int_0^{s\wedge t} \frac{S(u)J(u)}{Y_\cdot(u)} d\Lambda_j(u)
\end{aligned}$$

So using this, along with the expression for $\text{Cov}(\widehat{Q}_i(s) - Q_i(s), \widehat{Q}_i(t) - Q_i(t))$ given in theorem 2.5.3 we can calculate expressions for the variation process of $\sum_i \widehat{F}_i(u_i)$ for arbitrary $(u_1, \ldots, u_k)$. This is the form for both set of bounds. With some mild assumptions such as proposition 2.5.2 it can be shown that, with a suitable sequence of scaling constants a_n, Rebolledo's central limit applies (Rebolledo 1980) and hence confidence bands and intervals can be formed.

3.8 Summary

I have derived a set of bounds that, in some regions of the latent time space, improve the bounds on the joint survival function that can be inferred from competing risks data. These bounds can be obtained if we know the mapping ϕ, which is a strong assumption, and without specifying more structure, such as a frailty model, we can only obtain bounds on this function ϕ. This will be considered in the next chapter. How much these new bounds improve upon the existing Peterson bounds depends on which particular points in the latent time space are of interest. If we are considering a marginal survival then, as an extreme case, the bounds will converge if the line $\phi\{(t, \ldots, t)\}$ lies along the axis of the latent time that we are considering. This would be a very strong covariate effect. At the other extreme if the mapping $\phi\{(t, \ldots, t)\}$ does nothing, and coincides with the diagonal $(t, \ldots, t)$, then effectively we have a homogeneous sample and thus the new bounds will be the same as the Peterson bounds.

Chapter 4

Estimates of the covariate-time transformation

4.1 Introduction

Chapter 3 considered what could be inferred about the joint survival function, assuming that the covariate time transformation (CTT), $\phi(t, \mathbf{z})$, was known *a priori*. Clearly, this is not an assumption that can normally be made, hence, in this chapter, we will consider what properties this function can hold.

First we consider the implications of simplifying the covariate-time transform $\phi(\mathbf{t})$ to the case where $\phi_i(\mathbf{t}) = \phi_i(t_i)$. With this assumption we then show that the transformations are unique and are non-decreasing. The main result of the chapter is the bounds on the covariate-time transform in the case of a binary covariate. The chapter finishes by putting confidence intervals on these bounds, considers the limitations of the bounds and calculates them in a simulated data set where the 'true' transformations are known.

4.2 Elementary Properties

In a most general sense, the effect of covariates, Z, can be represented as a transformation of the time axes, where

$$S(t_1, \ldots, t_p|Z = z) = S(\phi_1(\mathbf{t}, z), \ldots, \phi_p(\mathbf{t}, z)|Z = z_0).$$

These transformations are not unique because any map which preserves the contours of the survival function, when $Z = z$, can be applied first and it will not affect the the joint survival. Hence the definition will be satisfied by any function, $\boldsymbol{\phi}(\mathbf{t}, z)$, which maps the contours $S(\mathbf{t}|Z = z_0) = k$ to the contours $S(\mathbf{t}|Z = z)$ for every value of $k, 0 \leq k \leq 1$.

Here we will consider the implication of the special case where $\phi_i(t_1, \ldots, t_p; z) = \phi_i(t_i; z)$ for all values of z, or equivalently, $\partial\phi_i/\partial t_j = 0,\ i \neq j$. In the notation of chapter 3 we have that $\boldsymbol{\Phi}(t)$ the transformation of the diagonal is equal, componentwise, to $\phi_i(t\mathbb{1})$, alternatively

$$\Phi_i(t) = \phi_i(t),$$

hence the inverse transformation of the general $\boldsymbol{\phi}(\mathbf{t})$ coincides with $\psi_i(t_i)$,

$$\boldsymbol{\phi}_i^{-1}(\mathbf{t}) = \psi_i(t_i).$$

4.2.1 Examples

Most of the models used in the literature (Cox and Oakes 1984, chapter 5) to describe multivariate and univariate survival distributions can be represented using a CTT of the form (4.1).

$$S(\mathbf{t}|Z = z) = S\{\phi_1(t_1, z), \ldots, \phi_k(t_k, z)|Z = z_0\} \tag{4.1}$$

Accelerated Failure Time

Accelerated failure time models (Kalbfleisch and Prentice 2002, chapter 7) can be represented using such CTTs, since their assumption is that

$$S(\mathbf{t}; z) = S_0 \left\{ f_1(z)t_1, \ldots, f_k(z)t_k \right\}$$

for some specified function, $S_0\{\cdots\}$. This is of the form (4.1) with $\phi_i(t_i, z) = f_i(z)t_i$, where the reference value, z_0, has to satisfy $f_i(z_0) = 1$.

Cox's Proportional Hazards model

The commonest model—the independent proportional hazards model (Cox 1972)—can also be represented since it assumes that

$$S(\mathbf{t}; z) \quad = \quad \exp\left\{ -\sum_i f_i(z)\Lambda_i(t_i) \right\}.$$

Hence the CTTs are functions, $\phi_i(t_i, z)$, such that

$$\begin{aligned} \Lambda_i \left\{ \phi_i(t_i, z) \right\} &= f_i(z)\Lambda_i \left\{ t_i \right\} \\ \Rightarrow \quad \phi_i(t_i, z) &= \Lambda_i^{-1} \left\{ f_i(z)\Lambda_i(t_i) \right\}, \end{aligned}$$

where Λ_i defines the cause-specific cumulative hazard functions conditional on $Z = z_0$ and z_0 is a reference value such that $f_i(z_0) = 1$.

Numerous other models, including the proportional odds model (Bennet 1983) and the additive hazards model (Aalen 1980), can also be described using the CTT framework.

4.2.2 Further Properties

A useful property under the assumption that $\partial\phi_i/\partial t_j = 0$ is that the functions $\phi_i(\cdot, \cdot)$ are unique.

Theorem 4.2.1. *If there exist functions $\phi_i(t_i; z)$ such that*

$$S(t_1, \ldots, t_p|Z = z) = S(\phi_1(t_1, z), \ldots, \phi_p(t_p, z)|Z = z_0),$$

and the cause-specific hazards are non-zero everywhere, then the functions ϕ_i are unique.

Proof. First, examine the case where $t_1 = \ldots = t_p = 0$, then

$$1 = S(0, \ldots, 0|Z = z) = S(\phi_1(0, z), \ldots, \phi_p(0, z)),$$

$$\Rightarrow \qquad \phi_i(0, z) = 0,$$

Hence, denoting the marginal survival function of T_i as $S_i(t|Z = z)$

$$S_i(t|Z = z) = S_i(\phi_i(t)|Z = z_0)$$

$$\Rightarrow \qquad \phi_i(t) = S_i^{-1}\{S_i(t|Z = z)|Z = z_0\}$$

Since these marginal functions clearly are unique, so are their composition, and inverses.

□

However their existence cannot be guaranteed, so it is useful to understand the implications of their existence. One useful property is that the Copula, and hence a large class of dependence measures, is invariant to the value of the covariate. A summary of recent developments in copula theory is given in Nelsen (1998).

Theorem 4.2.2. *There exist functions $\phi_i(t_i; z)$ such that*

$$S(t_1, \ldots, t_p|Z = z) = S(\phi_1(t_1, z), \ldots, \phi_p(t_p, z)|Z = z_0),$$

if, and only if, the survival Copula ($C_z(u_1, \ldots, u_p)$ s.t. $S(t_1, \ldots, t_p|Z = z) = C_z(S_1(t_1, z), \ldots, S_p(t_p, z))$) is invariant to values of z.

Proof. We know from theorem 4.2.1 that $\phi_i = R_i^{-1} \circ S_i$, where $R_i(t) = S_i(t|Z = z_0)$. Hence,

$$\begin{aligned}
& S(t_1, \ldots, t_p|Z = z) = S(\phi_1(t_1, z), \ldots, \phi_p(t_p, z)|Z = z_0) \\
\iff \quad & C_z[S_1(t_1, z), \ldots, S_p(t_p, z)] = C_{z_0}[R_1(\phi_1), \ldots, R_p(\phi_p)] \\
& \qquad = C_{z_0}[R_1\{R_1^{-1} \circ S_1(t, z)\}, \ldots, R_p\{R_p^{-1} \circ S_p(t, z)\}] \\
& \qquad = C_{z_0}[S_1(t_1, z), \ldots, S_p(t_p, z)] \\
\iff \quad & C_z = C_{z_0}.
\end{aligned}$$

□

Unfortunately, in a competing risks setting this is not very useful, per se, in estimating the functions ϕ_i, since it is well known that their marginal distribution functions are non-identifiable.

An elementary property of the functions $\phi_i(t)$ is that they are monotonic increasing functions

Theorem 4.2.3. *If* $S(t_1, \ldots, t_p \mid Z = z) = S(\phi_1(t_1, z), \ldots, \phi_p(t_p, z) \mid Z = z_0)$, *then* $\phi_i(u, z) > \phi_i(v, z)$, *for* $u > v$.

Proof. It follows that if sets A, B are such that $A \subset B$, then $\mathbb{P}(A) < \mathbb{P}(B)$, hence if $u > v$ then $\{T_i > u\} \subset \{T_i > v\}$, and therefore $\bigcap_{j \neq i}\{T_j > t_j\} \cap \{T_i > u\} \subset \bigcap_{j \neq i}\{T_j > t_j\} \cap \{T_i > v\}$. Taking the probability of these events

$$\begin{aligned}
& S(t_1, \ldots, u, \ldots, t_p \mid Z = z) < S(t_1, \ldots, v, \ldots, t_p \mid Z = z) \\
& \Rightarrow S(\phi_1(t_1, z), \ldots, \phi_i(u, z), \ldots, \phi_p(t_p, z) \mid Z = z_0) \\
& < S(\phi_1(t_1, z), \ldots, \phi_i(v, z), \ldots, \phi_p(t_p, z) \mid Z = z_0) \\
& \Rightarrow \phi_i(u, z) > \phi_i(v, z)
\end{aligned}$$

□

4.3 Bounds

Hereafter we will only consider the special case of a two-armed trial, where Z takes values in $\{0,1\}$; we set $z_0 = 0$ and simplify the notation so that $\phi_i(t) = \phi_i(t, z = 1)$. Now at any time point t it is possible to permute the indices referring to the causes of failure such that

$$\phi_1(t) \geq \phi_2(t) \geq \ldots \geq \phi_K(t),$$

which, given the ϕ_i are non-decreasing, is equivalent to

$$\psi_1(t) \leq \psi_2(t) \leq \ldots \leq \psi_K(t),$$

where $\psi_i = \phi_i^{-1}$. Assuming the ϕ are continuous functions, this ordering will hold over an interval $(u, v]$ containing t, and the number of such disjoint intervals is countable. Initially we will assume that there is just one interval: the entire real line.

Now note that, for any set $\mathcal{C}$ of causes of failure,

$$\bigcup_{i \in \mathcal{C}} \{C = i\} = \bigcup_{i \in \mathcal{C}} \bigcap_{j \neq i} \{T_i < T_j\} = \bigcup_{i \in \mathcal{C}} \bigcap_{j \notin \mathcal{C}} \{T_i < T_j\}, \tag{4.2}$$

since $\bigcup_{i \in \mathcal{C}} \bigcap_{j \in \mathcal{C}/\{i\}} \{T_i < T_j\}$ is the event that one of the T_is, restricted to $i \in \mathcal{C}$, is the minimum of the T_is, restricted to $i \in \mathcal{C}$, and this is always true. The right hand side of equation (4.2) is more mathematically convenient.

Consider $\mathcal{C} = \{1, 2, \ldots, m\}$ and observe that for $i \leq m < j$

$$T_i < T_j \Rightarrow \psi_i(T_i) < \psi_i(T_j) < \psi_j(T_j),$$

by the monotonicity and ordering of the ψ, hence

$$\bigcup_{i \leq m} \bigcap_{j \neq i} \{T_i < T_j\} \subset \bigcup_{i \leq m} \bigcap_{j \neq i} \{\psi_i(T_i) < \psi_j(T_j)\}. \tag{4.3}$$

This is useful as, under the assumption of equation (4.1) about the CTT

$$\mathbb{P}\{C \in \mathcal{C}|Z=1\} = \mathbb{P}\left\{ \bigcup_{i\in\mathcal{C}} \bigcap_{j\notin\mathcal{C}} \{\psi_i(T_i) < \psi_j(T_j)\} \middle| Z=0 \right\},$$

so we can use equation (4.3) to order the $\mathbb{P}\{C \in \mathcal{C}|Z\}$ for the different Z. Choosing a different $\mathcal{C} = \{m, m+1, \ldots, K\}$, we can reverse the ordering in (4.3), since for $j < m \leq i$,

$$\psi_i(T_i) < \psi_j(T_j) \Rightarrow T_i < T_j$$

hence

$$\bigcup_{i\geq m} \bigcap_{j\neq i} \{T_i < T_j\} \supset \bigcup_{i\geq m} \bigcap_{j\neq i} \{\psi_i(T_i) < \psi_j(T_j)\} \tag{4.4}$$

Now consider

$$\begin{aligned}
\sum_{i\leq m} Q_i\{t|Z=1\} &= \mathbb{P}\left\{ \bigcup_{i\leq m} \left(\{T_i < t\} \cap \bigcap_{j\neq i} \{T_i \leq T_j\} \right) \middle| Z=1 \right\} \\
&= \mathbb{P}\left\{ \bigcup_{i\leq m} \left(\{T_i < \phi_i(t)\} \cap \bigcap_{j\neq i} \{\psi_i(T_i) < \psi_j(T_j)\} \right) \middle| Z=0 \right\} \\
&> \mathbb{P}\left\{ \bigcup_{i\leq m} \left(\{T_i < \phi_i(t)\} \cap \bigcap_{j\neq i} \{T_i \leq T_j\} \right) \middle| Z=0 \right\} \\
&> \mathbb{P}\left\{ \bigcup_{i\leq m} \left(\{T_i < \phi_m(t)\} \cap \bigcap_{j\neq i} \{T_i \leq T_j\} \right) \middle| Z=0 \right\} \\
&= \sum_{i\leq m} Q_i\{\phi_m(t)|Z=0\}
\end{aligned}$$

where the first and last equalities are by definition, the second equality is through the CTT assumptions, the first inequality is through equation (4.3), and the second inequality is through the ordering of the ϕ. If we consider $i \geq m$ then equation (4.4)

allows us to reverse the inequalities and, to summarise, we have the two inequalities:

$$\sum_{i \leq m} Q_i\{\phi_m(t)|Z=0\} \quad < \quad \sum_{i \leq m} Q_i\{t|Z=1\}, \tag{4.5}$$

$$\sum_{i \geq m} Q_i\{t|Z=1\} \quad < \quad \sum_{i \geq m} Q_i\{\phi_m(t)|Z=0\}. \tag{4.6}$$

Since the Q_i are directly estimable these can be converted into bounds on the unknown functions ϕ_m.

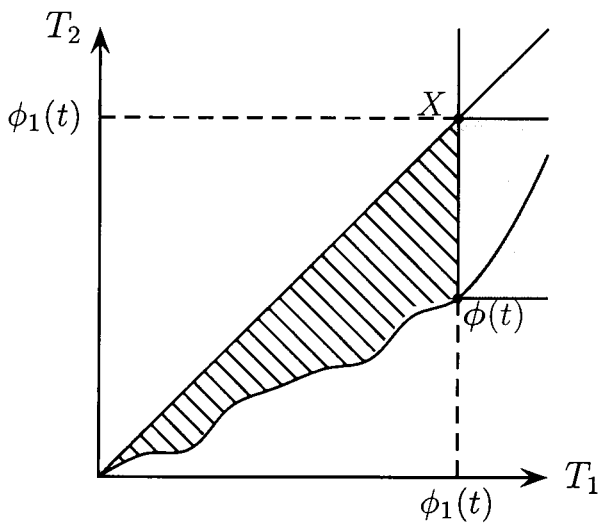


Figure 4.1: Illustration of the 2-d case

A geometrical interpretation of the 2-d case is shown in figure 4.1. This shows the plane of the two latent failure times and the curved line represents the line $\phi(t)$. The modelling assumption (4.1) means that $Q_1\{t|Z=1\}$ is the integral of the density, conditional on $Z=0$, over the region above the curved line, and to the left of the vertical line through X; $Q_1\{\phi_1(t)|Z=0\}$ is the integral over the region above the line $T_1 = T_2$, and to the left of the vertical line through X. The inequality (4.5) comes from the fact that the integral over the dashed region is non-negative. Observe that $Q_1\{t|Z=1\} + Q_2\{t|Z=1\}$ is equal to the integral over an area sandwiched between two boundaries: the two axes; a translation of the axes such that the image of the origin is $\phi(t)$. A similar region, but with the translated origin at X, gives the quantity

$Q_1\{\phi_1(t)|Z = 0\} + Q_2\{\phi_1(t)|Z = 0\}$. Hence equation (4.6) comes from the non-negativity of the integral over the solid region. This shows that the bounds are tight if the underlying joint distribution puts zero mass in either of the highlighted regions.

4.3.1 Ordering

Unfortunately, there is still some more work to be done since the ordering of the ϕ_i is not known in advance. If these bounds are to be used to check a model, or a further set of assumptions about the CTT, then this is not a problem. If the model produces estimates of the CTT then the estimates can be ordered, the bounds can be calculated, and if the estimates lie within the bounds then this would support the modelling assumptions.

The ordering does *not* need to be known in advance when there are *only two* causes of failure. For both causes the ϕ_i share a common bound, $c(t)$, which is obtained from the equality $\sum_{i=1}^{2} Q_i\{c(t)|Z = 0\} = \sum_{i=1}^{2} Q_i\{t|Z = 1\}$. For one cause this is the lower bound and for the other cause it is the upper bound. The other bounds, $b_i(t)$, are obtained from $Q_i\{b_i(t)|Z = 0\} = Q_i\{t|Z = 1\}$, and the correct ordering, and whether the bounds are upper or lower, becomes apparent when the three bounds are calculated. This is illustrated in section 4.6.

It is also true, in the case of two causes of failure, that the upper and lower bounds coincide if the two ϕ_i are equal. This is because, when equal, the ϕ_i must lie on the diagonal $(t_1 = t_2)$. Hence it is always possible to calculate the bounds and identify the ordering in the case of two causes of failure.

In the case with three or more failures there is no fail-safe route to determining the ordering. A possible avenue for investigation uses the result shown in Heckman and Honoré (1989),

$$\dot{Q}_i(0|Z = z_1)/\dot{Q}_i(0|Z = z_2) = \dot{\phi}_i(0, z_1)/\dot{\phi}_i(0, z_2),$$

where the dot ($\dot{}$) notation indicates the partial time derivative. In a two armed trial we know that $\dot{\phi}_i(t, Z = 0) = 1$, so we have identified $\dot{\phi}_i(0, Z = 1)$. Given that $\phi_i(0, z) = 0$, knowing the derivative at the origin, under continuity assumptions, gives the ordering of the ϕ_i immediately after the origin. Unfortunately there is no method to determine if the ϕ_i subsequently change their order in the case of three or more causes of failure. This is because in more that two-dimensions the line $(\phi_1(t), \ldots, \phi_k(t))'$ does not have to intersect the line $(t, \ldots, t)'$ when the ordering changes, whereas it does in two dimensions.

4.4 Confidence intervals

To obtain confidence intervals on the bounds in equations (4.5) and (4.6), consider the random process

$$\sum_{i \leq m} \left\{ \widehat{Q}_i(t|Z = 1) - \widehat{Q}_i(u|Z = 0) \right\} = \widehat{D}_m(t, u).$$

Using the counting process approach of Andersen et al. (1993) it is shown in chapter 2 that the two left-hand terms converge, asymptotically, to Gaussian processes. Unfortunately, the covariation process is complex and does not exhibit independent increments, however a simulation approach described in Lin (1997) can be used to calculate confidence limits $\widehat{a}_\alpha(u), \widehat{b}_\alpha(u)$ such that $\mathbb{P}\{D(t, u) > \widehat{a}_\alpha(u)\} = 1 - \alpha/2$ and $\mathbb{P}\{D_m(t, u) < \widehat{b}_\alpha(u)\} = 1 - \alpha/2$ for a fixed value of t, where D_m is the function estimated by $\widehat{D}_m$. It follows that the roots of the equations, $\widehat{a}_\alpha(u), \widehat{b}_\alpha(u) = 0$, provide pointwise $(1 - \alpha)100\%$ confidence intervals for the bounds in equation (4.5). An identical argument for the bounds in (4.6) applies.

It is not clear if confidence bands can be formed for a continuum of values of v. The problem is that the bounds are of the form $G^{-1}(H(v))$, where G and H

are functions that are estimated with random error; how to cope with a convolution of two such processes and subsequently form confidence intervals is unknown. The delta method may be of use.

4.5 Limitations

It is worth pointing out that it is impossible to calculate the bounds in equations (4.5) and (4.6) for all time points and choices of index, m. This is because, if the covariate has *any* effect, there will exist choices of ordering the causes of failure and m such that

$$\mathbb{P}\{1 \leq C \leq m | Z = 1\} > \mathbb{P}\{1 \leq C \leq m | Z = 0\}.$$

In these cases it will be impossible to solve, for u, the relevant equation

$$\sum_{i \leq m} Q_i(u|Z=0) = \sum_{i \leq m} Q_i(t|Z=1), \tag{4.7}$$

for values of $t > t_\infty$ such that

$$\sum_{i \leq m} Q_i(t_\infty|Z=1) = \sum_{i \leq m} Q_i(\infty|Z=0) = \mathbb{P}\{1 \leq C \leq m | Z = 0\}.$$

The bound will explode at this point, t_∞, and consequently be of no practical use.

Fortunately, the lack of a solution to (4.7) does not invalidate the bounds since in scenario 1 the solution, u to (4.7) provides an upper bound to $\phi_i(t)$; at t_∞ we have that $\phi_i(t_\infty) < \infty$, and hence it is perfectly logical, if rather uninformative, to keep ∞ as an upper bound for $t > t_\infty$. The negative aspects of this are that the point t_∞ is invariant to sample size, and hence we will always have to cope with a guarantee of infinitely large bounds on the covariate-time transformation in at least one of the latent times.

Even more unfortunate is the fact that it apears fruitless to try and use these bounds in conjunction with the theory of chapter 3 for marginal distributions. This is

because a sufficient condition for the alternative bounds of chapter 3 to improve upon the conventional Peterson bounds for the T_i marginal distribution is that the T_i axis is nearer to the line $\mathbf{\Phi}(t)$ than to the diagonal $t\mathbb{1}$. Unfortunately this is precisely the same condition required for the upper bound, for ϕ_i, to explode to infinity.

4.6 Illustration

These bounds are illustrated on the data set of chapter 3, where there were two causes, indexed by C, and a binary covariate, Z, with marginal exponential distributions with hazards as displayed in table 4.1.

	$Z=0$	$Z=1$
$C=1$	1	1.5
$C=2$	2	2.5

Table 4.1: Marginal hazards

A dependency is induced by a frailty term with a gamma distribution with mean of 1 and a variance of 2.

From this simulated distribution the crude incidence function was estimated. Since the resulting estimate of the crude incidence function, $\hat{Q}(t)$ is a right continuous increasing function, the inverse function was defined to be

$$\hat{Q}^{-1}(p) = \min\{t : \hat{Q}(t) \geq p\}.$$

From this figure 4.2 was produced which illustrates the theoretical bounds. Further along the time axis the true function $\phi_2(t) = 1.25t$ intersects its upper bound, although this is due to the random error associated with the estimates of the cumulative incidence functions.

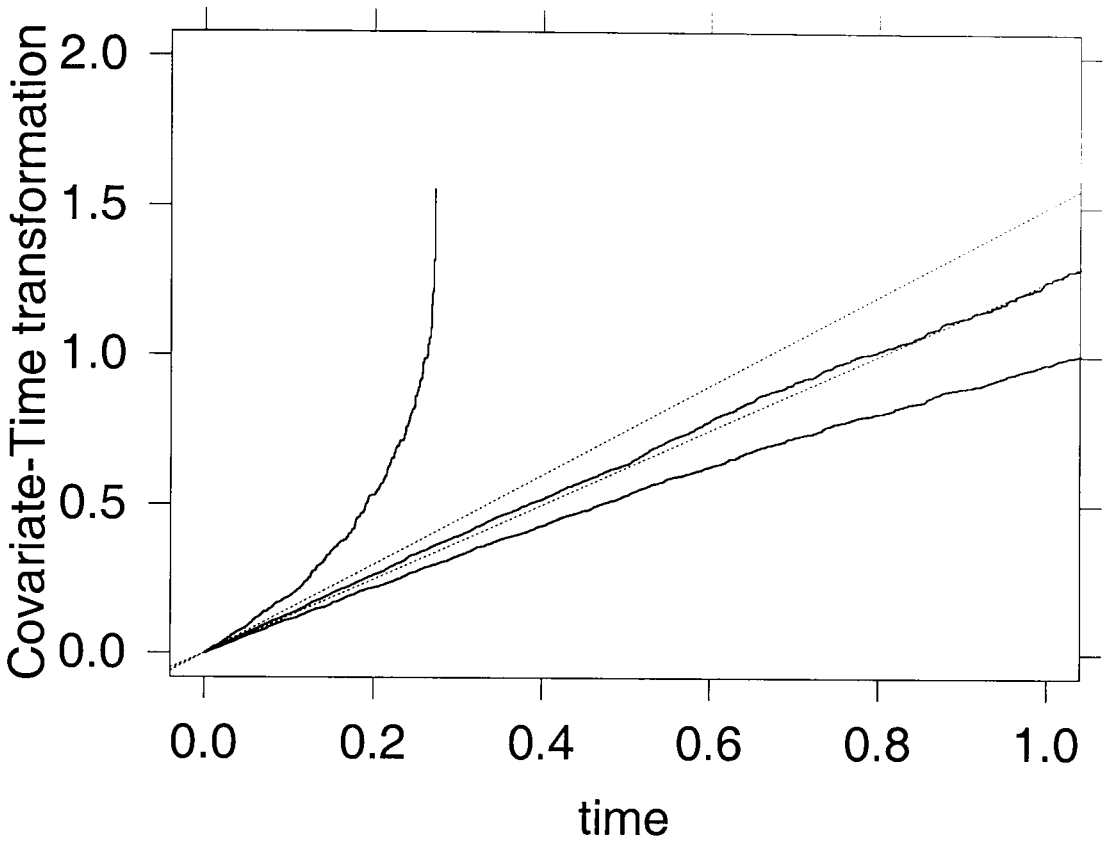


Figure 4.2: Bounds for $\phi(t)$, with the true value

4.7 Summary

This chapter has considered in general the covariate-time transformation. The first result is that the assumption of a simplified covariate-time transformation, $\phi_i(\mathbf{t}; z) = \phi_i(t; z)$, is equivalent to the assumption that the Copula of the dependence structure within the latent failure times is invariant to the covariates. In this case, it is shown that such functions ϕ_i are unique and are non-decreasing. In the case of a binary covariate, there are bounds on the functions ϕ_i that can be consistent with (perfect) competing risk information. We have considered the question of deriving confidence intervals for

these bounds, and have considered what conditions are needed for the bounds to be of practical use and not explode to infinity. The next chapter will illustrate how to use the results of this chapter and those of chapter 3 on an existing data set.

Chapter 5

Application to a two-armed trial

5.1 Background

In this chapter we will apply the results of chapters 3 and 4 to a two-armed trial. The data are in appendix A which are taken from Hoel (1972) and record the survival times of mice which received a radiation dose. The mice were then randomised into two treatments: the control was being kept in ordinary lab conditions and the treatment was being kept in a germ-free environment. There were three possible causes of death: thymic lymphoma, reticulum cell sarcoma and other.

5.2 Models

For illustrative purposes we will concentrate our attention on the sarcoma cause of failure. Two models were fitted which treated all other causes of failure as censorings: a Cox proportional hazards model (Cox 1972) and a Weibull model with a log link. The Weibull is both an accelerated failure time model, where the covariate-time transform is a straight line, and is also a proportional hazards model but where the hazard function is

constrained to belong to a two-parameter set of functions. Explicitly the hazard function is

$$\lambda(t;z) = \alpha\left(\exp(\lambda_0 + \beta z)t\right)^{\alpha}.$$

So the time transformation is, $t \mapsto \exp(\beta)t$, whereas the log hazard ratio is $\alpha\beta$. For this analysis $z = 0$ refers to the laboratory conditions group and $z = 1$ refers to the germ-free conditions group.

In the Weibull model, assuming a fixed shape parameter $\widehat{\alpha} = 6.94$, a 95% confidence interval for the log hazard ratio is $(-1.89, -3.10)$. This can be compared to the Cox model which has a 95% confidence interval for the log hazard ratio of $(-1.34, -2.72)$. The remaining parameters in the Weibull model, α and λ_0, had confidence intervals of $(5.63, 8.58)$ and $(1.41, 1.55) \times 10^{-3}$, respectively. This indicates that a germ-free environment lowers mortality.

5.3 Covariate-time transformations

Using the theory of chapter 4 we can see how plausible these two models are given the data.

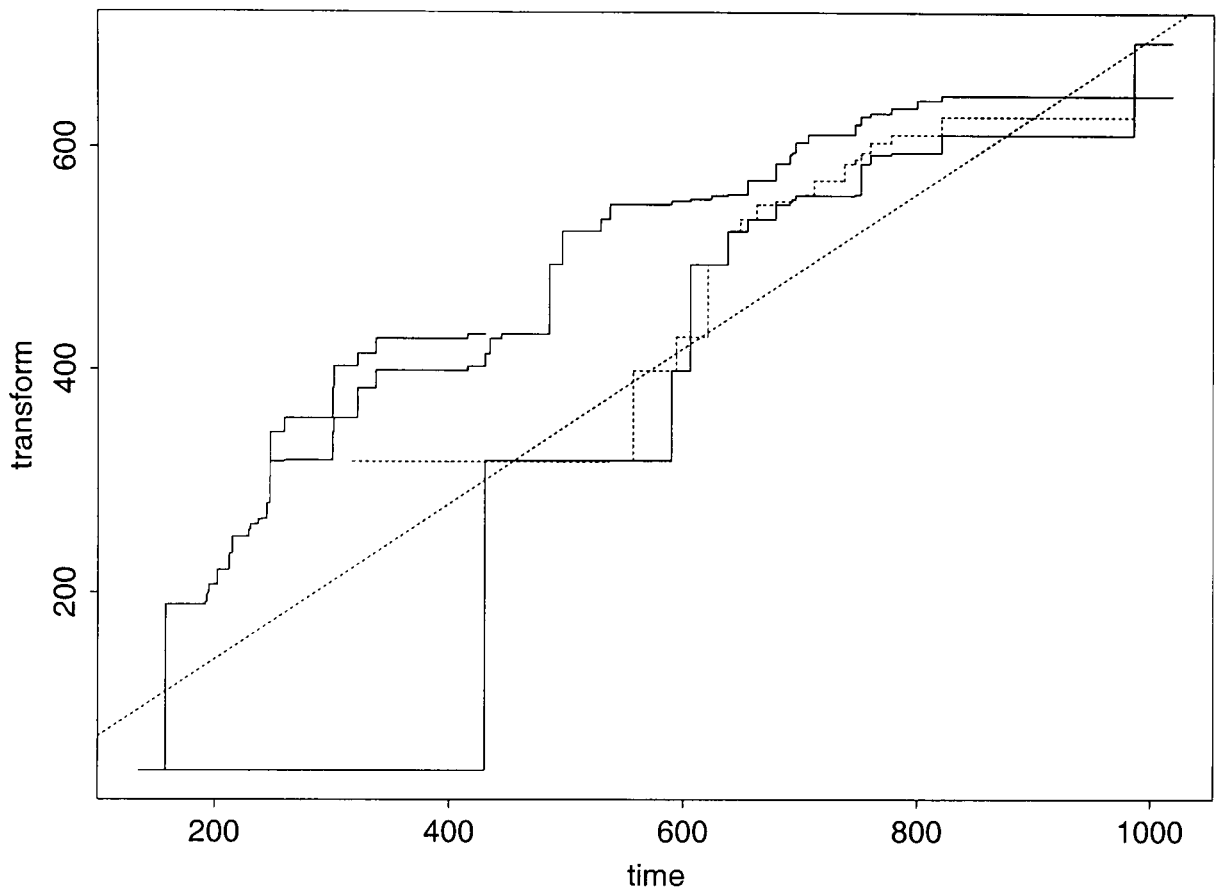


Figure 5.1: Bounds and estimates of the time transformation

This is shown in figure 5.1. There are three solid lines, although the uppermost pair coincide initially and split apart approximately at the point (250,300). Under the model that treats sarcoma and lymphoma as two latent failure times with 'other' being treated as uninformative censoring, these form the bounds outlined in chapter 4 for the covariate time transformation where the lower set of bounds refers to the sarcoma latent failure time, and the upper set of bounds refers to the lymphoma latent failure time. These were obtained by forming the crude incidence functions for the two groups and without making any modeling assumptions.

Since the uppermost line explodes to infinity fairly early on—just after 400 days, where it finishes in figure 5.1—and, earlier on, coincides with the middle line, there is not a sufficient amount of information to be able to use the bounds on the lymphoma time. Henceforth we will concentrate on the sarcoma time. The straight, dotted line which is the line $y = \exp(-0.359)x$ represents the time transformation as predicted by the independent Weibull model. As we can see it does the best it can for a straight line trying to lie in a non-concave region, but is clearly not suitable for this data. The remaining dotted line gives the time transformation as given by the Cox proportional hazards model. This is

$$t \mapsto \widehat{\lambda}_0^{-1}\{\exp(-2.03)\widehat{\lambda}_0(t)\},$$

where $\widehat{\lambda}_0$ is the standard estimate of the baseline hazard for the laboratory conditions group and, since it is an increasing step function, its inverse is defined as

$$\widehat{\lambda}_0^{-1}(x) = \min\{t : \widehat{\lambda}_0(t) \geq x\}.$$

This estimate lies mostly within its bounds and only goes outside for a period of approximately twenty days just after 600 days.

5.4 Marginal Survival estimates

Now when we come to estimate the marginal survival function of the sarcoma failure time, we see that we must be in a region where the alternative bounds derived in chapter 3 are wider than the conventional Peterson bounds. This is shown in figure 5.2.

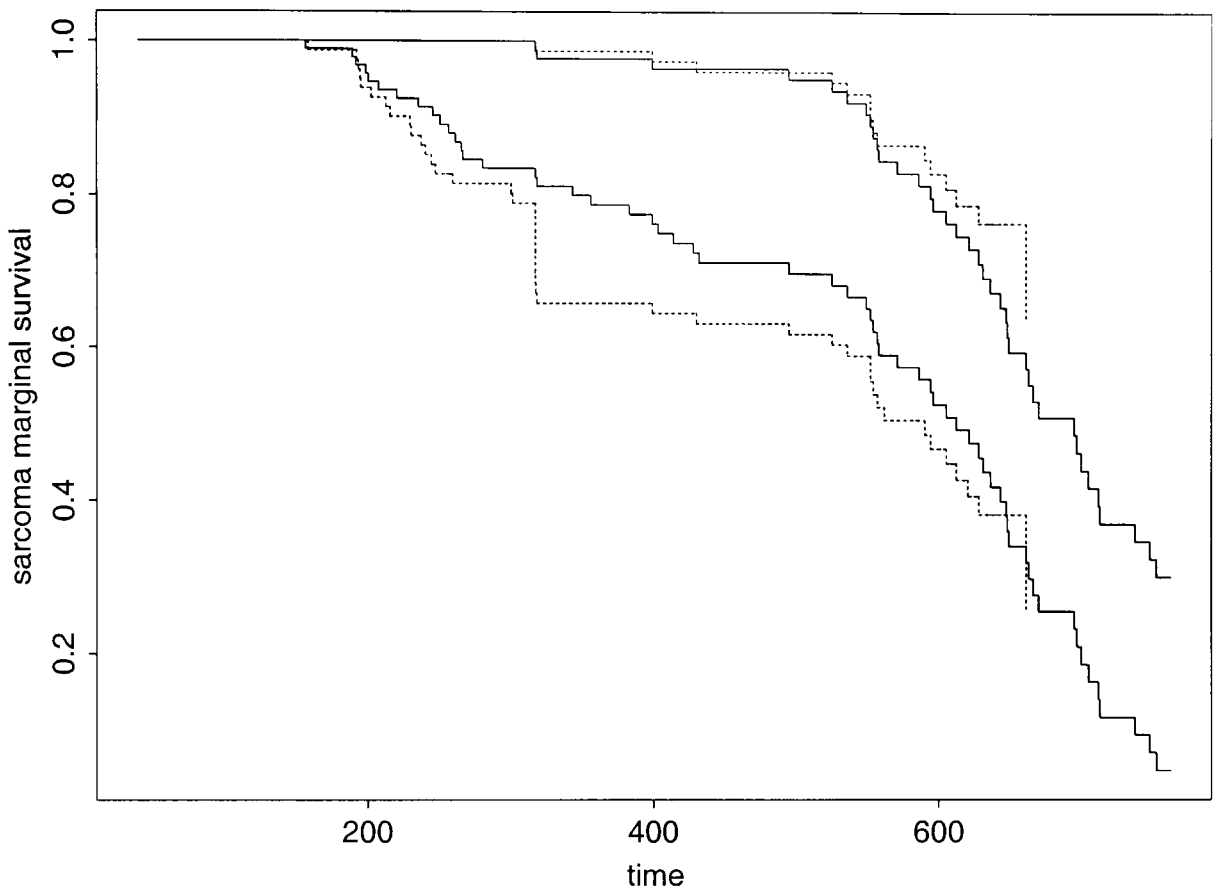


Figure 5.2: Sarcoma marginal survival

Here the solid lines are the bounds provided for sarcoma survival in the laboratory conditions group by the Peterson bounds and the dotted lines represent the alternative bounds derived in chapter 3 which use the Cox proportional hazards model to estimate the covariate-time transformation. We can clearly see that the alternative bounds are wider, and hence of no use, compared to the conventional bounds. However for a short period at around 620 to 650 days the lower alternative bound is higher than the conventional bounds. This is incorrect and it is propsosed that this is a consequence of the proportional hazards assumption being wrong. The proportional hazards assumption

forces the estimate of the covariate-time transformation to lie outside the correct region, at roughly the same time period, which is shown in figure 5.1.

5.5 Summary

This chapter has applied the results of chapters 3 and 4 to a two-armed trial. The bounds on the covariate-time transformation are useful in comparing two models, the Weibull and the Cox proportional hazards model, and indicate that the Weibull model is not flexible enough to model the sarcoma latent failure time. The data illustrate well the payoff between increased information about the covariate-time transformation and lack of information about the marginal survival curves. There are useful, reasonably narrow, bounds for the sarcoma time-transformation, but as a consequence the alternative bounds for the marginal sarcoma latent time are wider than the Peterson bounds. For the lymphoma cause, there is very little information on the covariate-time transformation since the bounds explode to infinity shortly after 400 days which is less than half of the 1000 days which are under consideration. If there was accurate knowledge of this covariate-time transformation then it could be used to calculate the narrower bounds derived in chapter 3 for the marginal lymphoma latent time distribution.

Chapter 6

Generalised identifiability for competing-risks with covariates

6.1 Fundamental problem

The main problem with competing risks data is that they can only provide information on the cause-specific survival functions. In a latent failure time setting, the natural objects of interest are the joint density or joint survival function. The problem is that it is a one-way street between the joint survival and the cause-specific survival. There are infinitely many distinct joint survival function which will give a specified set of cause-specific survivals.

This was proven in Tsiatis (1975). For any model of the joint survival function there exists a model which exhibits independence between the latent failure times that produces the same cause-specific survival functions. The proof rests on examining the derivatives of the cause-specific survival functions, rather the cause-specific survival functions themselves. This is because of the convenient mathematical relation-

ship between the derivative of the cause-specific survival function and the joint survival function.

$$\begin{aligned}
\frac{\partial F_i(t)}{\partial t} &= \lim_{\delta \to 0} \frac{-1}{\delta} \mathbb{P}(t < T^{\min} \leq t + \delta \cap C = i) \\
&= \lim_{\delta \to 0} \frac{-1}{\delta} \mathbb{P}(t < T_i \leq t + \delta \bigcap_{j \neq i} T_j > t) \\
&= \lim_{\delta \to 0} \frac{\{1 - \mathbb{P}\left(T_i \geq t + \delta \cap_{j \neq i} T_j > t\right)\} - \{1 - \mathbb{P}\left(T_i > t \cap_{j \neq i} T_j > t\right)\}}{\delta} \\
&= \lim_{\delta \to 0} \frac{S(t, \ldots, t, t + \delta, t, \ldots, t) - S(t, \ldots, t)}{\delta} \\
&= \frac{\partial S}{\partial t_i}(t\mathbb{1}), \qquad (6.1)
\end{aligned}$$

where $\mathbb{1}$ represents a k-dimensional vector of 1s.

So, to sketch the proof in Tsiatis (1975), given any specified joint survival function $S(\mathbf{t})$ which gives cause-specific survivals,

$$F_i'(t) = \frac{\partial S}{\partial t_i}(t\mathbb{1}),$$

then the new, independent joint survival function,

$$S^*(\mathbf{t}) = \exp\left[-\sum_i \int_0^{t_i} \frac{-F_i'(u)}{\sum_j F_j(u)} du\right],$$

will give identical cause specific survival functions. By inspection,

$$\frac{\partial S^*}{\partial t_i}(t\mathbb{1}) = \frac{F_i'(t)}{\sum_j F_j(t)} \exp\left[-\sum_i \int_0^{t} \frac{-F_i'(u)}{\sum_j F_j(u)} du\right],$$

since the range of integration is common, swapping the summation and integration

$$\begin{aligned}
&= \frac{F_i'(t)}{\sum_j F_j(t)} \exp\left[\int_0^{t} \frac{\sum_j F_j'(u)}{\sum_j F_j(u)} du\right], \\
&= \frac{F_i'(t)}{\sum_j F_j(t)} \exp\left[\ln\left\{\sum_j F_j(t)\right\}\right] \\
&= F_i'(t).
\end{aligned}$$

All this is assuming that we have a homogeneous sample, which is rarely the case. Chapters 3 and 4 considered what happens in the case of a binary covariate, and can be extended to a finite, discrete set of covariates, by choosing a baseline level of the covariate and comparing the different levels of the covariate, as if it were a binary covariate and then taking the narrowest set of bounds. However with a continuous covariate, whose effect can be represented through a continuous mapping, more results have been obtained.

An important result is from Heckman and Honoré (1989) where it is shown that given a continuous covariate, and the assumption that the covariate-time transformation, $\phi(t, z)$, has the form

$$\phi(t, z) = (\phi_1(z)\Lambda_1(t_1), \phi_2(z)\Lambda_2(t_2), \ldots, \phi_k(z)\Lambda_k(t_k)),$$

which is essentially a proportional hazards assumption, then the joint survival function can be identified. To be able to use the constructive proof given in the paper would require that the covariate is observed over a very large range of values, thus demanding a huge sample size. A useful extra assumption is that the dependence between the latent failure times is induced by an unobserved covariate having an effect on the distribution. This means that the *kernel* of the distribution—the joint survival at a baseline level of the covariate—is in the form of an integral. This integral has the role of marginalising with respect to the (unknown) density of the unobserved covariate. Because the *kernel* is in the form of an integral it must be an analytic function, in other words it can be represented by a Taylor expansion. It is shown in Abbring and van den Berg (2003), that this weakens the assumptions on which identifiability is obtained.

This chapter will consider if it is possible to generalise the assumption of pro-

portional hazards to where the form of the covariate-time transformation is

$$\boldsymbol{\phi}(t,z) = (\phi_1(t_1,z), \phi_2(t_2,z), \ldots, \phi_k(t_k,z)).$$

The general answer seems to be negative. It is shown that if the kernel is known then the covariate-time transformation can be identified and vice versa where if the covariate-time transformation is known then the kernel is identified. Unfortunately, these two cannot then be tied together since it is shown that there are infinitely many pairs of kernel & transformation which agree with the cause-specific survival functions, but disagree on the general joint survival function.

6.2 Assumptions

It is assumed that we are considering a data set where the individual has a set of latent failure times $\{T_1, \ldots, T_k\}$ but the observations only consist of the minimum time $T^{\min} = \min\{T_1, \ldots, T_k\}$ and the cause of failure $C = \text{argmin}\{T_1, \ldots, T_k\}$. In addition to this there is assumed to be a set of l-dimensional covariates, Z, where $l \geq k-1$.

If interest is in the underlying latent failure-time joint-distribution then, without further assumptions about its functional form, it is impossible to make any inferences. Here we will examine the particular assumption that the joint survival function $S(\mathbf{t}|Z) = \mathbb{P}(T_1 > t_1, \ldots, T_k > t_k|Z)$ takes the functional form

$$S(\mathbf{t}|Z) = K\left[\phi_1(t_1,z), \ldots, \phi_k(t_k,z)\right],$$

where K is a function such that $K(\mathbf{0}) = 1$, and $K(\infty) = 0$. The conditions on ϕ_i are that $\phi_i(t,z)$ is increasing in t, there exists a z^* such that $\phi_i(t,z^*) = t$ for all i, and $\phi_i(0,z) = 0$ for all i, z. An example of this functional form is an independent, proportional hazards model, where $\phi_i(t_i,z) = \exp(\beta_i' z)\Lambda_i(t_i)$, and $K(\mathbf{x}) = \exp\left(-\sum_i x_i\right)$; this can be

generalised to a frailty model, as considered in chapter 7, by keeping the same form for ϕ_i but replacing K with the Laplace transform of the frailty distribution. The key assumption here is that the different latent failure times can always be split up from each other, in some manner, and we do not need any terms such as $\phi(t_1, t_2, z)$, say. As we are concerned with identifiability we assume that we 'know', without any random error, the set of *cause-specific survival functions*,

$$F_i(t, z) = \mathbb{P}(T^{\min} > t \cap C = i|z).$$

6.3 Identifiability Results

The fundamental question is then whether, given the functional form assumptions, there is a unique set of $\phi_i(t, z)$ and $K(\cdot)$ that gives rise to the cause-specific survival functions. An alternative phrasing of our question is: is there a bijection between the joint survival function and the functions in equation (6.1)?

One possible tactic in attempting to prove identifiability is to show that if we know the ϕ_i then we can identify the K, whereas if we know the K then we can find the ϕ_i. If we were then to iterate between finding K from a given ϕ, and then using this K to find a new 'improved' ϕ we may find that the pair (K, ϕ) would converge. If a fixed point were found then it would satisfy the assumptions and give the correct cause-specific survival functions. Unfortunately, this fixed point is not unique. We will show that there exists an infinite set of such stationary points all of which give different functions for the joint survival function, and hence we still have the yoke of non-identifiability.

To start with we show that indeed, given either K or ϕ, the other can be found.

Theorem 6.3.1. *Given the general assumptions, along with assuming that the func-*

tions $\phi_i(t,z)\ i = 1,\ldots,k$ *are known and that the mapping* $\Phi : \mathbb{R}_+ \times \mathbb{R}^l \mapsto \mathbb{R}_+^k :$ $(t,z) \mapsto (\phi_1(t,z),\ldots,\phi_k(t,z))$*, is a surjection, then the function* $K(x_1,\ldots,x_k)$ *can be identified from the cause-specific survival functions.*

Proof. From the definition of the cause-specific survival function it is clear that $\sum_i F_i(t,z) = S(t\mathbb{1}|z)$, and this can be evaluated. So if we can find a $t(\mathbf{x}) \in \mathbb{R}_+$ and a $z(\mathbf{x}) \in \mathbb{R}_+^l$ which maps to $\mathbf{x} = (x_1,\ldots,x_k)$ under the mapping Φ, then by definition $K(x_1,\ldots,x_k) = \sum_i F_i(t(\mathbf{x}),z(\mathbf{x}))$. But as we assume that Φ is a surjective mapping, such $t(\mathbf{x})$ and $z(\mathbf{x})$ exist. □

The next theorem shows that if the joint survival function, K, is known then we can identify the time-transformations ϕ.

Theorem 6.3.2. *Assuming that the function* $K : \mathbb{R}^k \mapsto [0,1]$ *is known, has continuous, non-zero first derivatives and that*

$$\frac{\partial K}{\partial x_i}(\mathbf{\Phi}(t,z)) = O\left(\frac{\partial F_i}{\partial t}(t,z)\right),$$

as t *goes to infinity, then* $\phi_i(t,z)$ *can be found.*

Proof. Using (6.1), we have defined a set of first order differential equations,

$$\frac{\partial \phi_i}{\partial t}(t,z) = \frac{\partial F_i}{\partial t}(t,z) \Big/ \frac{\partial K}{\partial x_i}(\phi_1(t,z),\ldots,\phi_k(t,z)).$$

Our assumptions about the derivatives of K imply that the right hand side is bounded and continuous. This, along with the boundary conditions, $\phi_i(0,z) = 0$, satisfies standard conditions for a unique solution to exist (Brauer and Nohel 1967). □

If the modelling assumptions are true then the assumptions of theorem 6.3.2 must hold since ϕ_i and its derivatives exist; although this is a rather circular argument as we can never *know* that any modelling assumptions are true with complete certainty.

One would hope that if a pair of (ϕ, K) were found which simultaneously satisfied theorems 6.3.1 and 6.3.2 then this would be unique. Unfortunately this is not the case in general. We show this by considering two possible joint survival functions K and $\widetilde{K}$; next, we define a mapping $\mathbf{f} : \mathbb{R}^k \mapsto \mathbb{R}^k$, which takes the contours of $\widetilde{K}$, $\{\mathbf{x} : \widetilde{K}(\mathbf{x}) = c\}$ to the contours of K, $\{\mathbf{x} : K(\mathbf{x}) = c\}$. With this, we show that $\partial f_i / \partial x_j = 0$, for $i \neq j$, is a necessary and sufficient condition to obtain the general agreement, $K\{\phi_1(t_1, z), \ldots, \phi_k(t_k, z)\} = \widetilde{K}\{\widetilde{\phi}_1(t_1, z), \ldots, \widetilde{\phi}_k(t_k, z)\}$. Finally, we show that the information available, in the form of equation 6.1, does not limit the choice of f to satisfy this orthogonality condition and we provide an explicit counter-example.

Definition 6.3.1. *Given two functions $K, L : \mathbb{R}^k \mapsto \mathbb{R}$, define the set $\mathcal{C}(K, L)$ as:*

$$\mathcal{C}(K, L) = \{\mathbf{f} : K[\mathbf{f}(\mathbf{x})] = L[\mathbf{x}]\},$$

where $\mathbf{f}$ has both domain and range, $\mathbb{R}^k$.

The next theorem shows that if we have two pairs (ϕ, K) and $(\widetilde{\phi}, \widetilde{K})$, which both give the same cause-specific survival functions, then we can choose a member of $\mathcal{C}(K, \widetilde{K})$ which relates ϕ and $\widetilde{\phi}$.

Theorem 6.3.3. *Given two kernels K and $\widetilde{K}$, if there exist ϕ and $\widetilde{\phi}$ such that the resulting cause-specific survival functions are identical (equivalently they satisfy both theorems 6.3.1 and 6.3.2), then there exists $\mathbf{f} \in \mathcal{C}(K, \widetilde{K})$ such that*

$$\mathbf{f}[\widetilde{\mathbf{\Phi}}(t, z)] = \mathbf{\Phi}(t, z), \tag{6.2}$$

where $\mathbf{\Phi}(t, z) = (\phi_1(t, z), \ldots, \phi_k(t, z))$ as defined in theorem 6.3.1 and a similar definition for $\widetilde{\mathbf{\Phi}}$.

Proof. Since $\mathbb{P}(T^{\min} > t|z) = \sum_i F_i(t, z)$ is identified, we must have that

$$K[\mathbf{\Phi}(t, z)] = \widetilde{K}[\widetilde{\mathbf{\Phi}}(t, z)].$$

Given a mapping $\mathbf{f} \in \mathcal{C}(K, \widetilde{K})$, it satisfies $K[\mathbf{f}(\mathbf{x})] = \widetilde{K}[\mathbf{x}]$, this implies that

$$K[\mathbf{\Phi}(t, z)] = K[\mathbf{f}(\mathbf{\Phi}(t, z))].$$

This does not imply equation (6.2) since K is not a bijection, but we can apply any mapping to the image of $\mathbf{f}$ that preserves the contours of K, and we will obtain a new function that is also in $\mathcal{C}(K, \widetilde{K})$. In particular we can find such a mapping that gives equation (6.2). □

We can fix this version of the contour mapping and use it to define a relationship between ϕ and $\widetilde{\phi}$. We can do this without loss of generality since whenever we evaluate $\widetilde{K}[\widetilde{\phi}] = K[\mathbf{f}(\widetilde{\phi})]$, it is invariant to the choice of $\mathbf{f}$ within $\mathcal{C}(K, \widetilde{K})$.

At this point it will be convenient to make some further definitions. First, define the inverse relationship,

$$\widetilde{\mathbf{\Phi}}(t, z) = \mathbf{g}[\mathbf{\Phi}(t, z)]. \tag{6.3}$$

This leads to a relationship between ϕ and $\widetilde{\phi}$, since $\phi_i(t, z) = \mathbf{\Phi}_i(t, z)$, the ith component of $\mathbf{\Phi}$. Hence

$$\widetilde{\phi}_i(t, z) = g_i[\mathbf{\Phi}(t, z)]. \tag{6.4}$$

Second, define the mapping $\Psi : \mathbb{R}_+^k \times \mathbb{R}^l \mapsto \mathbb{R}_+ \times \mathbb{R}^l$ such that $\Psi(\mathbf{t}, x) = (u, z)$ if, and only if, $\boldsymbol{\phi}(\mathbf{t}, x) = (\phi_1(t_1, x), \ldots, \phi_k(t_k, x)) = \mathbf{\Phi}(u, z)$. And define similarly the function $\widetilde{\Psi}$.

Theorem 6.3.4. *The two mappings Ψ and $\widetilde{\Psi}$ coincide if, and only if, $\partial g_i / \partial x_j = 0$, for $i \neq j$.*

Proof. If $\Psi(\mathbf{t}, x) = (u, z)$ then by definition of $\Psi(\mathbf{t}, x)$,

$$\begin{pmatrix} \phi_1(t_1, x) \\ \phi_2(t_2, x) \\ \vdots \\ \phi_k(t_k, x) \end{pmatrix} = \begin{pmatrix} \phi_1(u, z) \\ \phi_2(u, z) \\ \vdots \\ \phi_k(u, z) \end{pmatrix} \tag{6.5}$$

and also if $\widetilde{\Psi}(\mathbf{t}, x) = (u, z)$ then, similarly,

$$\widetilde{\boldsymbol{\phi}}(\mathbf{t}, x) = \widetilde{\boldsymbol{\Phi}}(u, z).$$

Now using equation (6.4) obtains

$$\begin{pmatrix} g_1\{\phi_1(t_1, x), \phi_2(t_1, x), \ldots, \phi_k(t_1, x)\} \\ g_2\{\phi_1(t_2, x), \phi_2(t_2, x), \ldots, \phi_k(t_2, x)\} \\ \vdots \\ g_k\{\phi_1(t_k, x), \phi_2(t_k, x), \ldots, \phi_k(t_k, x)\} \end{pmatrix} = \begin{pmatrix} g_1\{\phi_1(u, z), \phi_2(u, z), \ldots, \phi_k(u, z)\} \\ g_2\{\phi_1(u, z), \phi_2(u, z), \ldots, \phi_k(u, z)\} \\ \vdots \\ g_k\{\phi_1(u, z), \phi_2(u, z), \ldots, \phi_k(u, z)\} \end{pmatrix}. \tag{6.6}$$

But, substituting the left side of 6.5 into the right side of 6.6 we see that,

$$g_i\{\phi_1(t_i, x), \phi_2(t_i, x), \ldots, \phi_k(t_i, x)\} = g_i\{\phi_1(t_1, x), \phi_2(t_2, x), \ldots, \phi_k(t_k, x)\},$$

for all i. Ignoring the ϕs and focusing on the gs, we see this is saying that,

$$g_i(t, t, \ldots, t) = g_i(a, b, c, \ldots, t, \ldots, x, y, z),$$

where the t is in the ith position on the right hand side. So the value of g_i only depends upon its ith argument. Assuming that $\mathbf{g}$ has derivatives, this can hold, in general, if, and only if, $\partial g_i / \partial x_j = 0$ for $i \neq j$. □

Now under our assumptions $S(t_1, \ldots, t_k|x) = \mathbb{P}(T^{\mathsf{min}} > u|z)$ where $\Psi(\mathbf{t}, x) = (u, z)$. Hence we need Ψ and $\widetilde{\Psi}$ to coincide otherwise the two models will give different probabilities to events such as $\{T_1 > t_1, \ldots, T_k > t_k|x\}$.

To recapitulate, it is assumed that we have two pairs (ϕ, K) and $(\widetilde{\phi}, \widetilde{K})$, which give identical cause-specific survival functions. Theorem 6.3.3 shows that there must exist a function $\mathbf{f}$ such that $K[\mathbf{f}(\mathbf{x})] = \widetilde{K}[\mathbf{x}]$, and that relates the covariate-time transformations $\mathbf{\Phi}(t, z) = \mathbf{f}(\widetilde{\mathbf{\Phi}}(t, z))$. Theorem 6.3.4 shows that to get agreement for all values of $(\mathbf{t}, z)$—effectively identifiability—a necessary and sufficient condition is that this $\mathbf{f}$, or rather its inverse $\mathbf{g}$, has to have a diagonal derivative matrix.

The next theorem shows that in general there exist multiple pairs (ϕ, K) which give the same cause-specific survival functions, but do not satisfy the conditions of theorems 6.3.3 and 6.3.4.

Theorem 6.3.5. *Given a pair, (ϕ, K) which is consistent with theorems 6.3.2 and 6.3.1, there exists a mapping $\mathbf{g} : \mathbb{R}_+^k \mapsto \mathbb{R}_+^k$ which defines a new pair $(\widetilde{\phi}, \widetilde{K})$, by means of theorem 6.3.3 and equations (6.2) and (6.3), and which does not satisfy $\partial g_i/\partial x_j = 0$ for $i \neq j$.*

Proof. Starting with the ith component of equation (6.3), and taking its derivative with respect to time,

$$\begin{aligned}
\frac{\partial \widetilde{\phi}_i}{\partial t}(t, z) &= \sum_j \frac{\partial \phi_j}{\partial t}(t, z) \frac{\partial g_i}{\partial x_j}\big(\mathbf{\Phi}(t, z)\big) \\
&= \frac{\partial \mathbf{\Phi}}{\partial t}(t, z) \wedge [D]^i\big(\mathbf{\Phi}(t, z)\big), \\
&= \left[\frac{\partial \mathbf{\Phi}}{\partial t}(t, z) D\big(\mathbf{\Phi}(t, z)\big)\right]^i, \qquad (6.7)
\end{aligned}$$

where $D(\cdot)$ is the matrix whose jth row, ith column, is defined to be $\partial g_i/\partial x_j$, and $[D]^i$ denotes the ith column of this matrix. Here $\wedge$ denotes the standard inner product.

Similarly taking the definition of $\widetilde{K}$ and taking the derivative with respect to x_i obtains,

$$\frac{\partial \widetilde{K}}{\partial x_i}\left(\widetilde{\mathbf{\Phi}}(t,z)\right) = \sum_k \frac{\partial f_k}{\partial x_i}\left(\widetilde{\mathbf{\Phi}}(t,z)\right)\frac{\partial K}{\partial x_k}\left(\mathbf{f}\left[\widetilde{\mathbf{\Phi}}(t,z)\right]\right)$$

by (6.2)

$$= \sum_k \frac{\partial f_k}{\partial x_i}\left(\widetilde{\mathbf{\Phi}}(t,z)\right)\frac{\partial K}{\partial x_k}\Big(\mathbf{\Phi}(t,z)\Big)$$

since $\mathbf{f} = \mathbf{g}^{-1}$

$$\begin{aligned} &= \left[D^{-1}\right]_i\big(\mathbf{\Phi}(t,z)\big) \wedge \nabla K\big(\mathbf{\Phi}(t,z)\big), \\ &= \Big[D^{-1}\big(\mathbf{\Phi}(t,z)\big)\nabla K\big(\mathbf{\Phi}(t,z)\big)\Big]_i \ , \end{aligned} \tag{6.8}$$

where $[D^{-1}]_i$ denotes the ith row of the inverse of the matrix D. Here ∇ denotes the standard gradient operator.

Substituting equations (6.7) and (6.8) into equation (6.1) we see, that

$$\left[\frac{\partial \mathbf{\Phi}}{\partial t}(t,z)\right]^i\Big[\nabla K\big(\mathbf{\Phi}(t,z)\big)\Big]_i = \left[\frac{\partial \mathbf{\Phi}}{\partial t}(t,z)D\big(\mathbf{\Phi}(t,z)\big)\right]^i\Big[D^{-1}\big(\mathbf{\Phi}(t,z)\big)\nabla K\big(\mathbf{\Phi}(t,z)\big)\Big]_i \tag{6.9}$$

Now if we regard $\partial\mathbf{\Phi}/\partial t$ and ∇K as two arbitrary, fixed, vectors it can be seen that equation (6.9) is a set of k equations on the k^2 elements of matrix D. In general we can find an infinite number of solutions which do not have D as a diagonal matrix. □

As a solid example it can be verified that the matrix

$$D(\mathbf{\Phi}) = \begin{pmatrix} 1 & a & & & \\ 1 & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix},$$

where

$$a = \frac{\partial \phi_2}{\partial t}\frac{\partial K(\mathbf{\Phi})}{\partial x_1} \Big/ \left\{ \frac{\partial \phi_1}{\partial t}\frac{\partial K(\mathbf{\Phi})}{\partial x_2} + \frac{\partial \phi_2}{\partial t}\frac{\partial K(\mathbf{\Phi})}{\partial x_2} - \frac{\partial \phi_1}{\partial t}\frac{\partial K(\mathbf{\Phi})}{\partial x_1} \right\},$$

and all the remaining elements are zero, satisfies equation (6.9).

6.4 Summary

In this chapter there are some results of limited use, which show that with less assumptions about the form of the covariate-time transformation we lose identifiability without extra information about the kernel joint survival. Hopefully this sheds some extra light on what is the absolute weakest set of assumptions which give identifiability in the competing risks setting. At the moment, for practical purposes, the weakest assumption is one of proportional hazards with a frailty distribution inducing a dependency. A topic for future research would be to understand why, and if, this is indeed the weakest assumption and to discover if there were any different, but 'equally weak' assumptions.

Chapter 7

Frailty modeling

7.1 Constituent Theory

This chapter is based on the binding together of three ideas. The first idea is the theorem proved in Heckman and Honoré (1989) where it is shown that under the assumption of proportional hazards, the addition of covariate information permits the identifiability of the joint survival function in a competing risks model. The second idea, or rather a large body of contemporary statistical research, takes the very general modeling assumption of Heckman and Honoré (1989), namely that

$$S(t_1, \ldots, t_p|x) = K[\exp\{-\Lambda_1(t_1)\phi_1(x)\}, \ldots, \exp\{-\Lambda_p(t_p)\phi_p(x)\}],$$

and refines it to the specific assumption that the $K(\cdot)$ represents the operation of marginalisation with respect to some unobserved covariates, upon which the different causes would be conditionally independent: this is commonly referred to as frailty modeling. The third idea is the practical device in Lunn and McNeil (1995) which enables, in theory, the use of existing frailty software to fit non-independent competing-risks data with covariates: explicitly, if an individual is at risk from p causes of failure, and

we observe a failure at time t from cause k, say, then this can be represented by p individuals, with identical values for time and all the covariates, with the addition of a further covariate, CAUSE, which systematically takes different values from $\{1, 2, \ldots, p\}$ corresponding to the causes of failure, and all individuals are recorded as being censored with the exception of the 'replicate' with CAUSE $= k$.

7.1.1 Identifiability Theorem of Heckman and Honoré

Their result is phrased in terms of only two competing risks, but it is clear that it generalises to p causes in a trivial fashion. The basic assumption is that the joint survival function of two latent variables t_1, t_2 with covariates x, has the form,

$$S(t_1, t_2|x) = K[\exp\{-\Lambda_1(t_1)\phi_1(x)\}, \exp\{-\Lambda_2(t_2)\phi_2(x)\}],$$

where $K(\cdot, \cdot)$ is a continuously differentiable, non-negative function $[0,1]\times[0,1] \mapsto [0,1]$. They assume some normalisations: $\Lambda_i(1) = 1, \phi_i(x_0) = 1,\ i = 1, 2$, for some fixed x_0. It is taken that the data provide an estimate of the cause-specific survival,

$$\mathbb{P}(T_i > t, T_i = \min\{T_j\}|x) = F_i(t; x).$$

The normalisations are not an important imposition, since the same joint survival function is obtained whenever Λ is divided by a constant and ϕ is multiplied by the same constant, hence Λ_1 can be normalised; the normalisation, $\phi_i(x_0) = 1$ can be achieved by a rescaling of ϕ_1, by c say, which can be accommodated by defining a new $\widetilde{K}(\eta_1, \eta_2) = K(\eta_1^{-c}, \eta_2)$.

They go on to prove the identifiability of the cumulative cause specific hazards, Λ_i, the covariate functions, ϕ_i, and K, and hence the joint survival function. An outline proof is to take the ratio of $F_i'(t; x)$ at an arbitrary $x \neq x_0$ to $F_i'(t; x_0)$, where x_0 is the reference level. This is equal to the ratio of the first derivatives of K with respect

to the latent time, t_i. Taking the limit as $t \to 0$, we get $\phi_i(x)$. Next, setting $t = 1$ so that $\Lambda_i(1) = 1$, and letting the ϕ_i range over their support we identify K. Finally, to identify Λ_1, say, we find a value of x such that $\phi_1(x)(\neq 0)$ is fixed but $\phi_2(x) = 0$. From this we get $K = f(\Lambda_1(t)\phi_1(x))$, which can be inverted to find Λ_1.

Unfortunately, this merely proves the identifiability of the joint survival function, it does not provide a practical means to estimate the effect of the covariates, since $\phi(x)$ is shown to be equal to the limit, *as t tends to zero*, of some quantity and hence, in the proof, ignores most of the information in the data set. Also, the fact that we need ϕ to take values over its entire range—the positive reals—suggests that a lot of information needs to be provided to build up a reliable picture of the function K.

7.1.2 The Frailty Model

This area is most easily approached as a specific application of generalised linear mixed models, which are a generalisation of generalised linear models (G.L.M.s). G.L.M.s provide a very useful framework that assumes the response variable, y, comes from a two-parameter exponential family, and relates the expected value of y to a linear function of the covariates, $\eta(X) = X\beta$, using a specified link function, $g(\cdot)$, where $g(\mathbb{E}(y)) = \eta(X)$. A comprehensive summary of the theory behind classical G.L.M.M.s is provided in McCulloch and Searle (2001).

All proportional-hazards survival models can be represented within this framework, where each individual is represented by repeated observations of the random variable $N(t) = I(t >$failure time), and $Y(t) = I(t <$event time). This is referred to as a *counting process*. The random process $N(t)$ starts at zero and jumps to one at an observed failure time, or stays at zero if the individual is censored. The predictable or left continuous, process $Y(t)$ indicates whether or not the individual is still at risk. When

$dN(t)$ is used as the response variable it is assumed to follow a Poisson distribution with mean $Y(t)\lambda(t)$ where $\lambda(t)$ is defined as the hazard function. It can be seen that the likelihood for the counting process coincides with the conventional likelihood, since (in a very heuristic fashion)

$$\mathbb{P}(N(t), Y(t)) = \prod_{0<t} (Y(t)\lambda(t))^{dN(t)} \exp(-Y(t)\lambda(t)) = \lambda(T)^{\delta} \exp(-\int_0^T \lambda(t)dt),$$

where (T, δ) represent the traditional (time, status) way of representing the data. The abuse of notation in taking the product over a dense set is explained in detail in Andersen et al. (1993).

Within the G.L.M. framework, the canonical link function for a Poisson distribution is the log-link. This ties in with the proportional hazards models since the hazard function for an individual with covariates x is assumed to be $\lambda(t, x) = \exp(\beta x)\lambda_0(t)$ for some baseline hazard function $\lambda_0(t)$, so when we take the logarithm of the Poison mean we get $\log[\mathbb{E}\{dN(t)\}] = \beta x + \log \lambda_0(t) = \eta$, a linear function of the covariates—as required by the G.L.M. framework. When the baseline hazard function is unknown, as in Cox's proportional hazards (Cox 1972), the term $\log \lambda_0(t)$ is left as a piecewise constant function on the intervals $[t_{(i)}, t_{(i+1)})$, with values to be estimated, effectively a factor. If the baseline hazard is assumed to take a parametric form then it is fitted as an offset term.

The next step is the introduction of random effects to the linear predictor,

$$\eta_{ij} = x_{ij}\beta + z_{ij}b_j,$$

where b_j is an *unobserved* continuous random variable common to the *cluster* indexed by j. Typically there are some restrictions on the distribution of $b_j$to enable identifiability: conventionally, $\mathbb{E}(b) = 0$ or $\mathbb{P}(b < 0) = 1/2$. Index i refers to individuals within these clusters and it is assumed that the $\log \lambda_0(t)$ term is absorbed into $x\beta$. The likelihood

for the observed data is then a function of the likelihood conditional on the b_j. The conditional likelihood can be represented by the likelihood for the conventional regression model which temporarily pretends the b_j are known. To obtain the likelihood, this function is then integrated with respect to the distribution of the random effects. The integration induces a dependency between individuals within a cluster which makes it a suitable framework for dealing with non-independent competing risks.

Aside: the well known non-identifiability problem of competing risks coincides exactly with the non-identifiability problem in G.L.M.M. when there are no-replicates—an individual cannot fail twice—and a lack of covariates.

The frailty model also has the convenient interpretation that the random effects represent some covariates which cannot be observed, although the assumption that these follow a normal distribution is questionable. There is also the restriction that, with a univariate random effect, the correlation which is induced must be positive. However in the case of multivariate random effects the covariance structure can be arbitrarily specified, thus opening up possibilities of an autoregressive structure with negative correlation or something even more exotic. However, the main problem with such a structure comes from the likelihood being in the form of an integral which, in general, has to be numerical evaluated over a multi-dimensional space, thus making it rather difficult to maximise accurately. The practicalities of such estimation will be considered later.

7.2 Penalised Quasi-Likelihood Estimation

In this section the problem of how to maximise a multi-dimensional integral will be addressed. The principal line of attack, at the present time, is to use a Laplace approximation to the integral which subsequently allows Newton-Raphson Schemes, or Fisher

Scoring, to maximise this *quasi*-likelihood numerically. Somewhat unsatisfactorily, an alternative perspective is to consider the quasi-likelihood as an ad hoc starting point and subsequently to analyse the properties of the resulting estimators.

We consider the log-likelihood of the data, y, conditional on the values of the random effects, b, and covariates, x, and denote this as $l_1(y; x, b)$. Note that there is a conflict of standard notation in the meaning of y or Y; for this section y denotes the response variable in a generalised linear mixed model and not the indicator variable of section 7.1.2. If we make the further assumption that the p random effects, which are unobserved, follow a p-dimensional multivariate normal distribution, $N[0, \mathbf{D}(\theta)]$, where θ represents a parameterisation of the covariance matrix, then the likelihood for the observed data is,

$$L(\theta, \beta) \propto |\mathbf{D}|^{-1/2} \int_{\mathcal{B}} \exp\left[l_1(b, \dots) - \frac{1}{2} b' \mathbf{D}^{-1} b\right] db.$$

Now, taking a second order Taylor expansion of the logarithm of the integrand about b_0 we get

$$\begin{aligned} L(\theta, \beta) \propto |\mathbf{D}|^{-1/2} \exp\left(l(b_0, \dots) - \frac{1}{2} b_0' \mathbf{D}^{-1} b_0\right) \\ \times \int_{\mathcal{B}} \exp\Big[\left(l'(b_0, \dots) + \mathbf{D}^{-1} b_0\right)^t (b - b_0) \\ + (b - b_0)^t \left(l''(b_0, \dots) + \mathbf{D}^{-1}\right) (b - b_0) + o(b^2)\Big] db. \end{aligned}$$

Now if b_0 is chosen to satisfy

$$l'(b_0, \dots) + \mathbf{D}^{-1} b_0 = 0,$$

which can be found by considering b to be a fixed effect coefficient in a standard G.L.M. framework, then this leads to an integral with a known value: the normalising constant of the multivariate normal distribution with inverse covariance, $-l''(b_0, \dots) - \mathbf{D}^{-1}$.

Using the exponential-family form of the conditional likelihood, it can be shown that the logarithm of this approximation reduces to the quasi log-likelihood which we define to be,

$$ql(\beta,\theta) = -\frac{1}{2}\log|\mathbf{I} + \mathbf{Z^t W Z D}| - l(y;x,b_0) - \frac{1}{2}b_0^t\mathbf{D}^{-1}b_0,$$

where $\mathbf{W}$ is the standard diagonal matrix of weights with $W_{ii} = 1/\text{Var}\{Y\}\, g^2\{\mathbb{E}(Y)\}$, from G.L.M. theory as summarised in chapter 5 of McCulloch and Searle (2001). For the purposes of maximisation it is assumed that the first term changes very slowly with values of the mean, $g^{-1}(X\beta + Zb)$, and hence it is ignored.

The basis of the algorithm proposed in Breslow and Clayton (1993), and derived by alternative means in Schall (1991) and Wolfinger (1993), is to iterate between maximising the quasi-log likelihood in terms of the coefficients (which as a side-effect gives a prediction of the random effects), and maximising in terms of θ, the parameterisation of the random effects variance, where at each stage it is assumed that the other parameters are fixed. To maximise with respect to the coefficients, we can use Fisher-scoring, where we define the *working vector*,

$$\widetilde{y} = \eta(x,b) + (y - \mathbb{E}[y|x,b])\, g'\left(\mathbb{E}[y|x,b]\right),$$

the first order Taylor expansion of $g(y)$ about $\mathbb{E}(y)$, and then iteratively solve

$$\begin{bmatrix} \mathbf{X^t W X} & \mathbf{X^t W Z} \\ \mathbf{Z^t W X} & \mathbf{Z^t W Z} + \mathbf{D}^{-1} \end{bmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{bmatrix} \mathbf{X^t W}\widetilde{y} \\ \mathbf{Z^t W}\widetilde{y} \end{bmatrix},$$

which has the appealing interpretation of transforming the response variable to a scale where least-squares estimation can be used. To estimate θ, the following equation has to be solved:

$$-1/2\left[(\widetilde{y} - \mathbf{X}\beta)^t\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{V}^{-1}(\widetilde{y} - \mathbf{X}\beta) - \text{trace}\left(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta_j}\right)\right] = 0,$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}$$

and $\mathbf{V} = \mathbf{W}^{-1} + \mathbf{ZDZ^t}$, the variance of the working vector $\widetilde{y}$. In the simple case where $\mathbf{D} = \theta\mathbf{I}$, independent random effects with a common variance, solving this equation reduces to iterating,

$$\widehat{\theta} = b^t b/(p-r),\ r = \theta^{-1}\text{trace}(\{\mathbf{Z^tWZ} + \mathbf{D}^{-1}\}^{-1}).$$

7.3 Partial Likelihood

Having justified how parametric and semi-parametric proportional-hazards survival models can be fitted within the G.L.M. framework, and hence can utilise the generalised framework of G.L.M.M.s for frailty models, it is appropriate to point out some problems. The main problem is that, within the unadulterated G.L.M. representation, the semi-parametric version requires a parameter to be estimated at each unique non-censored failure time. This translates into a large parameter space which increases in dimension at a rate proportional to the sample size. This has computational implications in that, regardless of the approach taken to estimation, the inverse of a large matrix will need to be repeatedly calculated. On a statistical level, the point of using a semi-parametric model is to be able to focus attention on the effect of the covariates without worrying about the baseline hazard. Fortunately there is a way to avoid this problem: the partial likelihood (Cox 1975).

If we take the expression below as a starting point for the likelihood as a function both of $\lambda_0(t)$, the baseline hazard, and η, the linear predictor,

$$L(\eta_i, \lambda_0(t_i); t_i, \delta_i) = \prod_i \left(\exp(\eta_i)\frac{\Lambda_{0i}}{t_i - t_{i-1}}\right)^{\delta_i} \exp\left(-e^{\eta_i}\sum_{j\leq i}\Lambda_{0j}\right),$$

where it is assumed that the failure times have been ordered so that, $i < j$ if, and only if, $t_i < t_j$, that $t_0 = 0$ and that $\Lambda_{0i} = \int_{t_{i-1}}^{t_i} \lambda_0(s)ds$. Hence $\Lambda_{0i}/(t_i - t_{i-1})$ will approximate the hazard function, $\lambda_0(t_i)$, as the time increments become smaller. If we assume that the coefficients, and the random effects, are known then if we take the derivative of the profile log-likelihood with respect to Λ_{0i}, we obtain

$$\delta_i/\Lambda_{0i} - \sum_{j \geq i} \exp(\eta_j),$$

which is equal to zero when

$$\Lambda_{0i} = \delta_i / \sum_{j=i}^{n} \exp(\eta_j).$$

If this is substituted into the full log-likelihood, then we obtain

$$\sum_i \left[\delta_i \eta_i - \delta_i \log \left(\sum_{j \geq i} \exp(\eta_j) \right) - \delta_i \log(t_i - t_{i-1}) - \exp(\eta_i) \sum_{j \leq i} \left(\delta_j / \sum_{k \geq j} \exp(\eta_k) \right) \right].$$

However it can be shown that the final term, upon changing the order of summation $(\sum_i \sum_{j \leq i} = \sum_j \sum_{i \geq j})$, equals the constant, $\sum_j \delta_j$, and can be ignored as can the constant term $\sum_i \delta_i \log(t_i - t_{i-1})$.

Hence we have shown that the likelihood is maximised, in term of the coefficients, by the maximiser of

$$\sum_i \left[\delta_i \eta_i - \log \left(\sum_{j \geq i} \exp(\eta_j) \right) \right],$$

which does not involve the baseline hazard. Given that the methods for estimating G.L.M.M.s uses the Laplace approximation to obtain the quasi-log-likelihood,

$$l(y; x, b_0) - \frac{1}{2} b_0^t \mathbf{D}^{-1} b_0,$$

it is clear that substituting the partial log-likelihood for l will give the same estimate for β, without needing to estimate the baseline hazard.

This can be derived in another more subtle fashion, by observing that

$$\mathbb{P}(\text{ individual } i \text{ fails at } t_i | \text{ individuals } i, i+1, \ldots, n \text{ are still at risk})$$
$$= \exp(\eta_i)\lambda_0(t_i) \Big/ \sum_{j \geq i} \exp(\eta_j)\lambda_0(t_i),$$

observing that the hazard functions cancel out reduces the expression to the partial-likelihood function defined above. Using this conditional probability, rather than the full density, is justified on the general grounds that if you have two sets of parameters ϕ, ψ such that the full likelihood $L(\phi, \psi|x) = L(\phi|\psi, x)L(\psi|x)$, where ϕ is a nuisance parameter, then it is valid to ignore the first factor and simply maximise the second factor, which does not depend on the nuisance parameters. The standard properties of the likelihood transfer to the partial likelihood, although there is the possibility of losing some inferential power if the two parameter sets are not orthogonal (Barndorff-Nielsen and Cox 1994, Cox 1975).

7.4 Data Editing

The commonest format, due to its conciseness, for competing risks data to be presented is for each individual to have a failure time $T \in \mathbb{R}^+$, a cause $C \in \{0, 1, \ldots, c\}$, where $C = 0$ represents a censored individual, along with a vector of explanatory variables Z. However, if we are using the concept of latent failure times then this really represents a set of c failure times $\{T_1, \ldots, T_c\}$, which all take the value T but are censored, with the one exception of T_C which is observed if $C \neq 0$, or is also censored if $C = 0$. In the counting process framework we have that, instead of a single $N_i(t)$ for individual i, there are c such processes, $N_{ij}(t)(j = 1, \ldots, c)$, along with the variables $Y_{ij}(t)$ (shifting back to the section 7.1.2 definition of Y) which do not vary across j—this is essentially the key distinguishing aspect of competing-risks survival analysis, as opposed to multivariate

survival analysis.

So if given a data set in the conventional format with rows corresponding to individuals, where row i is (T_i, C_i, Z_i), this should be converted to the matrix with c rows, where row j $(1 \leq j \leq c)$ consists of $(T_i, \delta_{ij}, j, Z_i)$ where $\delta_{ij} = I(C = j)$. The only remaining point is to explain how to use the new explanatory variables (j, Z) to represent the desired dependence structure of the model. If we assume that the covariates Z_i for individual i are explicitly a set of univariate random variables, $X_1, \ldots, X_p$, where a discrete variable, also known as a factor, with state space $\{1, \ldots, k\}$ is represented as a set of $(k-1)$ binary variables and any potential interactions are represented, then we need to create the correct design matrix. If the desired model is for the hazard function $\lambda_{ij}(t; Z_i)$, for individual i, cause j, to be of the form

$$\lambda_{ij}(t; Z_i) = \exp(X_{n_1}\beta_{1j} + X_{n_2}\beta_{2j} + \ldots + X_{n_q}\beta_{qj})h_{0j}(t)$$

for some subsequence $(n_1, \ldots, n_q)$ of $1, \ldots, p$, and a baseline hazard function $\lambda_{0j}(t)$ which varies between causes and if (j, Z), the data values, are labeled as

$$(CAUSE, X_1, \ldots, X_p)$$

then, using the notation employed in S-Plus (Wilkinson and Rogers 1973, Becker, Chambers and Wilks 1988, Chambers and Hastie 1992), we want the design matrix corresponding to

$$\text{CAUSE} : (X_{n_1} + \ldots + X_{n_q}) + \text{strata(CAUSE)}.$$

If we prefer that the values of $\beta_{n_1 j}, \ldots, \beta_{n_r j}$ do not change between causes of failure, then we need

$$X_{n_1} + \ldots + X_{n_r} + \text{CAUSE} : (X_{n_{r+1}} + \ldots + X_{n_q}) + \text{strata(CAUSE)}.$$

On the other hand, if we believe that $\lambda_{0j}(t) = w_j\lambda_0(t)$ so that the cause-specific baseline hazard functions are proportional to one another, then we need

$$\text{CAUSE}\backslash(X_{n_1} + \ldots + X_{n_q}) \textit{ or } (X_{n_1} + \ldots + X_{n_q}) \text{ \%in\% CAUSE},$$

and similarly if all the coefficients are constant between causes, and the cause-specific hazards are proportional then we have,

$$\text{CAUSE} + X_{n_1} + \ldots + X_{n_q}.$$

7.5 Practical Computing Issues

In summary, to fit a competing-risks survival analysis model, which represents any dependencies between the latent survival times by a random effects distribution, and represents the influence of covariates through a proportional hazards model we first edit the data as described in section 7.4 and obtain the appropriate design matrix. Then the relevant parameters, namely the fixed effect coefficients, the variance of the random effects, and the baseline hazard function, are estimated by maximising the penalised partial likelihood function as defined in section 7.3. This maximisation is performed using the algorithm described in section 7.2, which consists of estimating the fixed effects assuming that the random effects variance is known, and then estimating the variance assuming the fixed effects are known.

This iterative scheme, in algebraic terms, corresponds to defining a sequence $\sigma_{n+1} = g(\sigma_n)$, and calculating the limiting value. Geometrically this is shown in figure 7.1.

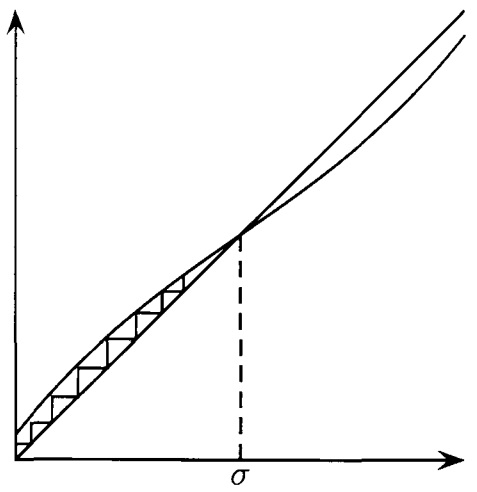


Figure 7.1: Geometric version of the algorithm

However, if the lines $y = x$ and $y = g(x)$ lie close together then it is clear that the sequence will converge at a slow rate. It is apparent that, due to the high proportion of censoring intrinsic in fitting a competing risks frailty model, the penalised quasi likelihood is rather flat, and hence we find ourselves in the slow convergence situation. In practice with a simulated data set of size 100, the number of iterations is in excess of a 100. An alternative is to use an interval bisection algorithm. The basic idea here is to have some means of determining whether or not σ lies in an interval (a, b); the first step is to start with a wide interval (a_0, b_0) which contains σ, then consider the two intervals $(a, [a+b]/2)$, and $([a+b]/2, b)$, clearly σ will only lie in only one of these intervals and we are able to determine which one; now we repeat this where at each stage the interval is bisected, and hence any desired margin of error can be achieved in a finite number of steps. This is shown in figure 7.2.

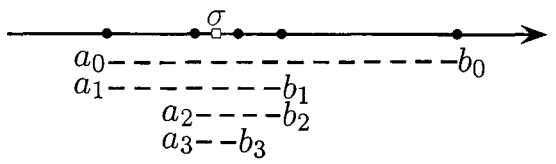


Figure 7.2: Geometric representation of the bisection algorithm

It is clear that the interval bisection algorithm hinges on the ability to determine whether the root of the equation lies within a specified interval. The proposed method of achieving this is based around performing one step of the previous, slow, algorithm and examining the new estimate to see if it is larger or smaller than the previous estimate. It is claimed that if the new estimate is larger, say, then the root of the equation is also larger than the previous estimate. Hence if the lower bound of an interval produces an increased estimate, and the upper bound produces a decreased estimate then the root lies within the interval. Theorem 7.5.1 provides necessary conditions, namely a region where the Hessian matrix is negative definite, for the proposed method to work.

Theorem 7.5.1. *Given a well behaved log-likelihood function,* $l(\alpha, \beta) : \mathbb{R}^p \times \mathbb{R} \mapsto \mathbb{R}$, *which has a negative definite Hessian matrix,* H, *if we define the functions*

$$f : \mathbb{R} \mapsto \mathbb{R}^p$$

such that

$$\frac{\partial l}{\partial \alpha_i}(f(\beta), \beta) = 0, \ i = 1, \ldots, p \tag{7.1}$$

and

$$g : \mathbb{R} \mapsto \mathbb{R}$$

such that

$$\frac{\partial l}{\partial \beta}(f(\beta), g(\beta)) = 0, \tag{7.2}$$

and define the vector, (α_0, β_0) *to satisfy*

$$\beta_0 = g(\beta_0),\ \alpha_0 = f(\beta_0),$$

then

$(g(\beta) - \beta)(\beta_0 - \beta) > 0$ *if, and only if,*

$$\sum_{i,j} f_i' \frac{\partial^2 l}{\partial \alpha_i \partial \alpha_j} f_j' > \frac{\partial^2 l}{\partial \beta^2}.$$

Proof. Consider the function,

$$h(\beta) = g(\beta) - \beta.$$

By the Mean Value Theorem and assuming that g, and therefore h, is continuous, there exists $\beta' \in [\beta, \beta_0]$ such that

$$h'(\beta') = \frac{h(\beta) - h(\beta_0)}{\beta - \beta_0} = \frac{g(\beta) - \beta}{\beta - \beta_0},$$

hence $h'(\cdot) < 0$, or equivalently $g'(\cdot) < 1$, is a necessary and sufficient condition.

Taking the derivative with respect to β of (7.2) we obtain,

$$\begin{aligned} \sum_{i=1}^{p} f_i' \frac{\partial^2 l}{\partial \alpha_i \partial \beta} + \frac{\partial^2 l}{\partial \beta^2} g'(\beta) &= 0 \\ \Rightarrow -\left(\sum_{i=1}^{p} f_i' \frac{\partial^2 l}{\partial \alpha_i \partial \beta} \right) \Big/ \frac{\partial^2 l}{\partial \beta^2} &= g'(\beta) \end{aligned} \tag{7.3}$$

Similarly taking the derivative with respect to β of (7.1) we get,

$$\begin{aligned} \sum_{j=1}^{p} \frac{\partial^2 l}{\partial \alpha_i \partial \alpha_j} f_j' + \frac{\partial^2 l}{\partial \alpha_i \partial \beta} &= 0,\ i = 1, \ldots, p \\ \Rightarrow -\sum_{j=1}^{p} \frac{\partial^2 l}{\partial \alpha_i \partial \alpha_j} f_j' &= \frac{\partial^2 l}{\partial \alpha_i \partial \beta} \end{aligned} \tag{7.4}$$

Hence substituting (7.4) in the left hand side of (7.3) we get,

$$\begin{aligned} g'(\beta) &= \left(\sum_{i,j} f_i' \frac{\partial^2 l}{\partial \alpha_i \partial \alpha_j} f_j' \right) \Big/ \frac{\partial^2 l}{\partial \beta^2} \\ &= (a^t H a)/(b^t H b) < 1, \end{aligned}$$

where $a = (f_1', \ldots, f_p', 0)$ and $b = (0, \ldots, 0, 1)$. By assumption, $x^t H x < 0$ for all $x \in \mathbb{R}^{p+1}$, hence the condition is equivalent to

$$\sum_{i,j} f_i' \frac{\partial^2 l}{\partial \alpha_i \partial \alpha_j} f_j' > \frac{\partial^2 l}{\partial \beta^2}. \tag{7.5}$$

□

So in the notation of the proof $\sigma^2 = \beta$, the variance of the frailties, and α is the vector of coefficients. Equation (7.1) represents the step of holding the variance fixed and finding the maximising set of coefficients: the mapping f. Equation (7.2) takes these new coefficients, $f(\beta)$, assumes they are fixed, and maximises with respect to the variance the mapping g. The end result, that $(g(\beta) - \beta)(\beta - \beta_0) < 0$, implies that if g increases β, then β is less than the converged value β_0, whereas if g decreases β then the β_0 is less than β.

This result is of limited use as, clearly, it hinges upon the model and which starting values are used as to whether the condition is satisfied or not. The condition could be checked, with a numerical approximation to f_i', at each step, but it is not clear what to do if the condition fails. One observation is that, considering equation (7.4), if the parameters, α and β are orthogonal in the sense that $\partial l^2 / \partial \alpha \partial \beta = 0$, then the left hand side of (7.5) is zero and the inequality is satisfied.

When used on the data set considered in chapter 9 the estimated variance of the random effect, using the algorithm proposed here, was 0.560 with a standard deviation of 0.127. This is reasonably close to the estimate obtained from the S-Plus frailty software

(Therneau and Grambsch 2000) which gave an estimate of 0.527 with a p-value of 0.062 for the hypothesis that the variance is zero. A comparison of the fixed effects coefficients are plotted in figure 7.3.

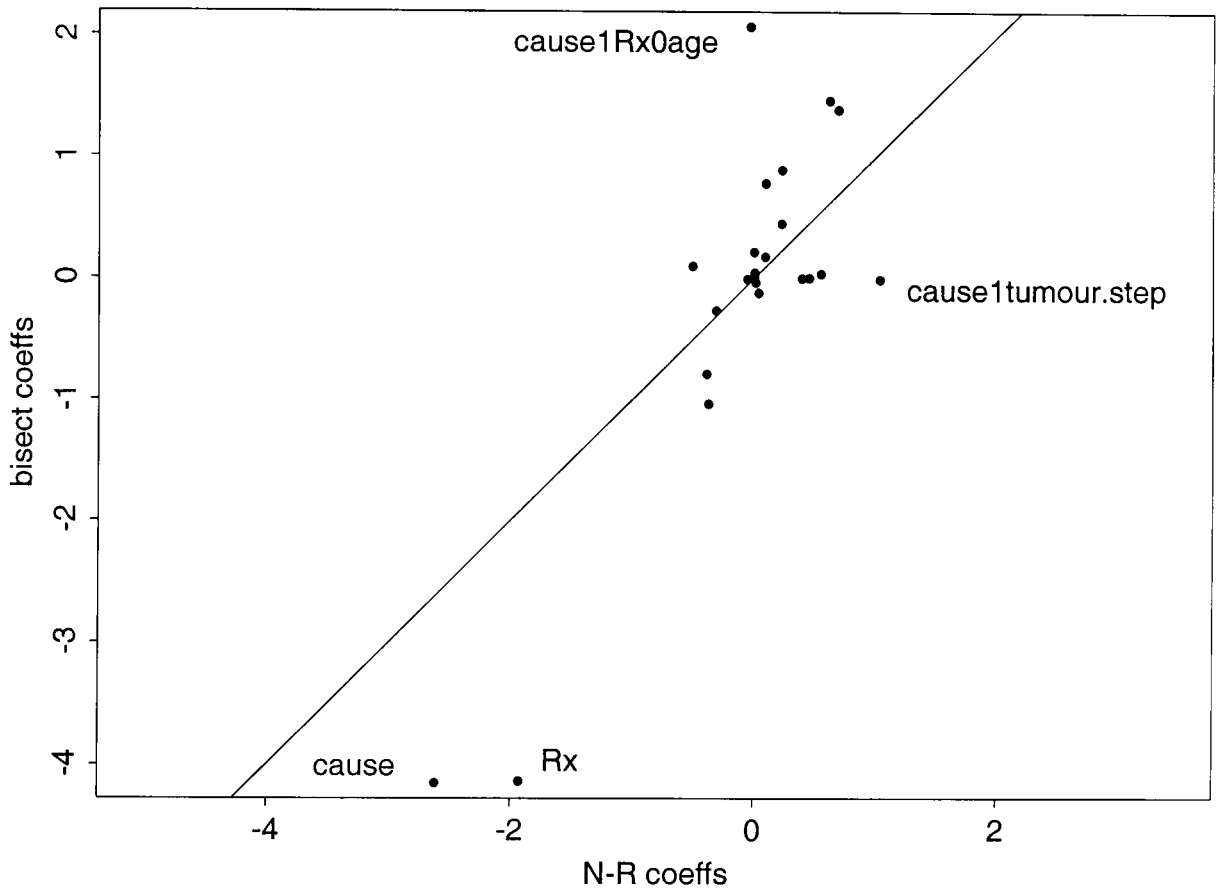


Figure 7.3: Scatter plot of two sets of fixed effects coefficients

As shown in chapter 9 the variance of the frailty distribution is sensitive to the choice of frailty distribution and the discrepancy between these two estimates is acceptable given the effect of changing the assumed frailty distribution. Also given the variety of ad hoc approximations which are proposed in the literature, as regards calculating a likelihood, this method may have a suitable role in practical model selection

given its speed.

There are other alternative approaches to the interval bisection algorithm, such as the secant method or the Fibonacci search method, however, limited practical experience indicates that any improvements in computation time are very modest and that the guaranteed convergence of the bisection method can be lost with the secant method.

7.6 Current Software

Although the proposed alteration to the established algorithm does decrease the computation time when both algorithms have been written in S code by the author, the optimal method in terms of absolute computation time, presumably achieved through optimal use of C code, is to use the frailty software of Therneau and Grambsch (2000) which is readily available within S-plus. To use this should be as simple as adding a term such as

```
frailty( clusterindex )
```

to the desired formula as discussed in section 7.4. However there are, at present, several bugs in the software.

First, the software defaults to a slightly dated method of forming the relevant design matrices when a `frailty` term is in the formula. This forms an incorrect design matrix when there are interactions, such as CAUSE/$\{\dots\}$, which is a matrix of sub-column-rank and thus the matrix inversion steps encounter a singular matrix and fail. This has been corrected using the standard code from the `coxph` software which is in use when there are no frailty terms. The altered software is in appendix B

Secondly, it does not respond correctly when the variable which indexes the frailty clusters is anything other than the simple sequence $1, 2, \dots, m$, where m is the

number of clusters. In particular, even when a cluster should be excluded due to missing values, the software still attempts to estimate and integrate out a frailty b_j. The practical solution is to settle upon a cleaned data set with no missing values and only then index the frailty clusters.

Another criticism is that it has proved practically impossible to obtain a value that corresponds to the integrated full likelihood when the fraily distribution is assumed to be normal. This would be useful in obtaining profile likelihood functions so as to calculate confidence intervals for the variance of the frailty distribution.

On the positive side, when the frailty distribution is assumed to follow a gamma distribution, rather than a normal distribution, the integration of the conditional likelihood can be done in closed form. This means the full likelihood can be maximised by a generic maximisation algorithm. This was done and compared to the result using the frailty software, and close agreement was found. In the case of the normally distributed frailty the software, as a default action, appoximates a block of the Hessian matrix with a diagonal matrix so as to save computational time. This was compared to using the full Hessian matrix and was found to agree well with the approximation and did save substantial computation time.

7.7 Summary

This chapter has considered the practical implementation of a model which assumes the latent failure times have a dependency which is induced by a univariate frailty variable. The method of estimation is closely based upon the theory of generalised linear mixed models as can easily be seen when the counting process formulation is modeled as a Poisson random variable. From the theory of the generalised linear mixed model it easier, on a practical level, to use the penalised partial likelihood although there are

several *ad hoc* approximations which have not been fully examined. A minor alteration to the estimation algorithm has been proposed which certainly reduces the number of iterations but, at present, cannot compete, in terms of computational time, with the existing frailty software.

Chapter 8

Pólya trees

8.1 Introduction

Pólya trees were introduced by Ferguson (1974) as an intermediate step between Dirichlet processes, which were, and still commonly are, the default choice of tool for Bayesian non-parametric analyses, and more general tail-free processes (Schervish 1995, section 1.6.2, pp. 60–72).

The name originates from the Pólya urn, where there is a urn containing a fixed number, b, of black balls and a fixed number, r, of red balls. A ball is drawn, replaced and then an additional ball of the same colour is added to the urn. This random sequence of balls is exchangeable, thus invoking De Finetti's theorem (De Finetti 1937/1964), and the probability of drawing a black ball, say, can be shown to follow a beta distribution with parameters corresponding to the original number of balls, (b, r) (Mauldin, Sudderth and Williams 1992). Now consider a tree with two arms from each node which is extended to an infinite number of levels and branches. Now 'ascend' up the tree from the root and at each node draw from a Pólya urn associated with that node; if the ball

is black go left, if the ball is red go right. If the nodes at each level correspond to a partition of the sample space of a random variable, where each extra level is a refinement of the previous level, then this is a mechanism for simulating from a random distribution (as opposed to a fixed distribution of random variables). This random distribution is referred to as a Pólya tree.

The principal attraction of Pólya trees is their use as a Bayesian non-parametric tool. When used to model an unknown or, equivalently, a random distribution they improve upon the normal Dirichlet process since a density realised from a random Pólya tree is finite with probability one. The Dirichlet process will almost surely give a discrete distribution, which is an incorrect imposition in many models. In addition, Pólya trees are highly tractable when the prior distribution is updated to a posterior distribution having observed data. The mathematics which underpin Pólya tree theory is covered in Lavine (1992), Mauldin et al. (1992), Lavine (1994), and a more readable work is Walker and Mallick (1997) which covers some practical applications.

This first section of this chapter will formally define a Pólya tree and sketch the main results concerning the posterior distribution and the sample space being the set of continuous distributions.

The next section aims to investigate how to set the parameters of the Pólya tree as a prior distribution so as to reflect any prior beliefs. This is considered in two ways. A mean prior distribution is chosen, f, and then one considers, marginally at a fixed point y, the distribution of the random variable, $f_\infty(y)/f(y)$, where f_∞ represents a density sampled from the Pólya tree. It is shown that, with a particular choice of parameters, this follows a gamma distribution with mean 1, and a variance of our choosing. Secondly, a more ad hoc means of considering the strength of the prior, is to say that any density sampled from a Pólya tree can be approximated by a normal distribution. It is

approximated in the sense that we can choose two, convenient, predetermined intervals in the sample space, I_1, I_2, examine what probabilities the random density attaches to these intervals, and then choose the two parameters of the normal distribution so that the probabilities coincide. We then consider the distribution of the two, random parameters of the approximating normal distribution.

The following section then considers the posterior distribution. Partial results concerning the density of f_∞/f are obtained, but the question of whether the posterior density is consistent remains unanswered.

Fortunately, for most practical purposes the interest lies in a quantity which can be represented as an integral with respect to an unknown density, rather than the density itself. This can be the probability of a particular interval or the expectation of a random variable. In section 8.5 the question of integrating in practice is considered.

The penultimate section presents together some results concerning Pólya trees which do not fit naturally elsewhere in this chapter and which may appear rather esoteric.

8.2 Definitions and existing results

8.2.1 Definitions

Pólya trees are a means to specify priors over a space of distributions on an arbitrary measure space. They are suited to performing non-parametric analysis within in the Bayesian paradigm. A Pólya tree $\mathcal{P}$ is characterised by two objects $(\Pi, \mathcal{A})$.

The first, Π is a sequence of binary partitions of the sample space Ω, where $\Omega = \pi_0 \cup \pi_1$, $\pi_1 = \pi_{10} \cup \pi_{11}$, and in general each element of the partition, at level m, is denoted π_ϵ where ϵ is a binary sequence length m , where $\pi_\epsilon = \pi_{\epsilon 0} \cup \pi_{\epsilon 1}$. So, $\Pi = \bigcup_\epsilon \pi_\epsilon$.

The second is a sequence of random variables on the unit interval,

$$\mathcal{A} = \{C_\Omega, C_0, C_1, C_{00}, C_{01}, C_{000}, \dots\},$$

where C_ϵ represents $\mathbb{P}(X \in \pi_{\epsilon 0} | X \in \pi_\epsilon)$, where X is a random variable following a realisation of the Pólya tree distribution; $C_\Omega = \mathbb{P}(X \in \pi_0)$. Informally, the subscripts, ϵ, of the random probabilities, C_ϵ, denote which interval is being conditioned on, π_ϵ, and the value of C_ϵ gives the conditional probability of being in the $\pi_{\epsilon 0}$, rather than the $\pi_{\epsilon 1}$, sub-partition. Since a Pólya tree is intended to represent a distribution with uncertainty—effectively random—this explains why $\mathcal{A}$ is a collection of random variables rather than fixed constants. All the random variables in $\mathcal{A}$ are independent and for every ϵ,

$$C_\epsilon \sim Beta(\alpha_{\epsilon 0}, \alpha_{\epsilon 1}).$$

A beta distribution $Beta(\alpha, \beta)$ has density function

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \text{ on } [0,1];$$

this results in a mean of $\alpha/(\alpha+\beta)$ and a variance of $\alpha\beta/(\alpha+\beta)^2/(\alpha+\beta+1)$. For some of the material further on in this chapter, it is useful to note that, if desired, we could define $1 - C_\epsilon$, rather than C_ϵ in which case we just swap the parameters around:

$$C_\epsilon \sim Beta(\alpha_{\epsilon 0}, \alpha_{\epsilon 1}) \Leftrightarrow 1 - C_\epsilon \sim Beta(\alpha_{\epsilon 1}, \alpha_{\epsilon 0}). \tag{8.1}$$

So in this formulation, for every $m = 1, 2, \dots$ and every $\epsilon = \epsilon_1 \epsilon_2 \dots \epsilon_m$,

$$\mathcal{P}(\pi_{\epsilon_1 \epsilon_2 \dots \epsilon_m}) = \left(\prod_{\substack{j=1; \\ \epsilon_j = 0}}^{m} C_{\epsilon_1 \dots \epsilon_{j-1}} \right) \left(\prod_{\substack{j=1; \\ \epsilon_j = 1}}^{m} (1 - C_{\epsilon_1 \dots \epsilon_{j-1}}) \right),$$

where the factors are C_Ω or $1 - C_\Omega$ if $j = 1$, and $\epsilon_1 = 0$ or 1 respectively.

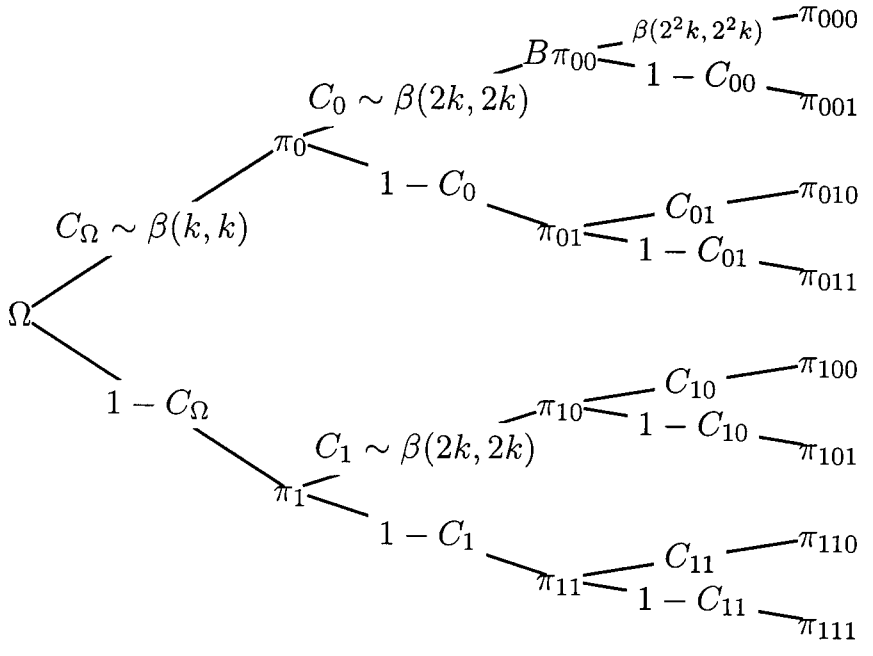


Figure 8.1: Relationship between π_ϵ and C_ϵ

8.2.2 Choice of hyper-parameters

The interdependence between the partition sequence, Π, and the conditional probabilities, $\mathcal{A}$, allows room for manoeuvre to simplify the distributions of $\mathcal{A}$. A convenient way to represent a desired mean prior is to let all the α_ϵ at level m be equal to some function of m. This simplification implies that $\mathbb{E}C_\epsilon = 1/2$. To accommodate this the partitions have to be chosen to have the prior mean probability of 2^{-m}.

Theorem 1.121 of Schervish (1995, pp. 66–68) shows that a sufficient condition for the limiting distribution of $f_n(x) = \mathcal{P}(\pi_{\epsilon_1 \ldots \epsilon_n})/\mu(\pi_{\epsilon_1 \ldots \epsilon_n})$ (μ is a dominating measure) to be finite μ-a.e., equivalently $\mathcal{P}$ is a continuous distribution, is that

$$\sup_n \int_{\mathcal{B}} \mathbb{E}[f_n^2(x)] d\mu(x) < \infty,$$

for all $\mathcal{B}$ which are measurable in the σ-algebra generated by Π.

If we let $\pi_n = \bigcup_{\epsilon = \epsilon_1 \ldots \epsilon_n} \pi_\epsilon$, then lemma 1.124 of Schervish (1995, p 68) shows

that if

$$\sum_{n=1}^{\infty} \sup_{\pi_\epsilon \in \pi_n} \text{Var}(C_\epsilon)/(\mathbb{E}C_\epsilon)^2 < \infty$$

then $\sup_n \int_{\mathcal{B}} \mathbb{E}[f_n^2(x)] d\mu(x) < \infty$. Hence in the special case considered here, $\text{Var}(C_\epsilon)/(\mathbb{E}C_\epsilon)^2 = 1/(2\alpha_n+1)$ where α_n is the parameter of the beta distribution common to level n. Hence sequences of the form $\alpha_n = cn^p$ for $p > 1$, constant c, or $\alpha_n = c\alpha^n$ for $\alpha > 1$ will satisfy this condition. A popular choice in the literature is $0.1n^2$.

8.2.3 Posterior Conjugacy

If data $\{X_1, \ldots, X_n\}$ are observed then updating the posterior distribution according to Bayes rule for $C_\epsilon = \mathcal{P}(X \in \pi_{\epsilon 0} | X \in \pi_\epsilon)$ is proportional to

$$C_\epsilon^{\sum_i I(X_i \in \pi_{\epsilon 0})} (1 - C_\epsilon)^{\sum_i I(X_i \in \pi_{\epsilon 1})} \times C_\epsilon^{\alpha_{\epsilon 0}} (1 - C_\epsilon)^{\alpha_{\epsilon 1}}.$$

So, we have prior-posterior conjugacy, where $\alpha_\epsilon \mapsto \alpha_\epsilon + \sum_i I(X_i \in \pi_\epsilon)$.

8.3 Interpretation of the strength of prior

8.3.1 Convergence of the density estimator

In this section we will consider the limiting distribution of $f_n(x)$. Given its form as an infinite product it is easiest to consider its logarithm. Now, defining $Y = \ln(Beta(\alpha, \beta))$, we can see that the transformed density function is $B(\alpha, \beta) e^{\alpha y} (1 - e^y)^{\beta - 1}$, where $B(\alpha, \beta)$ is the normalising constant of the form $\Gamma(\alpha + \beta)/\Gamma(\alpha)/\Gamma(\beta)$. The moment generating function, $\mathbb{E}(e^{tY})$ can easily be calculated,

$$\psi(t) = \int_{-\infty}^{0} B(\alpha, \beta) e^{(\alpha + t)y} (1 - e^y)^{\beta - 1} dy = B(\alpha, \beta)/B(\alpha + t, \beta).$$

The definition of $f_n(x)$ is

$$f_n(x) = \left(\prod_{\substack{j=1; \\ \epsilon_j=0}}^{n} C_{\epsilon_1 \ldots \epsilon_{j-1}} \right) \left(\prod_{\substack{j=1; \\ \epsilon_j=1}}^{n} (1 - C_{\epsilon_1 \ldots \epsilon_{j-1}}) \right) \Big/ \mu(\pi_{\epsilon_1 \ldots \epsilon_n}),$$

where $x \in \pi_{\epsilon_1 \ldots \epsilon_n}$ and μ is Lebesgue measure. The random variables, $C_{\epsilon_1 \ldots \epsilon_{j-1}}$ follow a $Beta(\alpha_j, \beta_j)$ distribution, so it is clear that without some further assumptions about (α_j, β_j) it is not possible to gain further insights.

Dropping the dependencies on x and using equation (8.1) to relabel the $C_{\epsilon_1 \ldots \epsilon_{j-1}}$, and $\mu(\pi_{\epsilon_1 \ldots \epsilon_{j-1}})$ so that we have,

$$f_n = \prod_{j=1}^{n} C_j / \mu_n,$$

we are free to choose $\alpha_j = \beta_j = k2^{j-1}$. We can now see that the moment generating function of $\ln(f_n) = Z_n$ is of the form

$$\begin{aligned}
\psi_{Z_n}(t) &= \exp(-\ln(\mu_n)t) \prod_{j=1}^{n} \psi_{\ln C_j}(t) \\
&= \exp(-\ln(\mu_n)t) \frac{\Gamma(2k)\Gamma(k+t)}{\Gamma(k)\Gamma(2k+t)} \cdot \frac{\Gamma(4k)\Gamma(2k+t)}{\Gamma(2k)\Gamma(4k+t)} \cdots \frac{\Gamma(2^n k)\Gamma(2^{n-1}k+t)}{\Gamma(2^{n-1}k)\Gamma(2^n k+t)} \\
&= \exp(-\ln(\mu_n)t) \frac{\Gamma(k+t)\Gamma(2^n k)}{\Gamma(k)\Gamma(2^n k+t)}
\end{aligned}$$

Now, μ_n is defined to be $F^{-1}((k+1)2^{-n}) - F^{-1}(k2^{-n})$ for a suitable integer $k(x,n)$, where F is the C.D.F. of the expected prior. So, assuming F is continuous with first derivatives, as $n \to \infty$, $\mu_n 2^n \to dF^{-1}/dq(F(x)) = 1/f(x)$.

Now given the result that $\mu_n = O(2^{-n})$ we need to consider the limit of $2^{nt}\Gamma(2^n k)/\Gamma(2^n k + t)$ as $n \to \infty$. Using Stirling's formula which states that

$$\lim_{n \to \infty} \sqrt{2\pi} n^{n+1/2} e^{-n} / \Gamma(n+1) = 1$$

and replacing 2^n with m, we obtain

$$\begin{aligned}\lim_{m\to\infty} m^t\Gamma(mk)/\Gamma(mk+t) &= \lim_{m\to\infty} \frac{(mk-1)^{mk-1/2}e^{-mk-1}m^t}{(mk+t-1)^{mk+t-1/2}e^{-mk-t-1}} \\ &= \lim_{m\to\infty} \left(1+\frac{t}{mk-1}\right)^{-(mk-1)} \left(1+\frac{t}{mk-1}\right)^{-1/2} \\ &\times \left(k+\frac{t-1}{m}\right)^{-t} e^t\end{aligned}$$

using the result that $(1+x/n)^n \overset{n\to\infty}{\longrightarrow} e^x$

$$= k^{-t}.$$

So, if we denote $c = \ln f(x)$ the limiting function of ψ_{Z_n},

$$\psi_Z(t) = \frac{e^{ct}\Gamma(k+t)}{k^t\Gamma(k)}. \tag{8.2}$$

Now examining $\Gamma(k+t)/k^t$, it can be seen that,

$$\begin{aligned}\Gamma(k+t)/k^t &= \frac{1}{k^t}\int_0^\infty x^{k+t-1}e^{-x}dx \\ &= \int_0^\infty x^k \left(\frac{x}{k}\right)^t e^{-x}\frac{dx}{x}\end{aligned}$$

using the substitution $y = \ln(x/k)$

$$= \int_{-\infty}^{\infty} (ke^y)^k \exp\left(-ke^y\right) e^{ty}dy.$$

Standard M.G.F. theory tells us that if $F(t)$ is the M.G.F. of $f(y)$, then $e^{ct}F(y)$ is the M.G.F. of $f(y-c)$. Hence we have found the distribution of $y = \ln(f_\infty/f)$, which is proportional to $\exp(-k(e^y-y))$. Making the change of variables $x = f_\infty/f = e^y$, we obtain the distribution proportional to

$$x^{k-1}e^{-kx}.$$

Hence we see that the random variable $w = f_\infty/f$ follows a gamma distribution with shape and scale parameter k. From this we can calculate pointwise confidence intervals for different values of k, which give an indication of the strength of the prior in terms of the multiplicative factor of deviation from the expected prior distribution. The upper and lower bounds of such $[0.025, 0.975]$ confidence intervals are shown in figure 8.2. The expected value of this distribution is 1, and its variance is $1/k$.

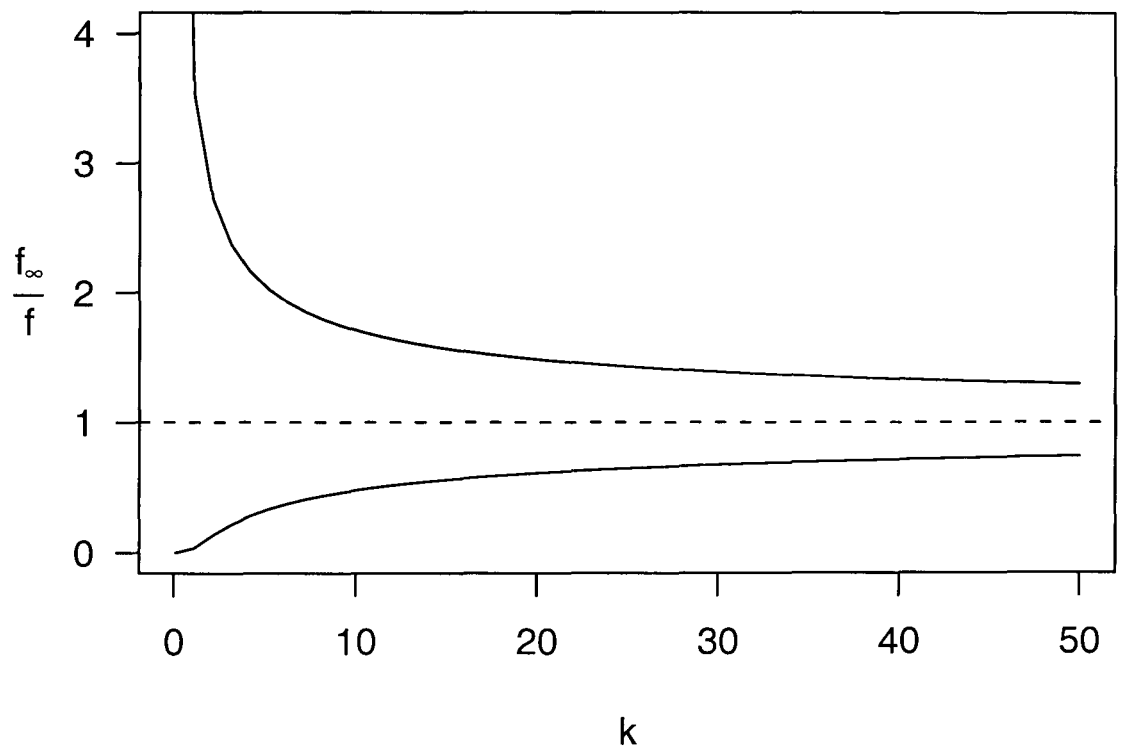


Figure 8.2: Solid lines: upper and lower bounds for the 95% C.I.s. Dashed line: expected value, 1.

8.3.2 Normal Approximation

This subsection makes the more specific assumption that any prior information can be described by setting the expected prior density to be a normal distribution, $N(\mu, \sigma^2)$ and then choosing a form for α_n that reflects any uncertainty. If we assume the partitions Π, at level n, have been set equal to the $j2^{-n}$-th quantiles of the expected prior, then in particular we have that

$$I_1 = \pi_0 = (-\infty, \mu],$$

$$I_2 = \pi_{01} \cup \pi_{10} = (\mu - \sigma\Phi^{-1}(0.75), \mu + \sigma\Phi^{-1}(0.75)],$$

where Φ denotes the C.D.F. of a standard normal density. Now the next step is to approximate any density realised from the Pólya tree to a normal distribution, $N(M, S^2)$, where the random variables M and S are defined to be those such that the random probability attached to the intervals, I_1 and I_2 by the random density, $\mathcal{P}(I_j)$, equals the probability given to these intervals by $N(M, S^2)$.

So M and S have the implicit definitions

$$C_\Omega = \Phi\left(\frac{\mu - M}{S}\right)$$

$$C_\Omega(1 - C_0) + (1 - C_\Omega)C_1 = \Phi\left(\frac{\mu + \sigma\Phi^{-1}(0.75) - M}{S}\right) - \Phi\left(\frac{\mu - \sigma\Phi^{-1}(0.75) - M}{S}\right)$$

These can be simplified slightly to an implicit definition for S which does not involve M, and an explicit definition for M which does involve S.

$$\Phi\left(\Phi^{-1}(C_\Omega) + \frac{\sigma}{S}\Phi^{-1}(0.75)\right) - \Phi\left(\Phi^{-1}(C_\Omega) - \frac{\sigma}{S}\Phi^{-1}(0.75)\right) \tag{8.3}$$

$$= C_\Omega(1 - C_0) + (1 - C_\Omega)C_1$$

$$M = \mu - S\Phi^{-1}(C_\Omega) \tag{8.4}$$

Note that μ only affects the definition of M; equation (8.3) is in terms of σ/S so σ just scales the distribution of S. Since S is almost surely positive $\mathbb{P}(M > \mu) =$

$\mathbb{P}(\Phi^{-1}(C_\Omega) < 0) = \mathbb{P}(C_\Omega < 0.5) = 0.5$, and hence μ is the median of M. A simulation of S variables, where $\alpha_n = k2^{n-1}$ with $k = 3$, as shown in figure 8.3, indicates that the median and mean of S is smaller than σ.

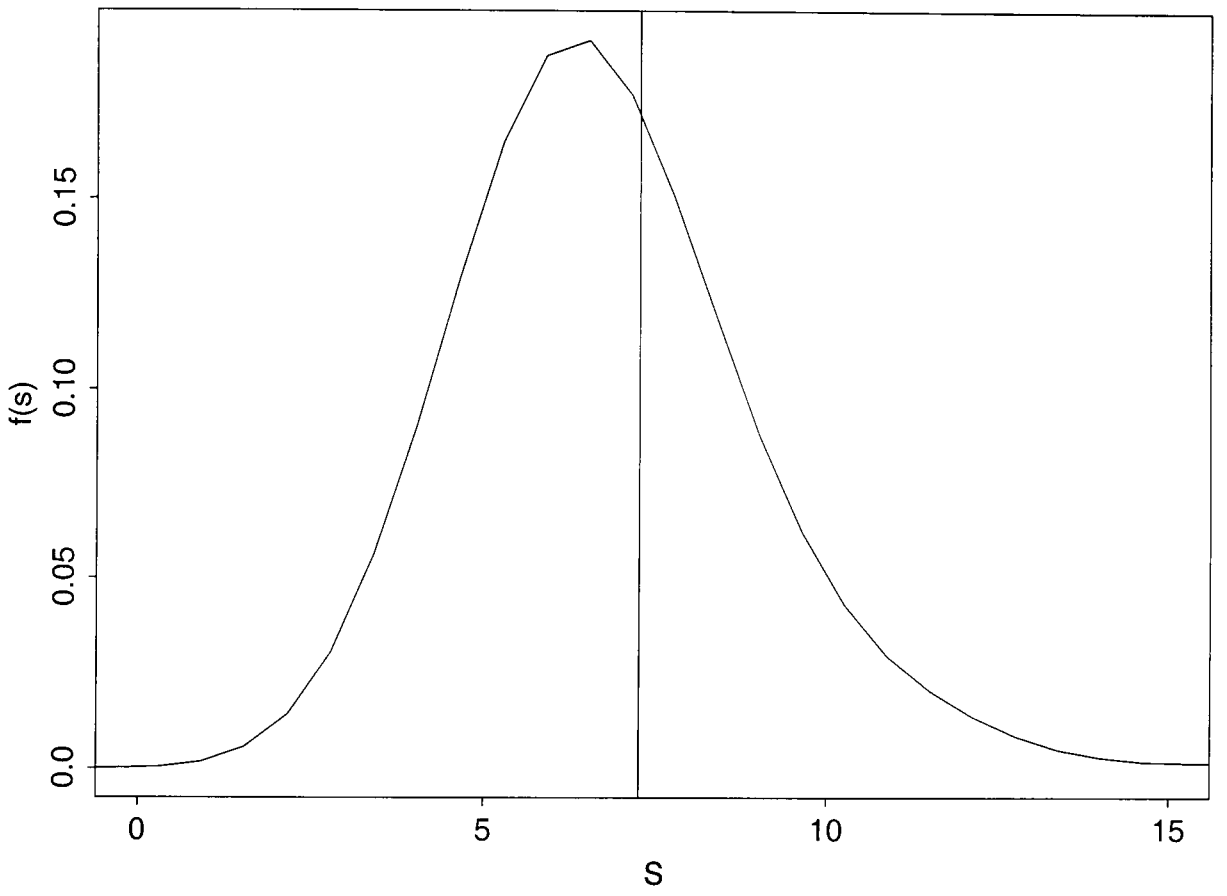


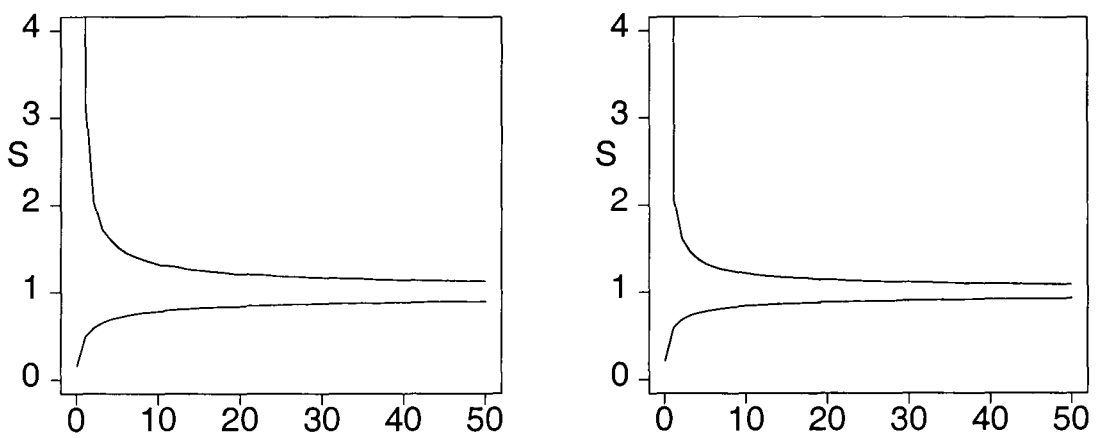
Figure 8.3: A kernel density estimate of S from a simulation size 1000. The vertical line gives σ

In chapter 9 a Pólya tree is used to model a frailty distribution. In this model confounding occurs between the location of the frailty distribution and the location of the fixed effects. To resolve this the value of C_Ω is fixed at $1/2$. This means that the median of any density realised from the Pólya tree is fixed at $\sup \pi_0$. In such cases

$M = \mu$ and equation (8.3) simplifies to

$$\frac{1}{2} + \frac{C_1 - C_0}{2} = 2\Phi\left(\frac{\sigma}{S}\Phi^{-1}(0.75)\right) - 1.$$

Observe that the right hand side equates to 0.5 if $S = \sigma$ and since the left hand side is symmetrically distributed about 0.5 also, this implies that the median of S, in this special case, is equal to σ. In figure 8.4, confidence intervals for S/σ are shown as functions of k, where k parametrises two choices of form for α_n: $k2^{n-1}$ and kn^2.



(a) $\alpha_n = k2^{n-1}$ (b) $\alpha_n = kn^2$

Figure 8.4: 95% C.I.s for S/σ.

8.4 Posterior

Considering the posterior distribution, in a pointwise fashion, the parameters $\{\alpha_\epsilon\}$ are transformed to $\{\alpha_\epsilon + N_\epsilon\}$ where N_ϵ is defined to be $\sum_i I(X_i \in \pi_\epsilon)$. Hence the moment generating function of $Y_\epsilon = \ln C_\epsilon$ becomes

$$\frac{B(\alpha_{\epsilon 0} + N_{\epsilon 0}, \alpha_{\epsilon 1} + N_{\epsilon 1})}{B(\alpha_{\epsilon 0} + N_{\epsilon 0} + t, \alpha_{\epsilon 1} + N_{\epsilon 1})}.$$

Using the result, $\Gamma(n+1) = n\Gamma(n)$, this can be simplified to

$$\left(\prod_{i=0}^{N_{\epsilon 0}-1} \frac{\alpha_{\epsilon 0}+i+t}{\alpha_{\epsilon 0}+i}\right)\left(\prod_{i=0}^{N_{\epsilon}-1} \frac{\alpha_{\epsilon 0}+\alpha_{\epsilon 1}+i}{\alpha_{\epsilon 0}+\alpha_{\epsilon 1}+i+t}\right) B(\alpha_{\epsilon 0},\alpha_{\epsilon 1})/B(\alpha_{\epsilon 0}+t,\alpha_{\epsilon 1}), \quad (8.5)$$

and if we want to consider the M.G.F. of $\ln(1-C_\epsilon)$ then we simply swap the subscripts $\epsilon 0$ and $\epsilon 1$ in (8.5).

It is helpful to consider the effect on the posterior distribution when a single value x is observed. First some definitions: we have a sequence of ϵ_i such that,

$$x \in \pi_{\epsilon_m} \subset \pi_{\epsilon_{m-1}} \subset \ldots \subset \pi_{\epsilon_2} \subset \pi_{\epsilon_1},$$

where m tends to infinity; also we have the counts observed 'before' the observation x, $N_\epsilon = \sum_i I(x_i \in \pi_\epsilon)$, for any ϵ

For any point of interest, $y \neq x_i, x$, in the support, Ω, there will be a level l such that

$$y \in \bigcap_{i=1}^{l} \pi_{\epsilon_i}, \text{ but } y \notin \bigcup_{i>l} \pi_{\epsilon_i}.$$

At each level, i, for $1 \leq i \leq l-1$, the moment generating function, $\psi(t)$ (on the $\ln(f_\infty(\cdot)/f(\cdot))$ scale) is multiplied by the factor

$$\left(\frac{\alpha_{\epsilon_{i+1}}+N_{\epsilon_{i+1}}+t}{\alpha_{\epsilon_{i+1}}+N_{\epsilon_{i+1}}}\right)\left(\frac{\alpha_{\epsilon_i 0}+\alpha_{\epsilon_i 1}+N_{\epsilon_i}}{\alpha_{\epsilon_i 0}+\alpha_{\epsilon_i 1}+N_{\epsilon_i}+t}\right);$$

at the level $i=l$, where the paths of x and y diverge, equivalently, $x, y \in \pi_{\epsilon_l},\ y \notin \pi_{\epsilon_{l+1}}$, the moment generating function $\psi(t)$ is multiplied just by

$$\frac{\alpha_{\epsilon_l 0}+\alpha_{\epsilon_l 1}+N_{\epsilon_l}}{\alpha_{\epsilon_l 0}+\alpha_{\epsilon_l 1}+N_{\epsilon_l}+t}.$$

Now we consider what is the effect of each of these factors. It is easily shown that $\lambda/(\lambda+t)$ is the M.G.F. of the density function, $\lambda\exp(\lambda x), x \leq 0$, so in effect, at

each level $i \leq l$ we subtract from the prior distribution for $\ln(f_\infty(\cdot)/f(\cdot))$ an independent exponential variable with mean $1/(\alpha_{\epsilon_i 0} + \alpha_{\epsilon_i 1} + N_{\epsilon_i})$.

The other factor is more complex and it does not correspond to the addition of a random variable, rather it is a transformation on the existing posterior distribution. For an arbitrary density function g it is true that

$$\int_{\Omega_X} e^{tx} \left(g(x) - \frac{g'(x)}{\lambda} \right) dx = \left(\frac{\lambda + t}{\lambda} \right) \int_{\Omega_X} e^{tx} g(x) dx + \left[-\frac{g(x)e^{tx}}{\lambda} \right]_{\Omega_X}. \quad (8.6)$$

Hence the effect of multiplying a M.G.F. by the factor $(1 + t/\lambda)$ is to transform the density function with

$$\mathcal{T}_\lambda : g \mapsto g - \frac{g'}{\lambda},$$

assuming that the density function function decays quickly enough in its tails. In the specific case considered here $\lambda = \alpha_{\epsilon_{i+1}} + N_{\epsilon_{i+1}}$ and g is the density function of $\ln(f_\infty(y|X)/f(y))$. Lemma 8.4.1 establishes that the final term in (8.6) does indeed vanish for an arbitrary number of iterations of this transformation, assuming that we start off with a log-gamma distribution. Hence by the uniqueness-inversion property of moment generating functions we can infer that this transformation does leave us with a density function.

Lemma 8.4.1. *Define the density function* $g_0(y) \propto \exp(-k(e^y - y))$, $y \in \mathbb{R}$, *for positive constant,* k. *Also, define the mapping* $\mathcal{T}_\lambda : \mathcal{F} \mapsto \mathcal{F}$, *such that* $f \mapsto f - f'/\lambda$, *where* $\mathcal{F}$ *is the space of density functions on* $\mathbb{R}$. *Then,*

$$\lim_{y \to \pm\infty} \mathcal{T}_\lambda^n(g_0)(y) e^{ty} = 0,$$

where $\mathcal{T}_\lambda^n$ *is the* n*th convolution of* $\mathcal{T}_\lambda$, *for* $n = 0, 1, \ldots$.

Proof. For positive a, b,

$$\lim_{y\to\infty} -ae^y + by = -\infty$$
$$\Rightarrow \lim_{y\to\infty} \exp(-ae^y + by) = 0$$

also

$$\lim_{y\to-\infty} -ae^y + by = -\infty$$
$$\Rightarrow \lim_{y\to-\infty} \exp(-ae^y + by) = 0$$

Hence for $n = 0$, $a = k$ and $b = t + k$ the result holds when $t > -k$. Since

$$\mathcal{T}_\lambda(\exp(-ae^y + by)) = \left(1 - \frac{b}{\lambda}\right)\exp(-ae^y + by) + \frac{a}{\lambda}\exp(-ae^y + (b+1)y),$$

it is clear that

$$\mathcal{T}_\lambda^n(g_0) = \sum_{r=0}^{n} \beta_r \exp(-ke^y + (k+r)y) \tag{8.7}$$

for some coefficients β_r. Therefore the result is true in general. □

Since the updating of the posterior is equivalent to repeatedly adding an independent variable and repeatedly performing the transformation $\mathcal{T}_\lambda$, it would be worrying if these operations—the adding and transforming—gave different results if the order in which they are performed were changed. Lemma 8.4.2 proves that this is not the case.

Lemma 8.4.2. *Given two independent random variables, X, Y with densities f_X, f_Y, and the mapping $\mathcal{T}_\lambda$ as defined in lemma 8.4.1, then*

$$\mathcal{T}_\lambda\left\{\int_{\Omega_X} f_X(x) f_Y(w-x)dx\right\}(w) = \int_{\Omega_X} f_X(x)\mathcal{T}_\lambda\left\{f_Y\right\}(w-x)dx.$$

Proof.

$$
\begin{aligned}
&\mathcal{T}_\lambda \left\{ \int_{\Omega_X} f_X(x) f_Y(w-x) dx \right\} (w) \\
&= \int_{\Omega_X} f_X(x) f_Y(w-x) dx - \frac{1}{\lambda} \frac{d}{dw} \int_{\Omega_X} f_X(x) f_Y(w-x) dx \\
&= \int_{\Omega_X} f_X(x) f_Y(w-x) dx - \int_{\Omega_X} f_X(x) \frac{f_Y'(w-x)}{\lambda} dx \\
&= \int_{\Omega_X} f_X(x) \left\{ f_Y(w-x) - \frac{f_Y'(w-x)}{\lambda} \right\} dx \\
&= \int_{\Omega_X} f_X(x) \mathcal{T}_\lambda \left\{ f_Y \right\} (w-x) dx.
\end{aligned}
$$

□

To summarise the logarithm posterior density divided by the expected prior density at a fixed point y, $\log\{f_\infty(y)/f(y)\}$, is a random variable which can be represented as

$$
\left(\prod_{m=1}^{L} \prod_{i=0}^{N_{\epsilon_m}-1} \mathcal{T}_{k2^{m-1}+i} \right) \{\log \mathsf{GAMMA}(k,k)\} - \sum_{m=1}^{L+1} \sum_{i=0}^{N_{\epsilon_{m-1}}-1} \mathsf{EXP}(k2^m + i), \quad (8.8)
$$

where it assumed that $y \in \pi_{\epsilon_L} \subset \pi_{\epsilon_{L-1}} \subset \ldots \subset \pi_{\epsilon_1}$; that N_ϵ denotes the number of observations in π_ϵ and that L refers to the highest level partion which contains y *and* has a count $N_{\epsilon_L} > 0$. The two expression $\mathsf{GAMMA}(\cdot)$ and $\mathsf{EXP}(\cdot)$ refer to gamma and exponential distributions with their parameters. There is an abuse of notation in

$$
\left(\prod_{\lambda} \mathcal{T}_\lambda \right) \{X\}
$$

which refers to a random variable with density equal to a convolution of the transformation $\mathcal{T}$ applied to the original density of X

$$
\mathcal{T}_{\lambda_1} \circ \mathcal{T}_{\lambda_2} \circ \ldots \circ \mathcal{T}_{\lambda_k} \{g_X\}.
$$

Considering equation (8.7) we can see that the density associated with the convolution of $\mathcal{T}$ in this particular case has the form

$$\sum_{r=0}^{R} \beta_r \exp(-ke^y + (k+r)y),$$

where $R = \sum_{m=1}^{L} (N_{\epsilon_m} - 1)$. This is a linear sum of densities of the logarithm of gamma distributions with scale k and shape $(k+r)$ and it could be postulated that this a mixture distribution. However the coefficients β_r, which depend upon the α_m and the observed counts N_ϵ, can become negative with increasing sample size, so we cannot apply this convenient interpretation.

So in summary the distribution posterior of the posterior density is modified by the convolution of the transformation $\mathcal{T}$ and is then added to a sequence of negative exponential variables as summarised in 8.8. The moments of the posterior are considered below although there is no clear conclusion.

It is easily proved by induction that, for an infinitely differentiable function M, for $n = 0, 1, 2, \ldots$,

$$\frac{d^n}{dt^n}\left\{\left(\frac{\lambda+t}{\lambda}\right)M(t)\right\} = \frac{d^n M}{dt^n}(t) + \frac{t}{\lambda}\frac{d^n M}{dt^n}(t) + \frac{n}{\lambda}\frac{d^{n-1}M}{dt^{n-1}}(t),$$

So if we consider M as a M.G.F. we see that the transformation $\mathcal{T}$ was arrived at by considering the effect of multiplying a M.G.F. by a factor $(1+t/\lambda)$. Hence we can use this expression to calculate moments of a transformed distribution. If Y is the transformation of X it follows that by evaluating at $t = 0$, $\mathbb{E}(Y^n) = \mathbb{E}(X^n) + n\mathbb{E}(X^{n-1})/\lambda$, for any integer, n. In terms of mean and variance, this is

$$\mathbb{E}(Y) = \mathbb{E}(X) + 1/\lambda$$

$$\mathsf{Var}(Y) = \mathsf{Var}(X) - 1/\lambda^2.$$

An EXP(λ) distribution has mean $1/\lambda$ and variance $1/\lambda^2$. A $\log$ GAMMA(k,k) distribution has mean $\Gamma'(k)/\Gamma(k) - \log k$ and variance $\left(\Gamma(k)\Gamma'(k) - \{\Gamma'(k)\}^2\right)/\Gamma^2(k)$.

So expressions for the expectation and variance of $\ln f_\infty(y)/f(y) = \mathcal{D}$ are below

$$\mathbb{E}(\mathcal{D}) = \frac{\Gamma'(k)}{\Gamma(k)} - \log k + \sum_{m=1}^{L} \sum_{i=0}^{N_{\epsilon_m}-1} \frac{1}{k2^{m-1}+i} - \sum_{m=1}^{L+1} \sum_{i=0}^{N_{\epsilon_{m-1}}-1} \frac{1}{k2^m+i}$$

$$\text{Var}(\mathcal{D}) = \frac{\Gamma'(k)}{\Gamma(k)} - \left(\frac{\Gamma'(k)}{\Gamma(k)}\right)^2 - \sum_{m=1}^{L} \sum_{i=0}^{N_{\epsilon_m}-1} \frac{1}{(k2^{m-1}+i)^2} + \sum_{m=1}^{L+1} \sum_{i=0}^{N_{\epsilon_{m-1}}-1} \frac{1}{(k2^m+i)^2}.$$

Obtaining asymptotic results on these expressions as the sample size increases is an unsolved problem. The problem is that as the level increases the count decreases and eventually becomes zero at some level, L. This level L increases with sample size and α_m increases with the level which means that we would expect the sum to converge, but what it converges to is unknown and there is no reason to think it converges to the correct value, $\log g/f$ where g is the 'true' sampling distribution. This is not too surprising considering the results of section 3.3 in Barron, Schervish and Wasserman (1999) which proves that a sufficient condition for consistency is that the parameters $\alpha_m = 8^m$, which is increasing far quicker than the choice of $\alpha = 2^m$ in this chapter. However, Barron et al. (1999) do not prove it is a necessary condition and there are no such results, at present, which give necessary *and* sufficient conditions for the consistency of the posterior density with a Pólya tree prior.

8.5 Integration with respect to a Pólya tree

When reporting on the results of a fitted model many important quantities can be expressed in terms of integrals with respect to the posterior density. For example, the probability that a random variable is less than zero, the expectation of a random variable, credible or confidence intervals, and expected utility. This section will consider

the practical integration of a Pólya tree which is considered up to a finite level. The main problem here is how to cope with the tail, or most extreme partitions. In practice we cannot extend the partitioning to an infinitely fine level as this would require infinite amounts of computer memory and processing. Fortunately our model is such that the parameters α_m increase with level and effectively says that as we examine the density conditional on some partition π_{ϵ_m}, this almost surely approaches a uniform density as the size (Lebesgue measure) of the partition decreases. This is exactly what we would anticipate in the case of a continuous distribution and in practice means we only need to monitor the Pólya tree up to a finite level. The literature seems to recommend a level of 8, or, equivalently, intervals with an expect prior probability of $2^{-8} = 0.00391$. The down side of this is that the tail partitions extend to $\pm\infty$ if the sample space is $\mathbb{R}$ and thus if the posterior puts significant mass in these intervals any approximation may be highly biased. In reality any numerical integration must consider a finite interval and hope that any region of the sample space which is ignored would only contribute a negligible amount to any integral which is being approximated. We consider how to choose such finite intervals.

The easiest way to perform integration with a realised Pólya tree as the integrating measure is to perform a version of the trapezium rule. That is to approximate $\mathbb{E}(f(x)|X \in \pi_\epsilon)$ with $[f(\sup \pi_\epsilon) + f(\inf \pi_\epsilon)]/2 = \mathcal{E}(f, \epsilon)$, and then an approximation, to level m, is

$$\sum_{\epsilon=\epsilon_1 \ldots \epsilon_m} \left\{ \mathcal{E}(f, \epsilon) \prod_{j=1,\ldots,m} C_{\epsilon_1 \ldots \epsilon_{j-1}} \right\}.$$

If $g(x)$ is the density function for a realised Pólya tree, and $H(x), h(x)$ are the C.D.F and density, respectively for the expected prior distribution, then we are using the

trapezium rule on

$$\int_{y=0}^{1} \frac{f[H^{-1}(y)]g[H^{-1}(y)]}{h[H^{-1}(y)]}dy = \int_{H^{-1}(y)=x\in\Omega} f(x)g(x)dx,$$

where the points of evaluation, in terms of y are $j2^{-m}, j = 0, \ldots, 2^m$. Given that g is almost surely finite, bounds on the error can be provided and these are $M4^{-m}/12$, where M is

$$\sup_{y\in[0,1]} \left| \frac{d^2}{dy^2} \left\{ \frac{f[H^{-1}(y)]g[H^{-1}(y)]}{h[H^{-1}(y)]} \right\} \right|.$$

However, this constant M, may not be finite if the range of integration is infinite. In practical terms, to be able to compute any estimate, we need to choose a constant l, such that integrating over $[-l, l]$ approximates integrating over $\mathbb{R}$. If the sample space is finite then the integrand of interest will be zero in $\mathbb{R}/[-l, l]$ and the two integrals will be equal. If the integral over $\mathbb{R}$ is finite then it must hold that the difference between the 'true' value and the 'truncated' value must have limit zero as l tends to infinity.

For each function $f(x)$, and lower end-point,a there is a value of l such that

$$(f(a) + f(l))/2 = \int_{a}^{\infty} f(x)g(x)dx,$$

where $g(x)$ is the density of X conditional on $X \in [a, \infty)$. This value of l depends upon the actual function . A sensible default choice would be to consider the identity function $f(x) = x$.

A useful part of Pólya tree theory is that the expected posterior density in each partition π_ϵ is the expected prior density scaled by the appropriate amount so that $\mathbb{E}[\mathcal{P}(\pi_{\epsilon 0}|\pi_\epsilon)|$ data $]$ equals the correct value,

$$\frac{\alpha_{\epsilon 0} + \sum_i I(X_i \in \pi_{\epsilon 0})}{\alpha_{\epsilon 0} + \alpha_{\epsilon 1} + \sum_i I(X_i \in \pi_\epsilon)}.$$

With a finite (or real) data set, at a certain level and above, and at a non-zero distance from an observed data point, the expected conditional probabilities will equal the prior

values since $I(X_i \in \pi_\epsilon) = 0$. Hence, the expected posterior density is a rescaled version of the expected prior density. This means to find the expected posterior density (away from the observed data values) we simply do a piecewise re-scaling of the expected prior density which would normally chosen to be of some convenient mathematical form (Lavine 1992); it is not necessary to consider each point, of a continuum of points, in turn.

So if we have chosen the standard normal as the prior, and are partitioning to level 8, then, considering the right tail, $\tau = \pi_{11111111}$ (eight ones) it is the interval $[2.66007, \infty)$, since $1 - \Phi(2.66007) = 2^{-8}$. Hence the task is to solve,

$$(2.66007 + l)/2 = \frac{1}{\mathcal{P}(\tau)} \int_{2.6607}^{\infty} \frac{x}{\sqrt{2\pi}} \exp(-x^2/2) dx,$$

where the right hand side is a rescaled version of the original expected prior distribution. Fortunately the integral has a closed form, $\exp(-(2.66007)^2/2)/\sqrt{2\pi}$, and so the root is

$$l = 0.02316/\mathcal{P}(\tau) - 2.66007.$$

Using the expected value of $\mathcal{P}(\tau)$, which is 2^{-8}, this is 3.2687.

8.5.1 Hermite Polynomial approach

A more systematic consideration would be to consider a basis of functions which span the space of functions with finite expectation with respect to normal measure. For the convenience of the mathematics, a sensible choice is the Hermite polynomials. These are defined to be

$$H_0(x) = 1 \qquad\qquad H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} \left(e^{-x^2} \right).$$

Now defining the definite integral,

$$\begin{aligned} I_n(a) &= \int_a^\infty H_0(x) H_n e^{-x^2} dx \\ &= (-1)^n \int_a^\infty \frac{d^n}{dx^n}\left(e^{-x^2}\right) dx = (-1)^{n-1} \frac{d^{n-1}}{dx^{n-1}}\left(e^{-x^2}\right)\Big|_{x=a} \\ &= e^{-a^2} H_{n-1}(a) \end{aligned}$$

Considering the standard normal density function,ϕ, we can represent $I_n(a)$ as

$$\begin{aligned} I_n(a) &= \sqrt{2\pi} \int_a^\infty H_n(x)\phi(\sqrt{2}x)dx \\ &= \sqrt{\pi} \int_{\sqrt{2}a}^\infty H_n(y/\sqrt{2}) d\phi(y) \end{aligned}$$

So to find an upper bound, l, which will give the correct value with respect to the expected value of $H_n(x/\sqrt{2})$, we need to find a suitable root, in terms of l, of the nth order polynomial

$$1/2\left[H_n(\inf \tau/\sqrt{2}) + H_n(l/\sqrt{2})\right] = \frac{e^{-(\inf \tau)^2/2}}{\sqrt{\pi}\mathcal{P}(\tau)} H_{n-1}(\inf \tau/\sqrt{2}), \qquad (8.9)$$

where $\inf \tau = 2.66007$, and $\mathcal{P}(\tau)$ is replaced by its expected value, 3.2687.

Figure 8.5 shows the smallest such root which is real and greater than $\inf \tau$, for $n = 1 \ldots 50$.

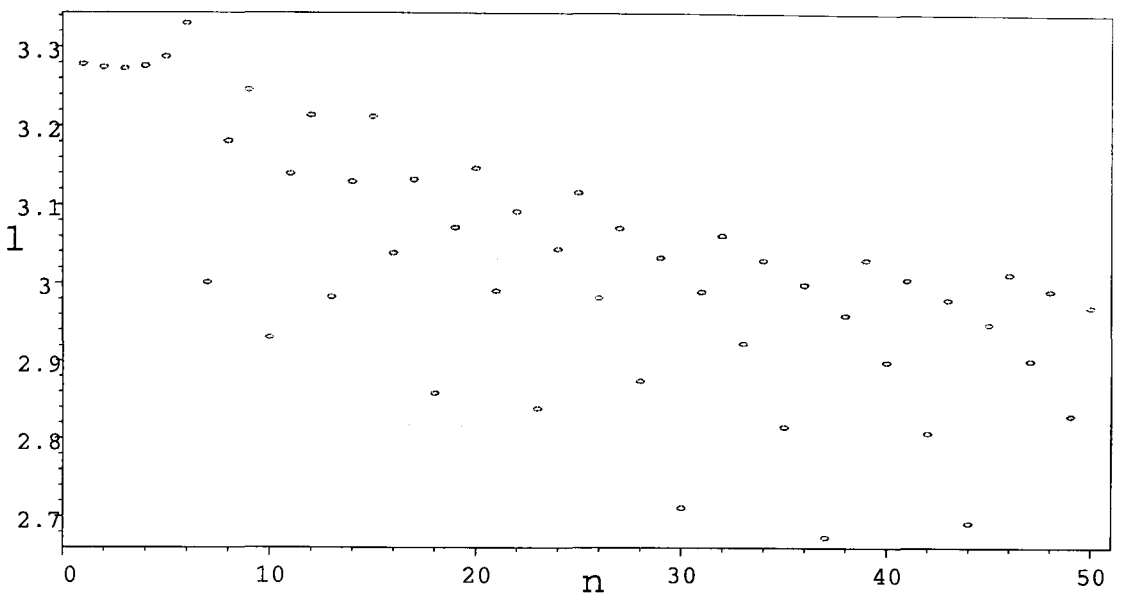


Figure 8.5: First 50 roots

However given that the first five values of l are near to 3.3, the roots nearest to this value were found and these are shown in figure 8.6.

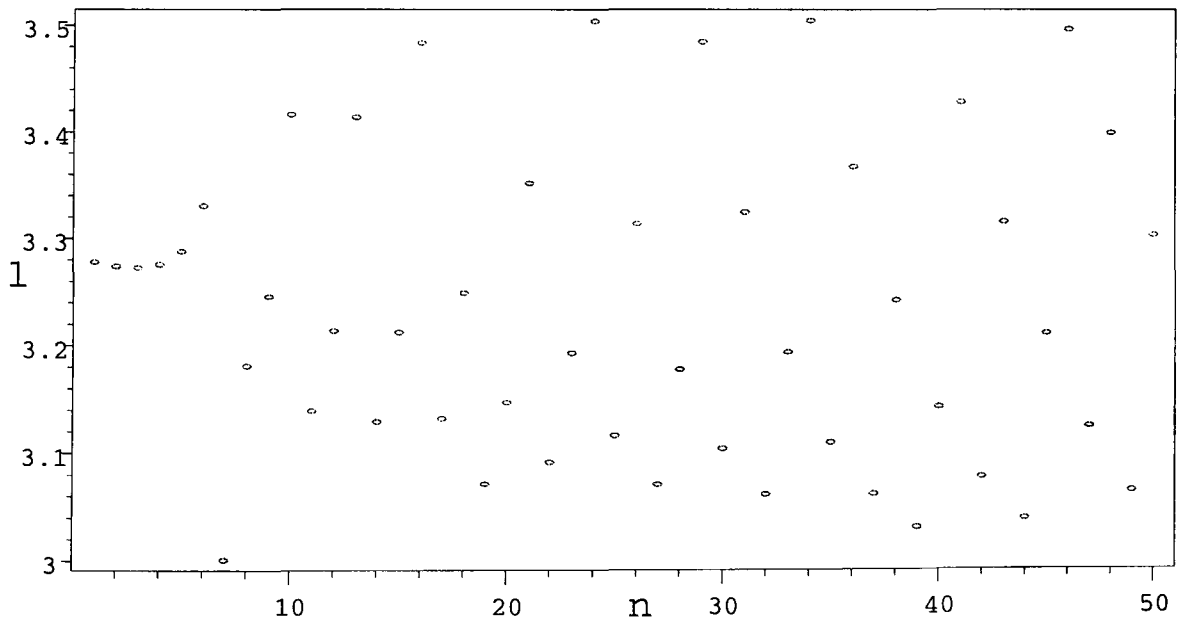


Figure 8.6: The roots nearest to 3.3

Unfortunately it does not offer any discernible improvement in the spread of values.

In practical terms, we want to choose a single value for l. Given the roots of (8.9), denoted as l_n, we can easily compute the percentage error which a different value of l will give:

$$e(n,l) = \left(\frac{H_n(\inf \tau/\sqrt{2}) + H_n(l/\sqrt{2})}{H_n(\inf \tau/\sqrt{2}) + H_n(l_n/\sqrt{2})} - 1 \right) 100\%.$$

Without further information of what function we want to find the expectation, a sensible approach would be to consider averaging e and e^2 over n. With the first function, which we call *bias*, we would like to find a value of l which gives a value of zero; with the second function, which we call *mean square error*, we would like to minimise with respect to l. These two functions are shown in figure 8.7, and they show that a sensible choice lies between 2.99 and 3.00 . However, it must be admitted that the implications of restricting the averaging to $n = 1 \ldots 50$ are unknown.

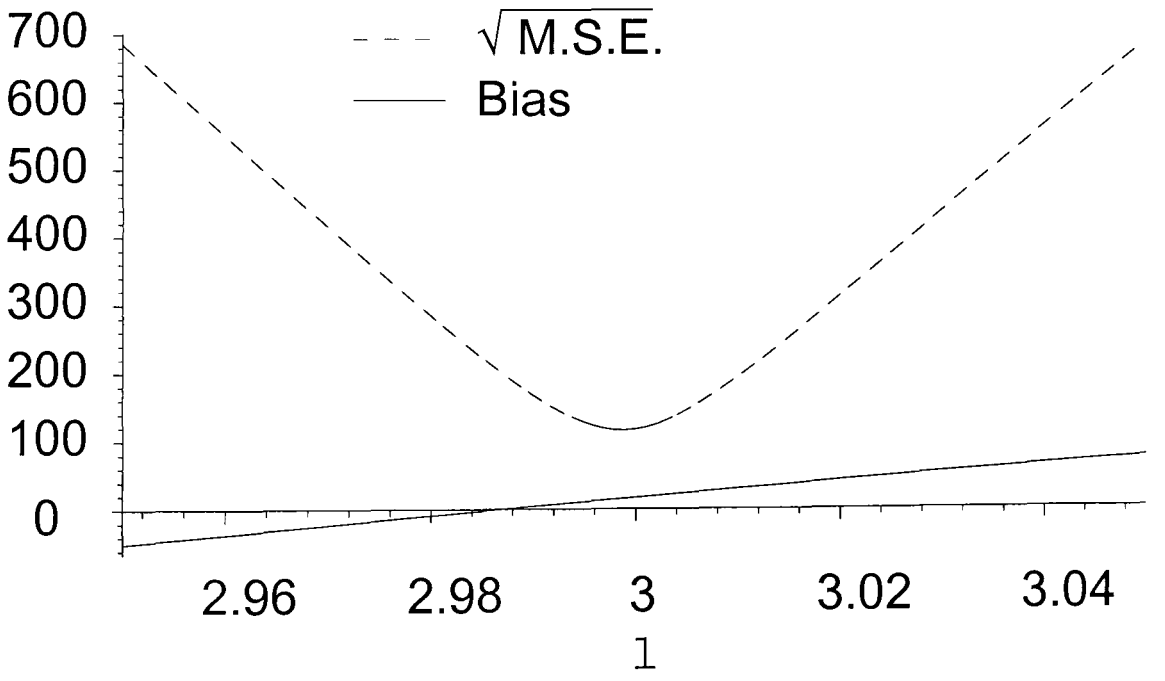


Figure 8.7: The bias and $\sqrt{}$(mean square error)

8.6 Miscellanea

The final section of this chapter contains three parts. The first part considers what happens to the precision or variance of the conditional probabilities C_{α_m} when data is observed. In the prior distribution the precision increases monotonically with level. However when we update to the posterior distribution the precision first *decreases* with the level, reaches a nadir and then increases, which is somewhat surprising at first sight. The second part of the chapter considers what happens if we transform a particular choice of Pólya tree by rotating the realised C.D.F. through a half-turn. It turns out that the Pólya tree distribution is invariant to such a transformation. The third part considers a generalisation of Pólya trees where the partitions are not fixed in advance but are randomly sampled. In this case the distribution of the C.D.F. is identical to the distribution of the inverse C.D.F.

8.6.1 Maximal variance

If we consider the variance of the conditional probabilities, Var(C_ϵ) as a function of the level, m, then for most choices of the sequence $\{\alpha_m\}$, Var(C_ϵ) will be monotonically decreasing, and hence have a maximum at the crudest partition. However, once data have been observed the posterior will not, in general follow this trend. The reason being that $\sum_i I(X_i \in \pi_\epsilon)$ will be smaller (non-increasing) as the level of π_ϵ increases. If a sequence of $\{\pi_\epsilon\}$ are nested then this will be true with certainty; if we are considering an arbitrary sequence with increasing level then this result will hold in probability as the level increases to infinity. Considering the updated posterior parameters, $\alpha_\epsilon + \sum_i I(X_i \in \pi_\epsilon)$, there is a trade-off between an increasing α_ϵ and a non-increasing $\sum_i I(X_i \in \pi_\epsilon)$, hence the observed Var$(C_\epsilon)$ will increase, reach a maximum, and finally decrease (so as to satisfy conditions for continuity). In computational terms, calculations can only be

performed to a finite level and we may not observe the decreasing variance.

A practical example is the choice $\alpha_m = cm^2$, hence the update posterior parameter is $cm^2 + \sum_j I(X_j \in \pi_{\epsilon_m})$. If it is assumed that the choice of prior distribution coincides with the data generating mechanism, then the updated parameter, with sample size n, has expectation

$$cm^2 + n2^{-m}.$$

This has derivative $2cm - n(\ln 2)2^{-m}$, so the maximum variance is achieved at the root of $2^{-m} = \frac{2c}{(\ln 2)n}m$. Examining figure 8.8 it is clear that increasing the sample size, n, or decreasing the constant, c, which reflects the overall, prior precision, will accentuate this effect.

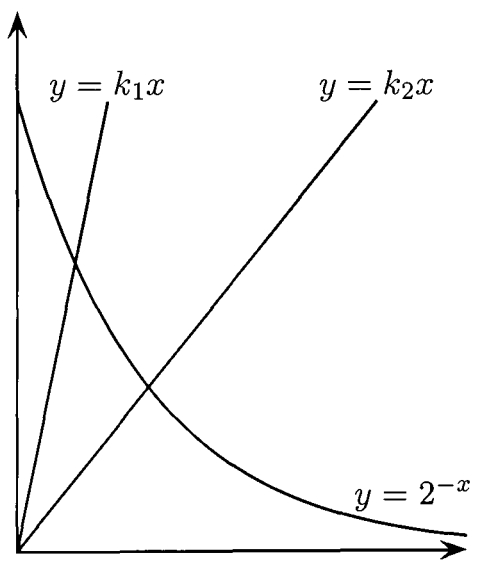


Figure 8.8: Geometric illustration

8.6.2 Two transformation theorems

The first theorem shows us that given a specific Pólya tree with the unit interval as the sample space then we can rotate the C.D.F. plane through π about the point $(1/2, 1/2)$

and the random C.D.F. will have the same distribution.

The second theorem shows that with a mixture of Pólya trees, on the unit interval, the inverse C.D.F. has the same distribution as the C.D.F.

Theorem 8.6.1 (Rotation). *Assuming the partition of Pólya tree is of the form*

$$\bigcup_{i,j}((j-1)2^{-i}, j2^{-i}],\ j = 1, \ldots, 2^i, i = 1, 2, \ldots,$$

and that the distributions associated with the C_ϵ are identical within each level, and are universally symmetric about 1/2, then for all values of $x, p, \in [0, 1]$,

$$\mathbb{P}\{\mathcal{P}(x) < p\} = \mathbb{P}\{\mathcal{P}(1-x) > 1-p\}.$$

Proof. Given that the distributions of C_ϵ are identical within levels this means that any permutation of the C_ϵ will give an identically distributed $\mathcal{P}$. In particular if we systematically swap *all* the 0s and 1s in the ϵ-suffix notation, at all levels, then we have effectively performed the transformation $X \mapsto 1 - X, \mathcal{P} \mapsto 1 - \mathcal{P}$, effectively rotating the $(X, \mathcal{P})-$plane through π about $(1/2, 1/2)$. □

Theorem 8.6.2 (Inverse). *If the partitions of a Pólya tree (defined on the unit interval) are not fixed, but randomly chosen, with*

$$\pi_{\epsilon 0} = (\inf \pi_\epsilon, (1 - Q_\epsilon) \inf \pi_\epsilon + Q_\epsilon \sup \pi_\epsilon]$$

and

$$\pi_{\epsilon 1} = ((1 - Q_\epsilon) \inf \pi_\epsilon + Q_\epsilon \sup \pi_\epsilon, \sup \pi_\epsilon],$$

where Q_ϵ and C_ϵ have independent, identical beta distributions, then

$$\mathbb{P}\{\mathcal{P}(x) < p\} = \mathbb{P}\{X(p) < x\}.$$

Proof. By definition the Q_ϵ and C_ϵ have independent, identical beta distributions, hence they can be exchanged at all levels, and values, of ϵ and the resulting '*mixed*' Pólya tree will have the same distribution. However this exchange is equivalent, on a macroscopic scale, to making the transformation $X \mapsto \mathcal{P}, \mathcal{P} \mapsto X$, a reflection about the diagonal. This is equivalent to obtaining the inverse C.D.F. □

Note that, despite the abundance of symmetry in the definition of a Pólya tree defined on $[0,1]$ with a fixed set of partitions $\{\pi_\epsilon\}$, the inverse C.D.F. is not, in general, identically distributed to the C.D.F. As a counter example consider values $p = 1/2,\ x = 1/4$. Without any assumptions, the event $\{X(1/2) < 1/4\}$ is equivalent to $\{\mathcal{P}(1/4) > 1/2\}$ or

$$\{C_\Omega C_0 > 1/2\}. \tag{8.10}$$

If the inverse C.D.F. were identical to the C.D.F. we would have $\mathbb{P}(\mathcal{P}(1/2) < 1/4) = \mathbb{P}(X(1/2) < 1/4)$. But the event considered on the left hand side is equivalent to

$$\{C_\Omega < 1/4\}. \tag{8.11}$$

Numerical calculations, or simulations, easily show that the probabilities of (8.10) and (8.11) are not, in general, equal.

8.7 Summary

In this chapter I have considered how to interpret the strength of a Pólya tree prior distribution and the practical issues of integration with the Pólya tree as the measure. I have proven, with a particular choice of parameters, that the prior density is distributed as the expected prior density multiplied by a gamma-distributed random variable. Limited results on the distribution of the posterior distribution have been obtained. I have

considered how to choose a finite range of integration to approximate an integral over $\mathbb{R}$ so that any error is minimised.

Chapter 9

Analysis of prostate cancer data set

9.1 Origins of the data

The data are published in Andrews and Herzberg (1985) and are down-loadable from

`http://lib.stat.cmu.edu/datasets/Andrews/T46.1`

They consist of patient records from a randomised clinical trial for patients with stage 3-4 prostatic cancer. There were four treatments: placebo, 0.2mg , 1.0mg, and 5.0mg of oestrogen. The endpoint considered was the survival time and survival status, for which there were 10 possible, and mutually exclusive events as shown in table 9.1, which also tells us that there were 506 patients in the trial.

code	description	count
0	alive	150
1	dead from prostatic cancer	130
2	dead from heart or vascular disease	96
3	dead from cerebrovascular disease	31
4	dead from pulmonary embolus	14
5	dead from other cancer	25
6	dead from respiratory disease	17
7	dead from other specific non-cancer cause	29
8	dead from unspecified non-cancer cause	7
9	dead from unknown cause	7
	TOTAL	506

Table 9.1: Table of endpoints

Along with survival time/status, tumour stage and treatment, there were recorded twelve pretreatment covariates: age, weight index, exercise performance rating, history of cardio-vascular disease, systolic blood pressure, diastolic blood pressure, electrocardiogram code, serum haemoglobin, size of primary tumour, combined index of tumour stage and histological grade (Gleason grade), serum, prostatic acid phosphatase in King-Armstrong units, bone metastases. The data were originally analysed in Byar and Corle (1977), and, Byar and Green (1980), which give further details of the variables recorded.

From the statistical perspective this is an interesting data set as the end-point is clearly a competing risks situation. Furthermore we would *a priori* expect there to be positive correlations between cancer, and cardiovascular disease, say, whether on medical grounds or due to an unobserved confounding variable such as a history of smoking.

Moreover, the data set is of a reasonable size to induce reasonable statistical power in any inference and we have a several continuous, and plausibly relevant covariates which allows us to make use of the identifiability results discussed in chapter 6.

9.2 Statistical analysis

The point of this chapter is to demonstrate the techniques which have been developed in the preceding chapters. A precursor to any formal analysis is to note that aside from the status codes of *alive* and *dead from prostatic cancer* all the remaining codes have 226 patients in total, and hence any finer stratification is unlikely to have sufficient statistical power. In addition, the focus of the trial was on the treatment of prostatic cancer, so with these two considerations any further analyses will just use three possible status codes: dead from prostatic cancer, dead from other causes, alive. The alive status will be considered, where appropriate, as an uninformative censoring.

The preliminary analysis will be in 9.3 where the tool of the crude incidence function will be used, along with the simultaneous confidence bands presented in chapter 2, as a means to compare the treatments. The second part, 9.4 will develop a regression model using the techniques described in chapter 7, where the data is augmented as in Lunn and McNeil (1995), but then is modeled using a gamma or a log-normal frailty distribution to try to capture any dependency between the two causes of failure. The final section, 9.5 will consider whether the assumption of the gamma or log-normal is appropriate by the use of Pólya tree theory as considered in chapter 8 to model the frailty distribution, and in addition will translate the model of 9.4 into a fully Bayesian framework.

9.3 Preliminary Analysis

The first analysis is presented in figure 9.1 and it shows the crude incidence function, $Q_k(t) = \mathbb{P}(\text{survival time} < t, \text{cause} = k)$, for all nine non-censoring causes of failure.

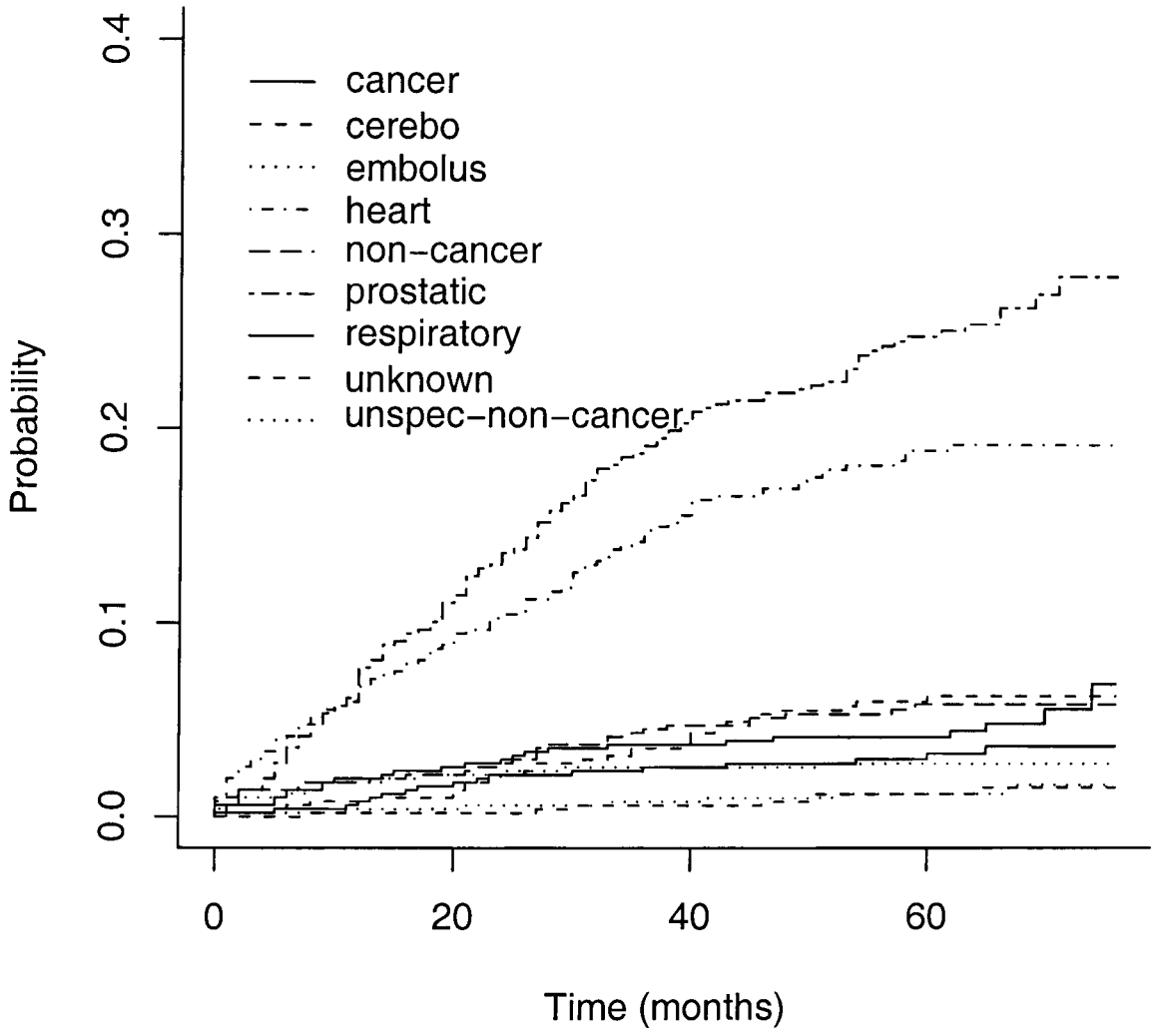


Figure 9.1: Point estimates of the crude incidence function for all causes

It clearly shows that prostatic cancer, and cardio-vascular disease, are the com-

monest observed causes of death in the trial. In figure 9.2 we can see that if we group together the remaining causes of failure then we have three causes which are all within the same order of magnitude in terms of mortality rates. The North-West graph shows the three causes together, and the remaining three graphs show the causes separately with their 95 % Hall-Wellner confidence bands.

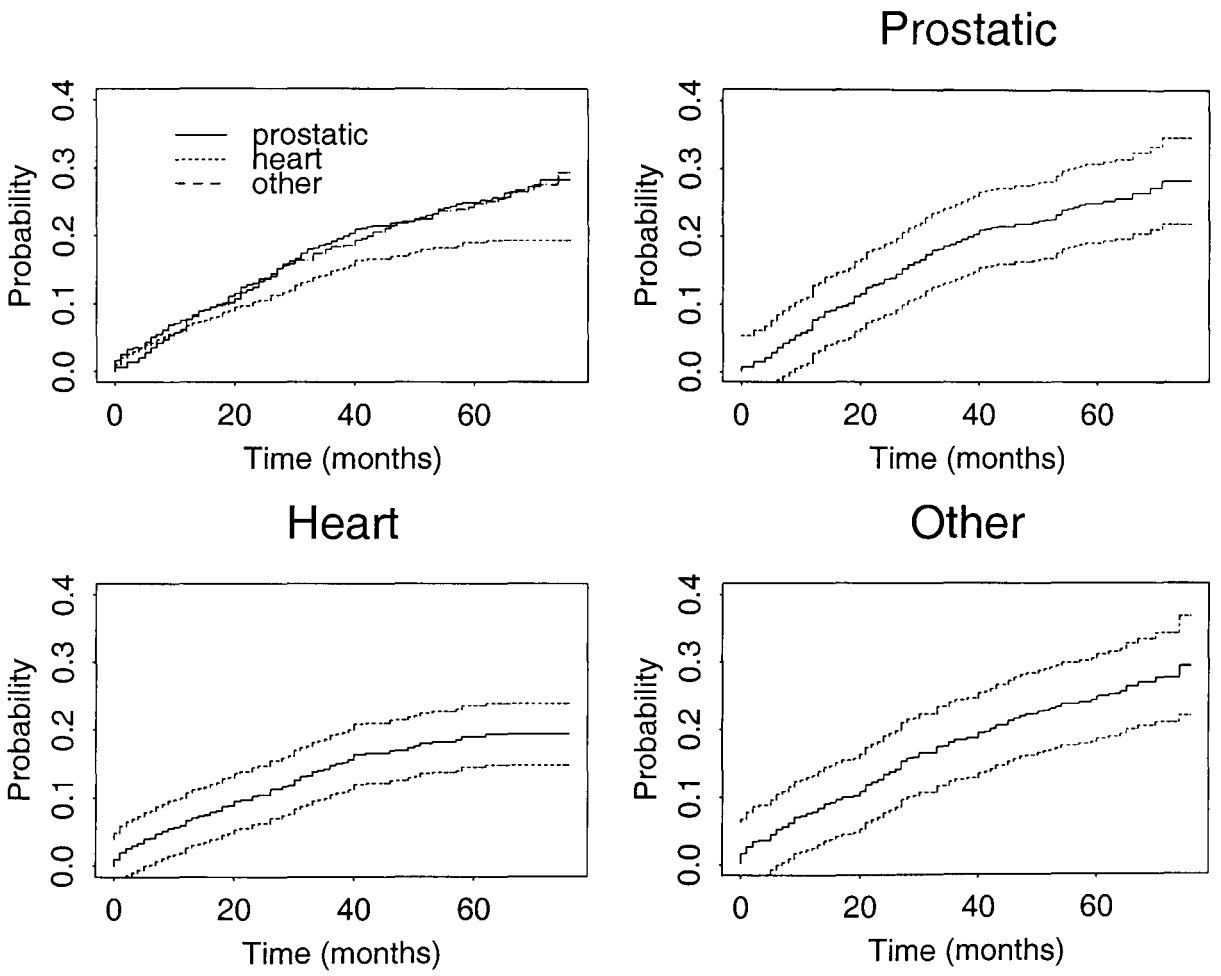


Figure 9.2: Crude Incidence for the individual causes with confidence bands

The next step is to compare the the treatment groups in terms of their effect on crude incidence, and on the cause specific hazard. In general we must consider separately the two sets of basic null hypotheses; H_a : a common cause-specific hazard

between the treatments; H_b : a common crude incidence between the treatments. This is because of the relationship

$$Q_k(t) = \int_0^t S(u-)d\Lambda_k(u),$$

where S is the overall survival, and $d\Lambda_k$ is the cause-specific hazard for cause k. It is possible for the S and the $d\Lambda_k$'s to be different between treatment groups but still produce a common crude incidence, and vice versa.

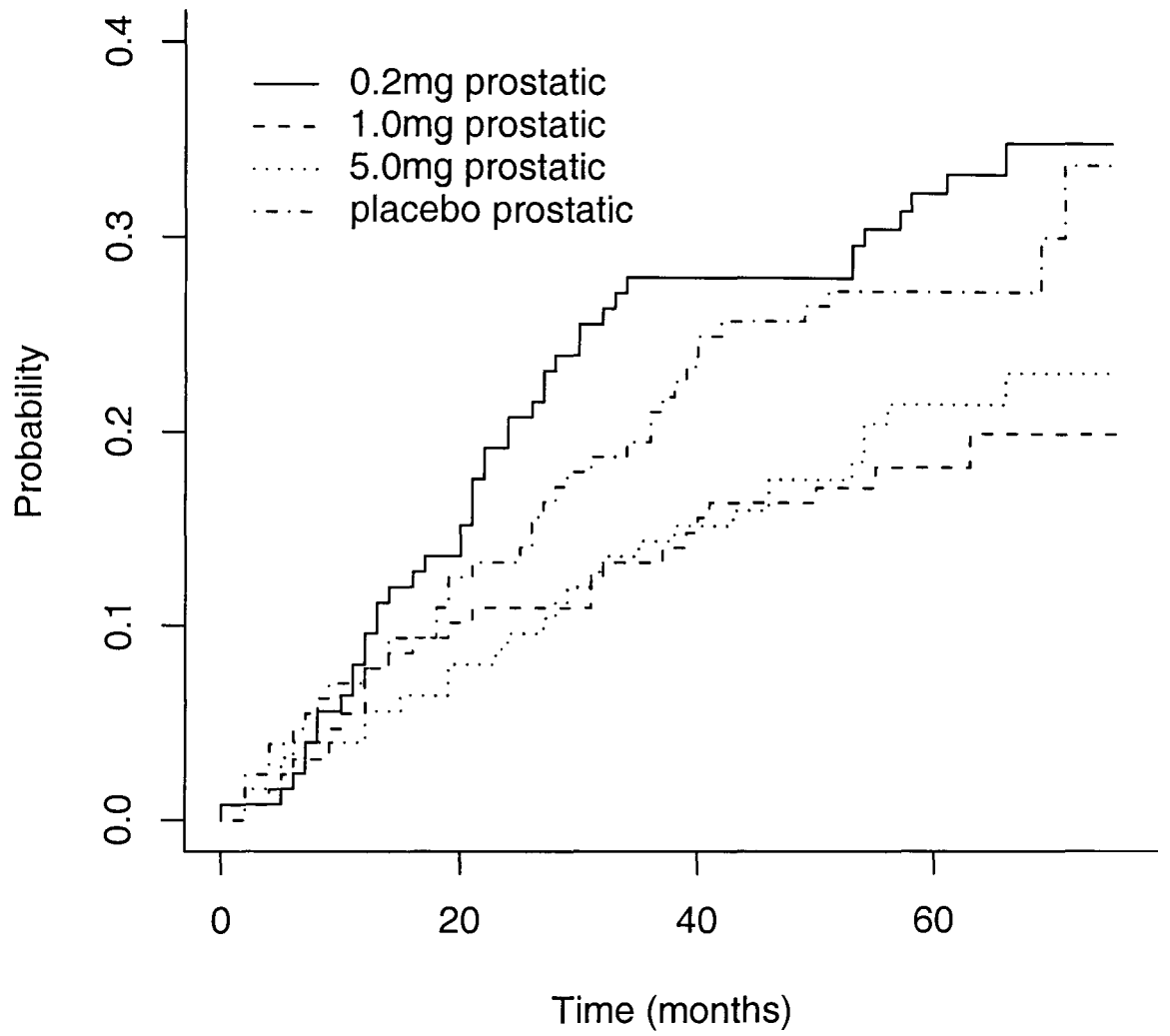


Figure 9.3: Crude Incidence for prostatic cancer stratified by treatment

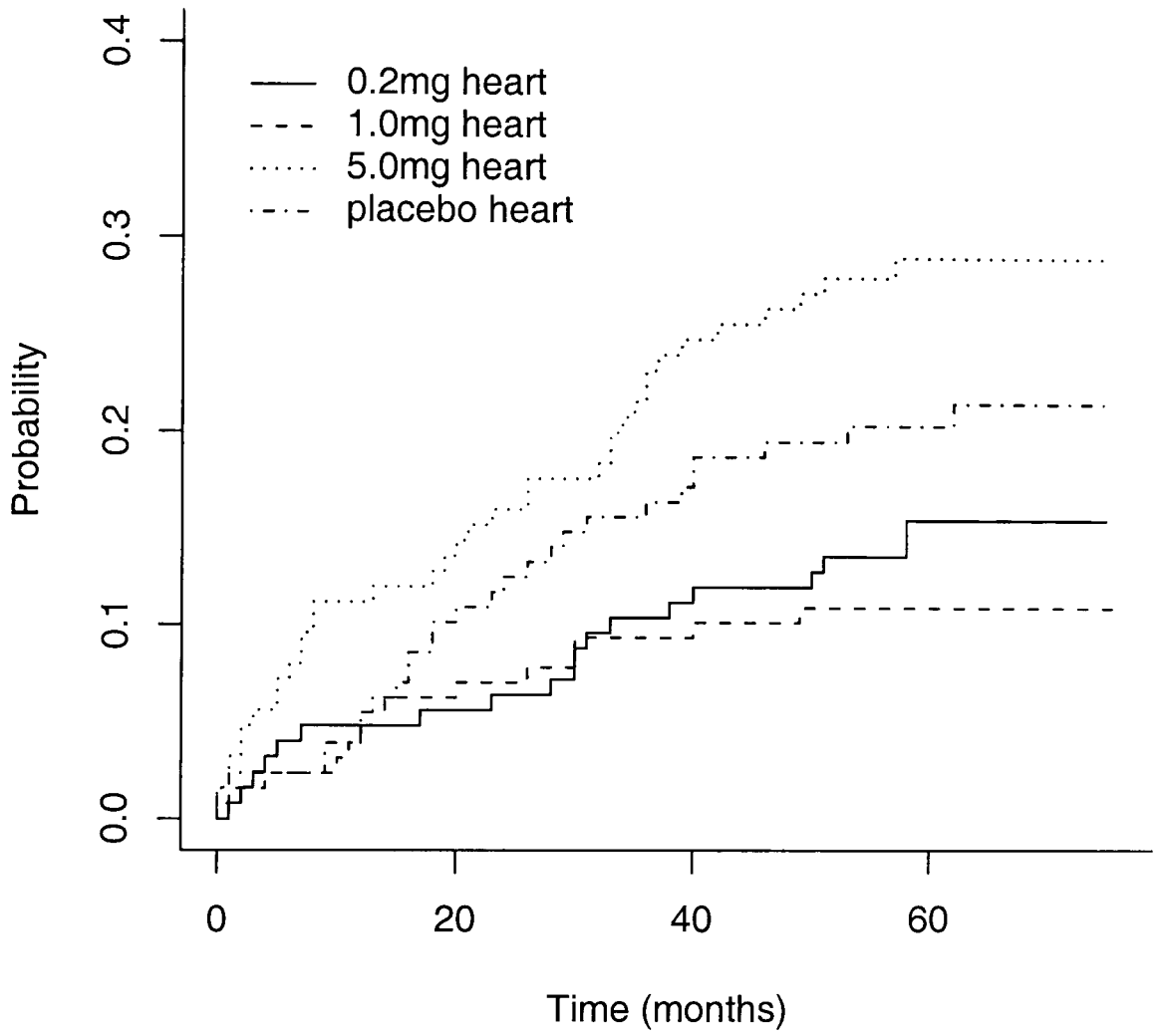


Figure 9.4: Crude Incidence for heart stratified by treatment

Figures 9.3 and 9.4 display the crude incidence curves for the separate treatment groups. The crude incidence curves are formally compared in table 9.2 which uses the chi-squared test developed in Gray (1988). To compare the cause-specific hazards we can use the well known log-rank test, which is also in table 9.2. Both tests were subject to a weighting parameter, ρ, which was considered with two values: $\rho = 0$ uses a constant

weight through out the time period, whereas $\rho = 1$ weights the data proportionally to $S(t)$, thus giving more weight to earlier observations.

	$\rho = 0$		$\rho = 1$	
cause	Q_k	Λ_k	Q_k	Λ_k
cancer	0.1370	0.2122	0.1350	0.2073
cerebo	0.5700	0.8027	0.5750	0.8037
embolus	0.1590	0.1431	0.1550	0.1388
heart	0.0021	0.0028	0.0022	0.0031
non-cancer	0.4060	0.3522	0.4020	0.3523
prostatic	0.0357	0.0355	0.0353	0.0374
respiratory	0.5570	0.6211	0.5550	0.6215
unknown	0.6850	0.5967	0.6860	0.6000
unspec-non-cancer	0.6810	0.6261	0.6810	0.6272

Table 9.2: Table of p-values comparing the treatment groups

The tests indicate that there is strong evidence of a treatment effect on heart/vascular mortality, and some evidence of an effect on prostatic cancer mortality. The p-values are unaffected by which test we perform and this is probably due to the overall survival being unaffected by the treatment groups, as shown in figure 9.5 with p-values of 0.0426 ($\rho = 0$) and 0.155 ($\rho = 1$) in the standard log-rank test. In these circumstances having common Λ_k's *is* equivalent to having common Q_k's.

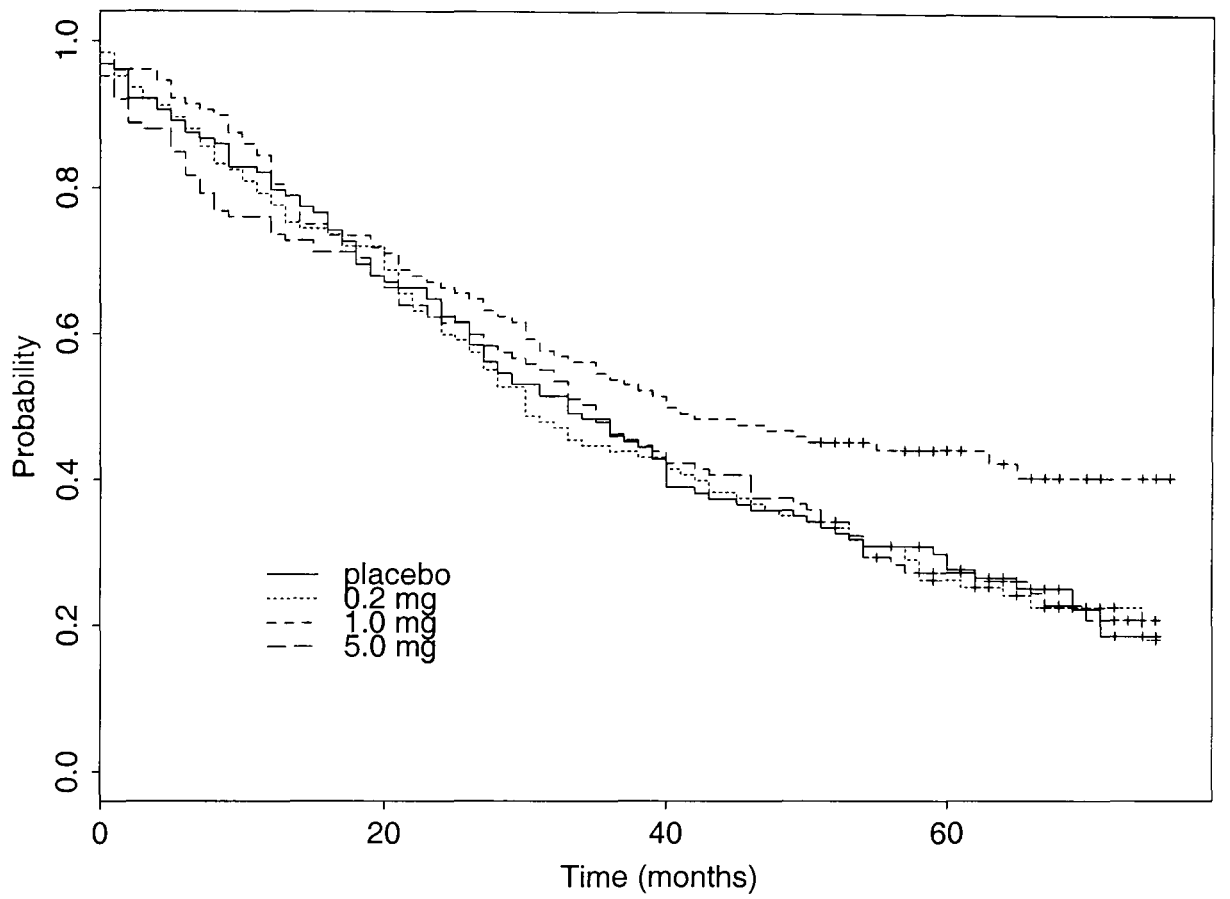


Figure 9.5: Overall survival by treatment group

9.4 Classical Frailty Regression

The basic framework used here is an extension of the Cox proportional hazards model (Cox 1972), as discussed in chapter 7. In brief, the model uses the counting process

$N_{ij}(t), Y_i(t)$, where ,

$$Y_i(t) = I(\text{individual } i \text{ at risk at time } t),$$

$$N_{ij}(t) = I(\text{individual } i \text{ failed of cause } j \text{ at time } T < t),$$

$$\mathbb{P}(dN_{ij}(t) = 1) = Y_i(t) \exp(\beta_j^T x + b_i)\lambda(t),$$

$$b_i \sim N(0, \sigma^2) \text{ or } \exp b_i \sim \Gamma(1/\sigma^2, 1/\sigma^2).$$

In practice, this model is fitted using first using the data augmentation technique of Lunn and McNeil (1995), and then instead of assuming independence, using the frailty software in S-plus, as described in Therneau and Grambsch (2000), to model a dependence between the causes of failure. This effectively changes the format of the data from a $k \times n$ matrix of counting processes to a kn vector of (X, δ) survival pairs. The fact that the model uses a β_j, rather than a common β means that every fixed effect has an interaction with the cause covariate. Each individual in the original, un-augmented, data set now spawns k individuals which form a cluster, or family, represented by one unobserved frailty, b_i.

As the primary interest was in prostatic cancer, the causes were now condensed into two possibilities, along with censoring: dead from prostatic cancer, dead from other causes, alive. The treatment was condensed into two levels: placebo & 0.2 mg, and, 1.0 mg & 5.0 mg . Other covariates which were included in the final model were: serum haemoglobin levels in grams per 10 ml; tumour size in cm^2; tumour step – an indicator of a combined index of tumour stage and histological grade exceeding 11; cardio – an indicator of a history of cardiovascular disease; age in years; bone metastases – an indicator variable; performance 1 – an indicator that the patient was confined to bed less than 50 % of the time; performance 2 – an indicator that the patient was confined to bed more than 50 % of the time. A summary of the marginal distribution of the covariates

is given in table 9.3. The continuous covariates all had a bell-shaped distribution, with the exception of tumour size, which was all positive and peaked at zero and would be better described with an exponential, rather than a normal, distribution.

variable	mean or proportion	s.d.
haemo	134	19.5
tumour size	14.6	12.3
tumour step	0.480	–
cardio	0.424	–
age	71.4	7.08
bone metastases	0.162	–
performance 1	0.0734	–
performance 2	0.0299	–

Table 9.3: Distribution of covariates

The form and choice of covariates was determined by a two-step procedure. The first step was of an exploratory nature and consisted of taking the martingale residuals from the null model with just one fixed effect for causes, and plotting them against any potential covariates as in Fleming and Harrington (1991) sections 4.5 and 4.6 –for example the cut point of 11 in tumour step was chosen by eye in this way. The second step was a more formal nested hypothesis testing procedure based on likelihood ratio statistics. A model was considered which had separate baseline hazards for the two causes, but was found to be unnecessary.

9.4.1 Gamma Frailty

For the gamma-frailty model the estimates and p-values are presented below. The principal effects, as summarised by the preliminary analysis, are in table 9.4 along with the non-significant effect of the gamma frailty. The effect if the covariates on the prostatic causes of failure is in table 9.5 and the effect on the 'other' causes of failure is in table 9.6.

variable	coef	s.e.	χ^2	DF	p-value
cause='other'	-4.34	2.03	4.60	1	0.032
Rx=oestrogen	-3.95	1.72	5.24	1	0.022
cause:Rx	0.936	2.49	0.14	1	0.707
frailty			0.02	0.02	0.670

Table 9.4: Estimates and p-values for the gamma frailty model: main cause/treatment effects

The crude incidence curve for 'other' causes will be an average of the 'heart' and the remaining non-prostatic causes which will be weighted towards the 'heart' curve due to patient numbers. As shown in figure 9.2 the prostatic curve is higher, indicating that there is a lower risk of failure from 'other' which is reflected in the negative coefficient. The treatment (labeled as Rx with two levels: oestrogen, control) lowers the risk and its effect, without accounting for interactions with covariates, does not significantly vary between the causes. The frailty terms are not statistically significant thus indicating, under an untestable assumption, that there is no dependency of this form. The estimated variance of the gamma distribution is 5e-5, and the profile log-likelihood gives a 95% confidence interval of [0, 0.429].

variable	coef	s.e.	p-value
haemo	-0.011	0.00511	0.032
tumour size	0.0397	0.00683	6.07E-9
tumour step	1.94	0.29	2.43E-11
cardio	-0.142	0.206	0.490
control : age	-0.0422	0.0153	0.006
Rx : age	-0.00295	0.0188	0.875
control : bone metastases	0.031	0.33	0.925
Rx : bone metastases	0.857	0.311	0.006
control : bed<50%	-0.741	0.48	0.122
Rx : bed<50%	0.445	0.54	0.410
control : conf/bed>50%	1.46	0.443	0.001
Rx: conf/bed>50%	0.866	0.549	0.115

Table 9.5: Estimates and p-values for the gamma frailty model: covariate effects for 'prostatic' cause

variable	coef	s.e.	p-value
haemo	-0.00475	0.00381	0.212
tumour size	-0.00503	0.00712	0.480
tumour step	-0.194	0.162	0.230
cardio	0.678	0.142	1.79E-6
control : age	0.0336	0.0181	0.064
Rx : age	0.078	0.0167	2.94E-6
control : bone metastases	1.07	0.305	4.67E-4
Rx : bone metastases	-0.904	0.378	0.017
control : bed<50%	0.155	0.34	0.647
Rx : bed<50%	1.23	0.316	1.04E-4
control : conf/bed>50%	-0.994	1.01	0.325
Rx : conf/bed>50%	-0.169	1.04	0.871

Table 9.6: Estimates and p-values for the gamma frailty model: covariate effects for 'other' causes

Comparing the two sets of coefficients, within 'prostatic' cause, haemoglobin has a statistically significant association with increased survival. Tumour size and tumour step are significant only for the prostatic cause, and both decrease survival. Age is significant for both causes; it *decreases* the risk in the prostatic control group (with non-significance in the treatment group), in the 'other' cause it has approximately twice the effect in increasing the log-odds in the treatment group compared to the control group (0.0780 versus 0.0336). Metastases only has a significant effect in the treatment group for prostatic which decreases survival, but in the 'other' cause it increases risk in the control group and decreases risk in the treatment group by a similar magnitude.

The performance indicator, is significant at level 2 for the control group in prostatic, and at level 1 for the treatment group in 'other', both decrease survival.

9.4.2 Log-normal Frailty

As an alternative model the random effects, on the multiplicative scale, were modeled with a normal distribution. The estimates of the fixed effects were very close to those of the model with a gamma frailty. This is shown in figure 9.6, where the two point estimates form a co-ordinate, and the crosses give error bars equal to twice the standard deviation. The points are very close to the dashed line of equality, and the error bars in each direction are of similar length.

The estimate of the variance of the log-normal frailty distribution is 0.527. This gives a starkly different conclusion about the question of dependence, as it is on the borderline of significance. To compare with table 9.4, there is a χ^2 statistic of 140.35 on 116 degrees of freedom giving a p-value of 0.062.

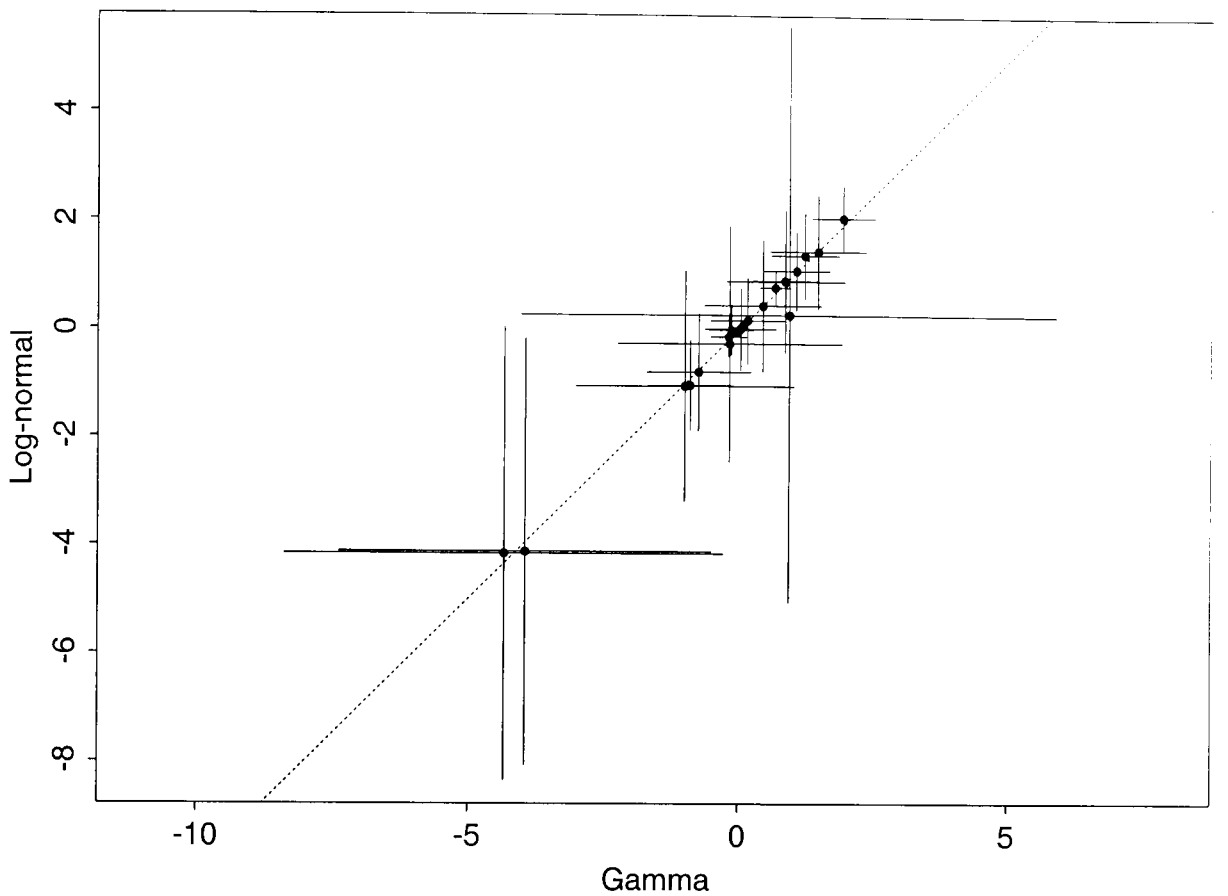


Figure 9.6: Comparison of fixed effects estimates

It is difficult to investigate this difference any further, due to the complexity of the S-plus code. The likelihood can be calculated explicitly in the case of the gamma frailty, this was then maximised using a general maximisation routine and was found to agree with the analysis. However in the case of the log-normal frailty, the basic idea used is a Newton-Raphson scheme which requires the inverse Hessian matrix. Because this matrix is nearly diagonal in the block associated with the frailties, the code makes this approximation and greatly reduces the computation time. It is possible to override this and the estimates are in broad agreement with each other. However it was not

possible to extract a profile log-likelihood to judge the estimate of the variance of the frailty distribution.

9.5 Pólya tree frailty analysis

The next modification to the model considered was to replace the parametric assumptions about the form of the random effects distribution, with the infinite-parametric framework of a Pólya tree. The only practical way to perform this was using M.C.M.C. simulation techniques, which permits a move to the full Bayesian framework. The partial likelihood was used instead of the full likelihood in computing the posterior distribution. This is equivalent to assuming that we only observe the order, and not the exact times, of the events. It avoids the extra computational burden of estimating the baseline hazard.

The priors for the fixed effects were all reference priors of $N(0, 10^2)$. The Pólya tree had its median constrained to be zero so as to give identifiability. The partitions for the Pólya tree at level m were $(q_k, q_{k+1}]$, where $\mathbb{P}(Z < q_k) = k2^{-m}$, and Z follows a standard normal; and the prior put on the probability, p, associated with each interval at level m was a $\beta(2^{m-1}/100, 2^{m-1}/100)$. This effectively centres the Pólya tree's prior on $N(0, 1)$, but says that the ratio of the 'realised to expected' density at any point (marginally) follows a gamma distribution, with mean 1 and variance 100.

The sampling procedure used was very similar to that described in Raftery and Lewis (1996) which is the Metropolis-Hastings algorithm applied to one parameter at a time using a symmetric, uniform random walk as the proposal distribution. For the fixed effects, a linear transformation of the parameters, $\beta' = A^{-1}\beta$ was sampled where A was such that $(XA)^T(XA) = I$ (the Gram-Schmidt orthonormalisation). Loosely, this reduces the correlation in the posterior distribution of the components of β' and this improves the mixing and convergence of the chain. The bounds of the uniform

proposal distribution were chosen to give a standard deviation equal to $2.3s.d.(\beta_j|\beta_{-j})$ where $s.d.(\beta_j|\beta_{-j})$ is the residual standard deviation of regressing β_j on the remaining parameters. This calculation was iterated three times, until it had stabilised. The length of the final simulation was 12,000 and was performed, in parallel, three times from different starting values and the convergence diagnostics from the CODA package (Best, Cowles and Vines 1995) indicated satisfactory convergence and mixing.

9.5.1 Results

The traces of one of the chains, along with a kernel density estimate of the posterior distribution (starting from iteration 2000) is shown in figure 9.7.

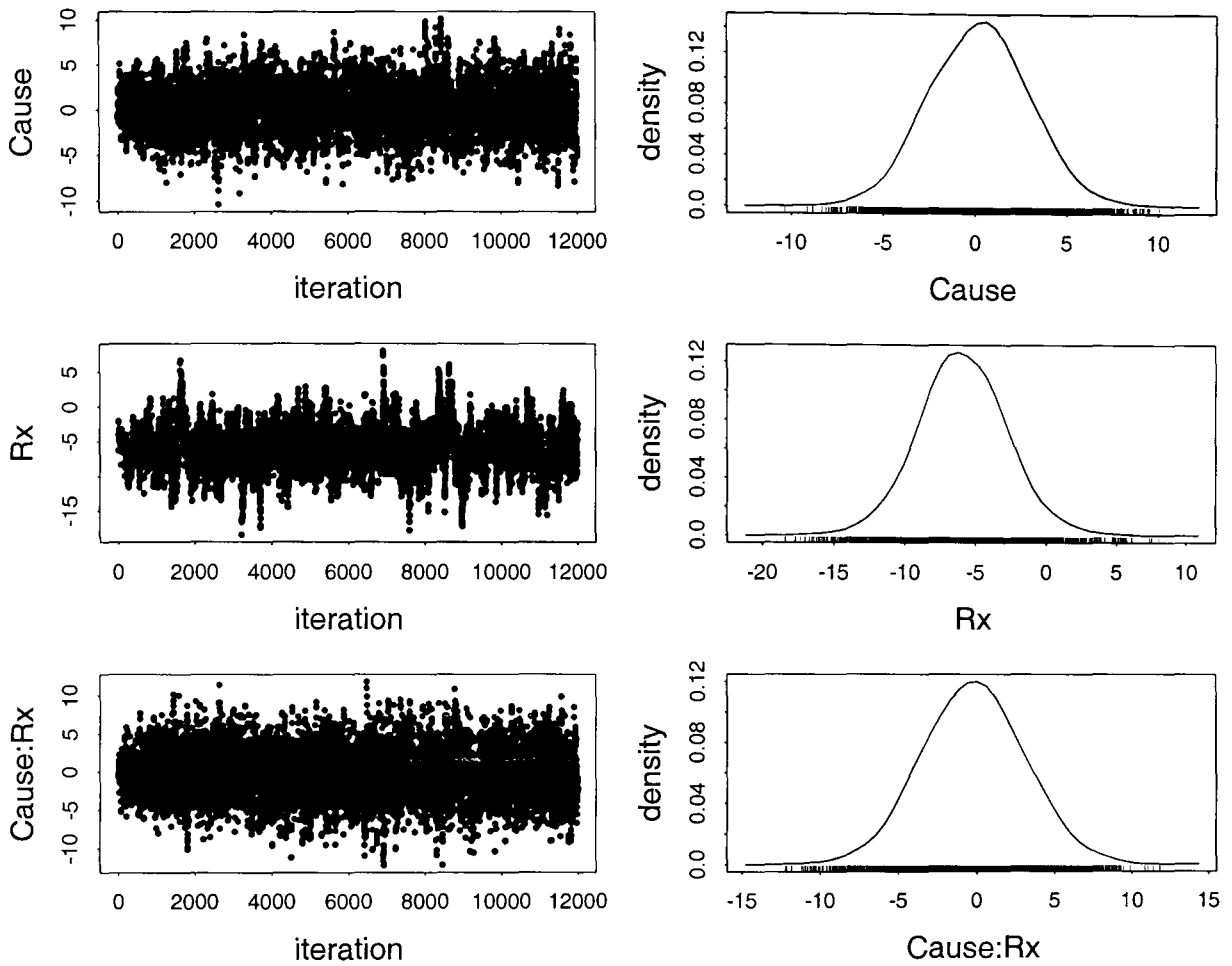


Figure 9.7: Main effects posteriors

The covariate effects for prostate are shown in figures 9.8 and 9.9.

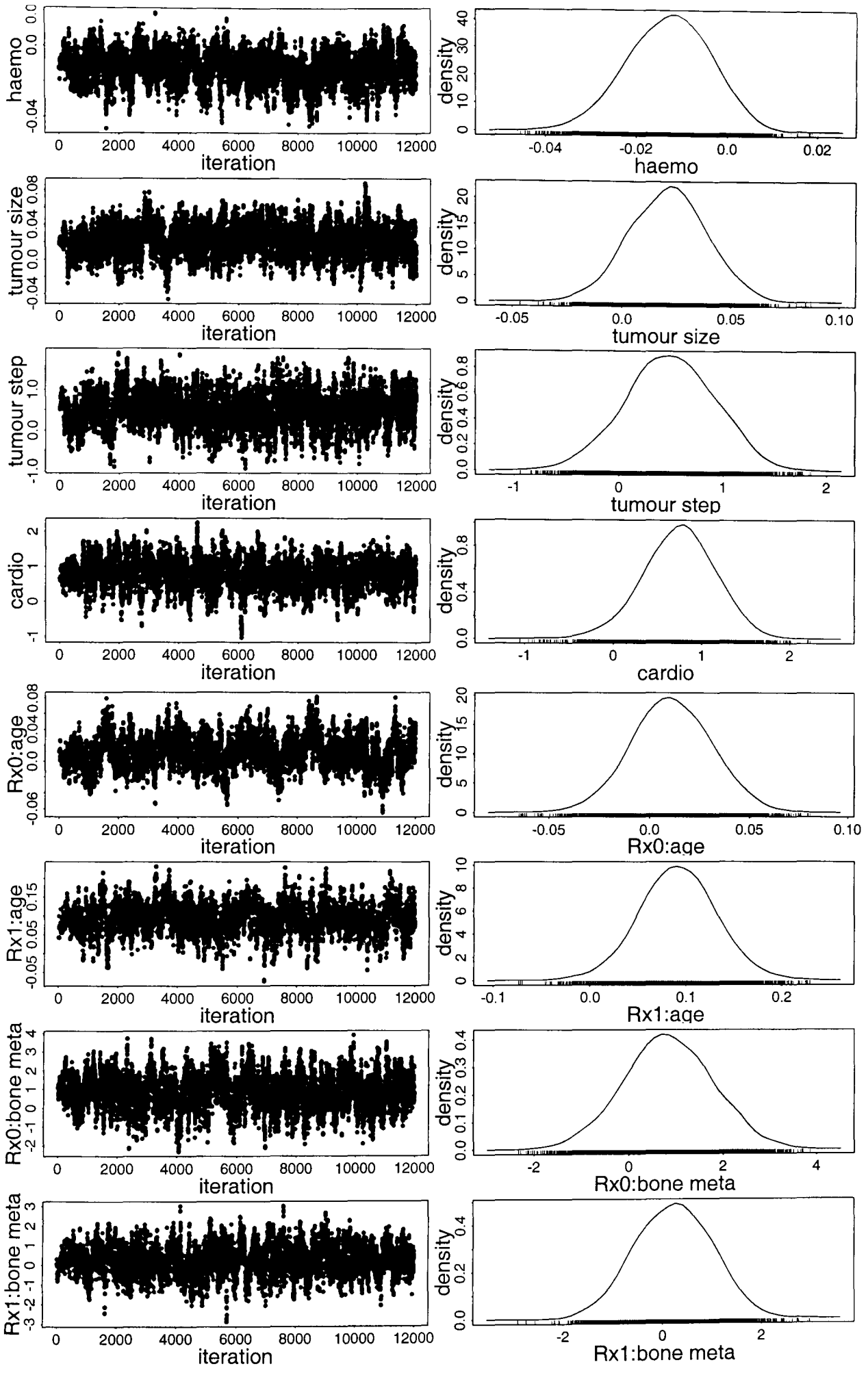


Figure 9.8: Prostate effects posteriors

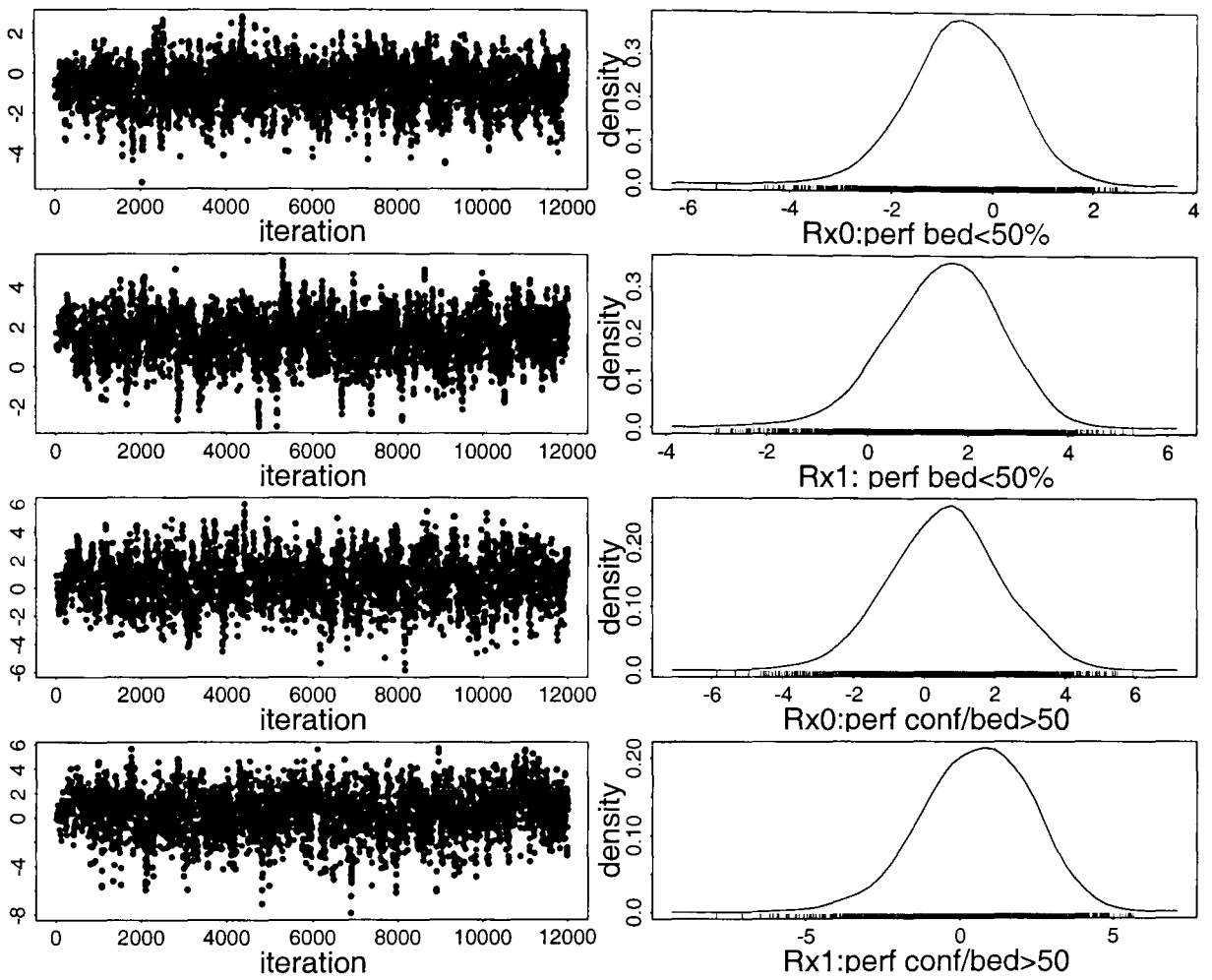


Figure 9.9: Prostate effects posteriors (continued)

The equivalent for 'other' cause is shown in figures 9.10 and 9.11.

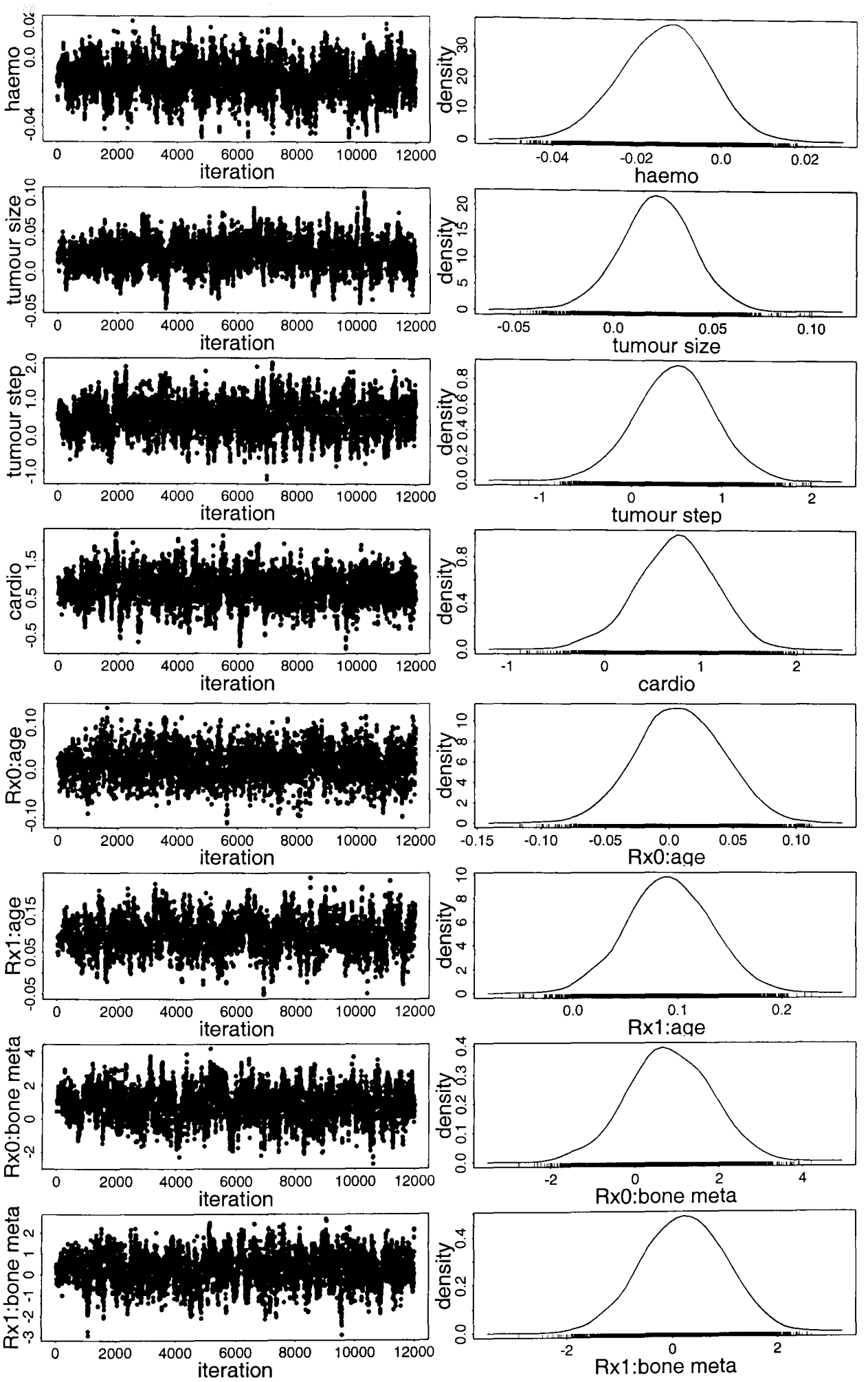


Figure 9.10: 'Other' effects posteriors

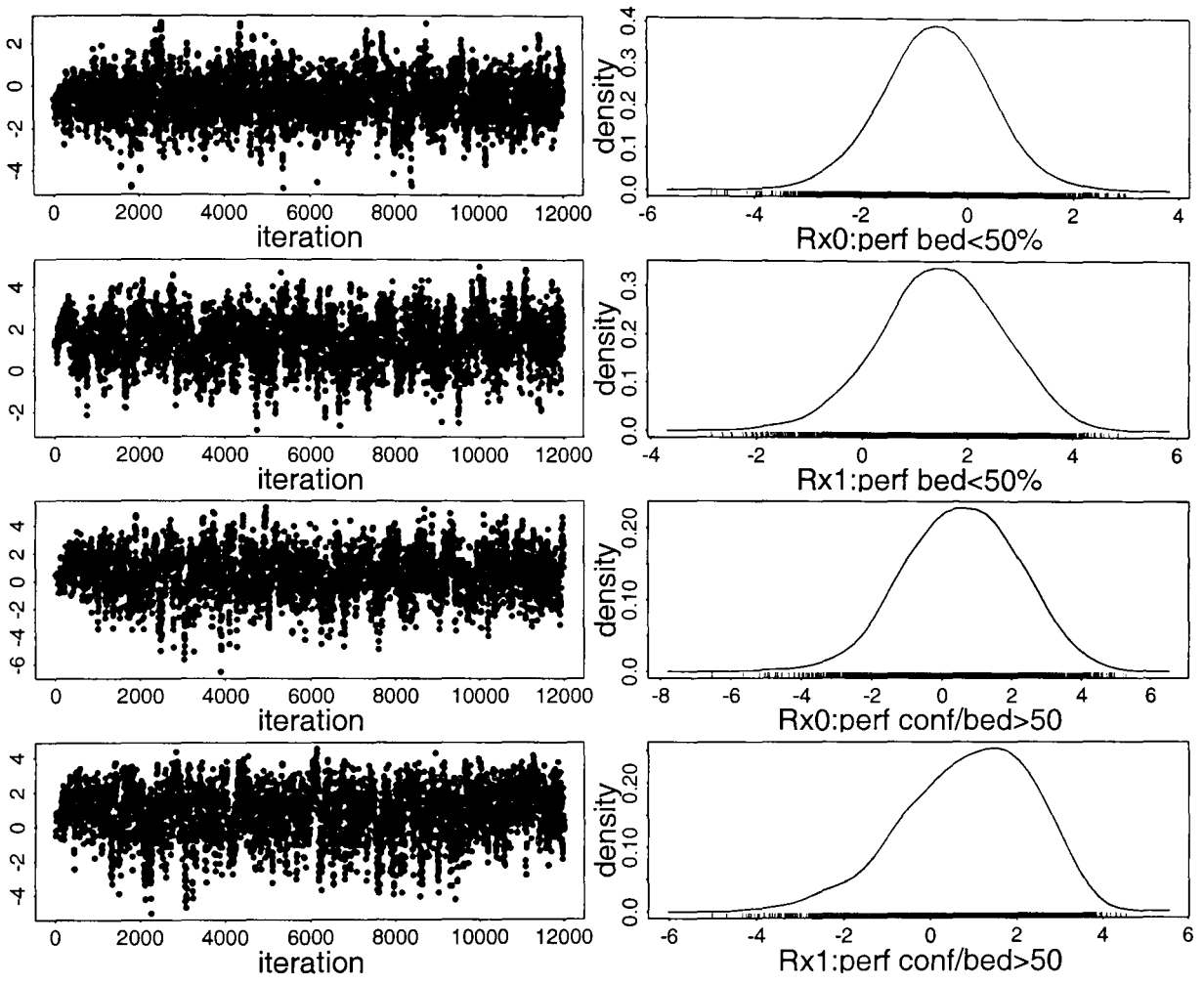


Figure 9.11: 'Other' effects posteriors (continued)

A Pólya tree was used to model the frailty distribution. The tree was constrained to have a median of zero with probability 1. To display the random distribution visually, each iteration of the M.C.M.C. re-samples the conditional probabilities which define the tree. With these it is possible to calculate the cumulative density at any set of point desired–effectively giving a realisation of the random cumulative density function. This was done at unit intervals between -3 and +3. With these the resulting cumulative density function is approximated with a piecewise linear function. Approximated pointwise confidence intervals can then be formed by discarding the most extreme proportion of

the estimates at any point. This is presented in figure 9.12 which give 90% confidence intervals, along with the empirical mean of the estimates.

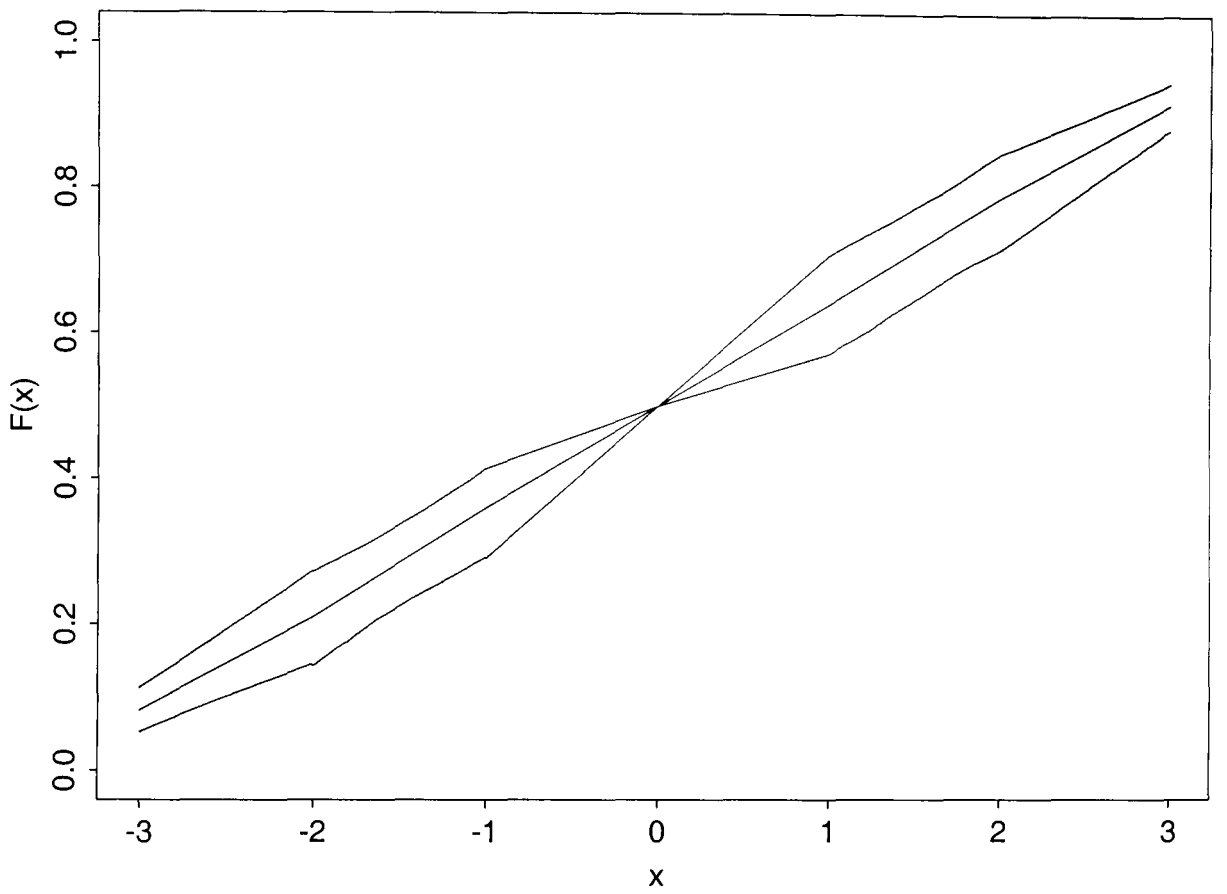


Figure 9.12: 90% pointwise confidence intervals for the C.D.F.

9.5.2 Sceptical Prior analysis

Ideally, to perform an Bayesian analysis which is not using a reference prior we would like to consult expert opinion and the existing literature. Unfortunately the current literature on clinical trials of prostate cancer are of poor statistical quality. The most recent large scale randomised control trials published by the European Organisation for Treatment and Research into Cancer (EORTC trials 30843 & 30853) (Sylvester, Denis, de Voogt

et al. 1998, de Voogt, Studer, Schröder, Klijn, de Pauw, Sylvester et al. 1998), have performed a Cox proportional hazards regression analysis, but as an intermediary step to forming a prognostic scoring system which is assessed by stratifying patients according to the score and performing log-rank test on the resulting groups. No confidence intervals or p-values were reported on the relevant coefficients. All that can be inferred from the literature is that roughly the same covariates are considered as having a potential prognostic influence.

In the light of this we have effectively had to pluck a prior out of thin air. If one follows the philosophy advised in Spiegelhalter, Myles, Jones and Abrams (2000), of considering the posterior to be a function of the prior and it being the responsibility of the analyst is to display this function, then we can justify asking the question, "What extremal priors will give substantial posterior mass near zero?"

To answer this we consider the canonical Normal-Normal prior-posterior case. If the data follow a normal distribution with mean μ_{obs} and precision (inverse of the variance) γ_{obs}, and the prior distribution is normal with mean and precision (μ_0, γ_0), then the posterior has parameters $((\gamma_0\mu_0 + \gamma_{\text{obs}}\mu_{\text{obs}})/(\gamma_0 + \gamma_{\text{obs}}), \gamma_0 + \gamma_{\text{obs}})$. So if we take the step of saying the maximum likelihood estimators of the fixed effects coefficients give us values of $(\mu_{\text{obs}}, \gamma_{\text{obs}})$, then what values of (μ_0, γ_0) will give posterior mass near zero? If we assume that μ_{obs} was positive then we formulate this question, in terms of Z-statistics, as finding the region in the (μ_0, γ_0)-plane such that,

$$\frac{\gamma_0\mu_0 + \gamma_{\text{obs}}\mu_{\text{obs}}}{\sqrt{\gamma_0 + \gamma_{\text{obs}}}} < \Phi^{-1}(1-\alpha), \tag{9.1}$$

where α is typically chosen to be 0.05 . If it was the case that $\mu_{\text{obs}}\sqrt{\gamma_{\text{obs}}} > \Phi^{-1}(1-\alpha)$ then this region is similar to the area below the line in figure 9.13

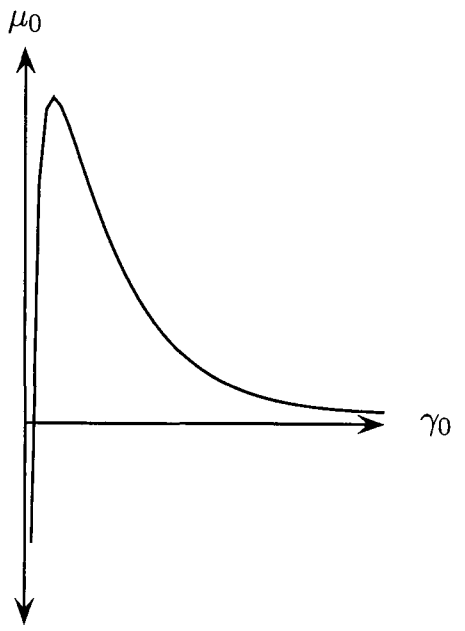


Figure 9.13: Region which gives a 'non-significant' posterior

If it is the case that $\mu_{\text{obs}}\sqrt{\gamma_{\text{obs}}} < \Phi^{-1}(1-\alpha)$, then a reference prior will give substantial mass near zero. This graph can be interpreted as a play-off between the influence of the prior's mean, and its precision. If the precision is close to zero then the mean needs to be a large negative number to counter the influence of the likelihood and drag the posterior distribution towards zero. As the precision increases, then the prior gains in influence on the posterior and as such the prior's mean needs to be close to zero to achieve the same effect on the posterior.

Our sceptical priors were chosen according to this argument. In particular, the apex of the curves in figure 9.13 were used as this represents the point of 'equal influence' between the prior's mean and precision. The location of the apex can be found in an analytical form by taking the derivative of 9.1 (converted to an equality with μ_0 as the subject) with respect to γ_0 and equating it to zero. The fixed parameters $(\mu_{\text{obs}}, \gamma_{\text{obs}})$ used were the estimates in the gamma-frailty model. If the parameter estimates were

'non-significant' then the reference prior of $N(0, 10^2)$ was retained. The prior means and standard deviations are in tables 9.7 and 9.8.

Main effects

Covariate	Mean	S.D.
Cause	-1.39	1.91
Rx	-1.01	1.25
Cause : Rx	0	10

Table 9.7: Sceptical priors for the main effects

	Prostate Effects		**Other Effects**	
Covariate	Mean	S.D.	Mean	S.D.
haemo	-0.00351	0.00482	0	10
tumour size	0.00121	0.00123	0	10
tumour step	0.0445	0.045	0	10
cardio	0	10	0.0312	0.0319
Rx0: age	-0.00653	0.00722	0	10
Rx1: age	0	10	0.00375	0.00385
Rx0 : bone meta	0	10	0.0958	0.101
Rx1 : bone.meta	0.133	0.148	-0.204	0.242
Rx0: bed<50%	0	10	0	10
Rx1: bed <50%	0	10	0.0876	0.0912
Rx0: conf/bed>50%	0.149	0.158	0	10
Rx1: conf/bed>50%	0	10	0	10

Table 9.8: Sceptical priors for covariate effects

Sceptical Prior Results

The traces and density estimates are given below. The main effects are in figure 9.14, the prostate effects are given in 9.15 and 9.16, the other effects are given in 9.17 and 9.18. The prior densities are given as dotted lines.

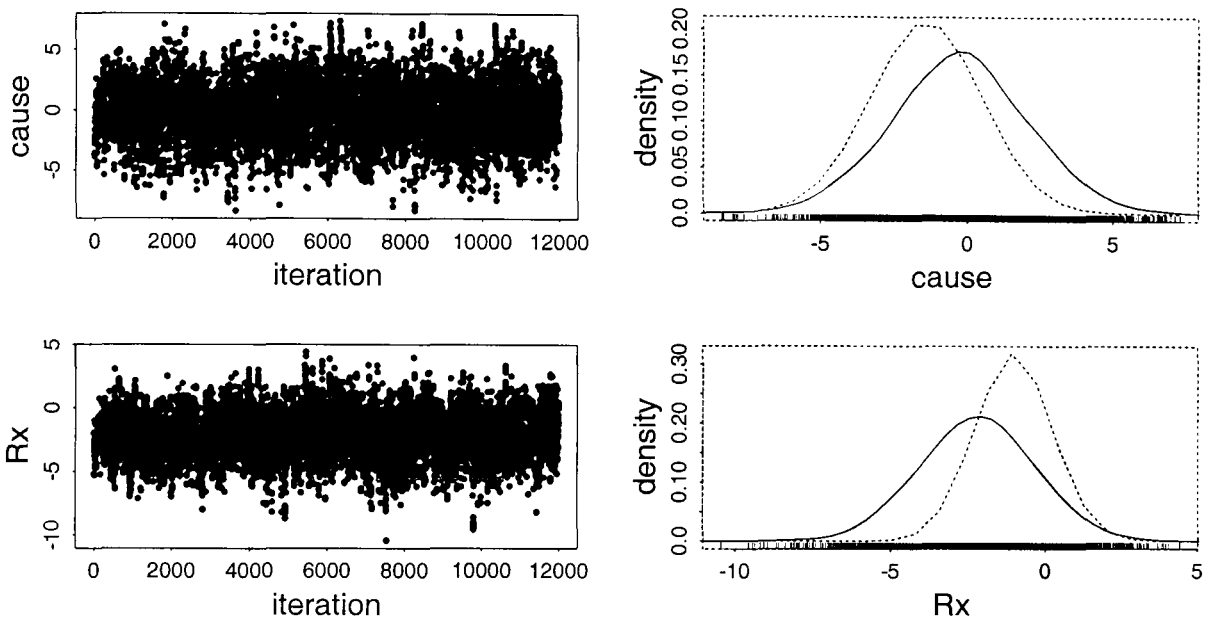


Figure 9.14: Sceptical main effects posteriors

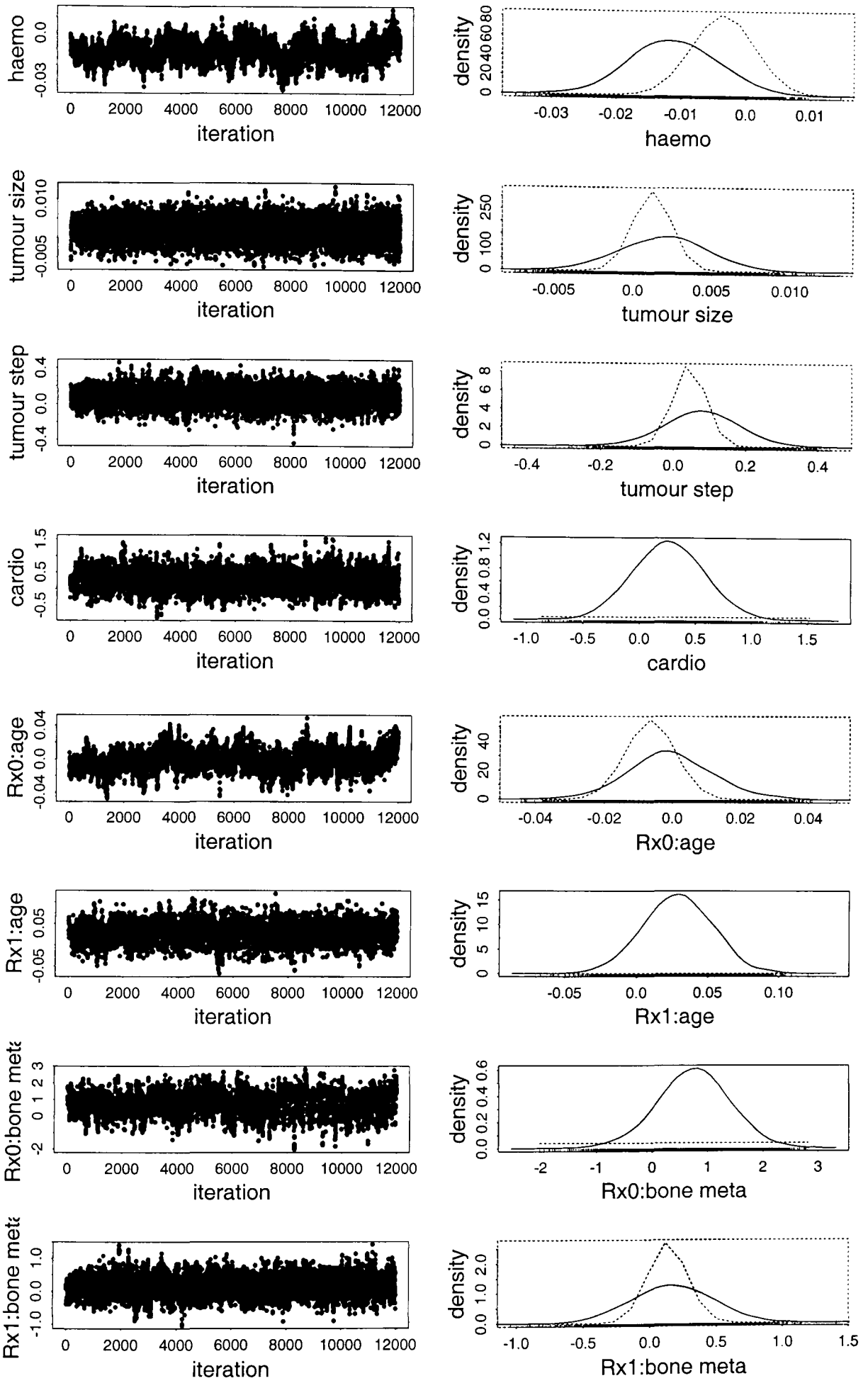


Figure 9.15: Sceptical prostate effects posteriors

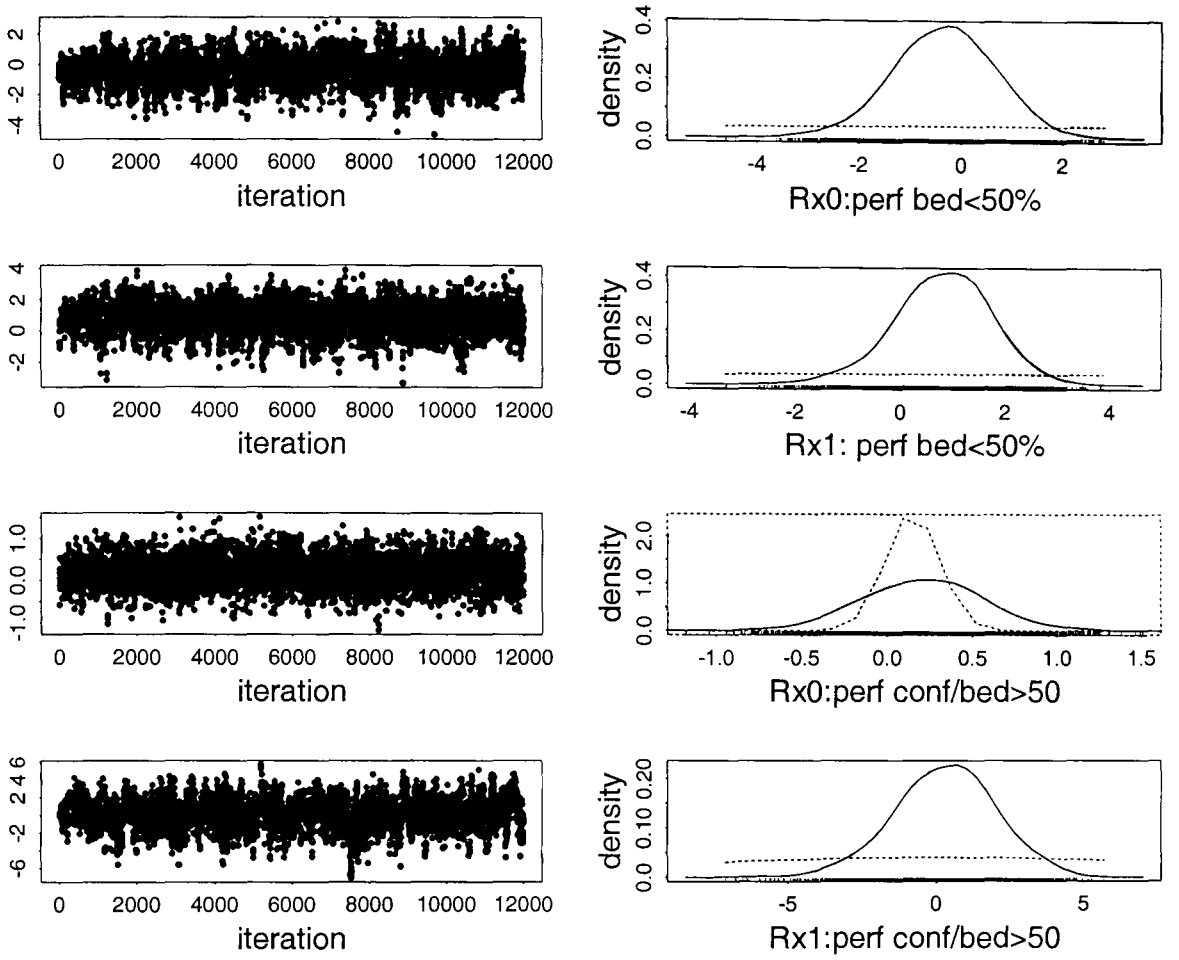


Figure 9.16: Sceptical prostate effects posteriors (continued)

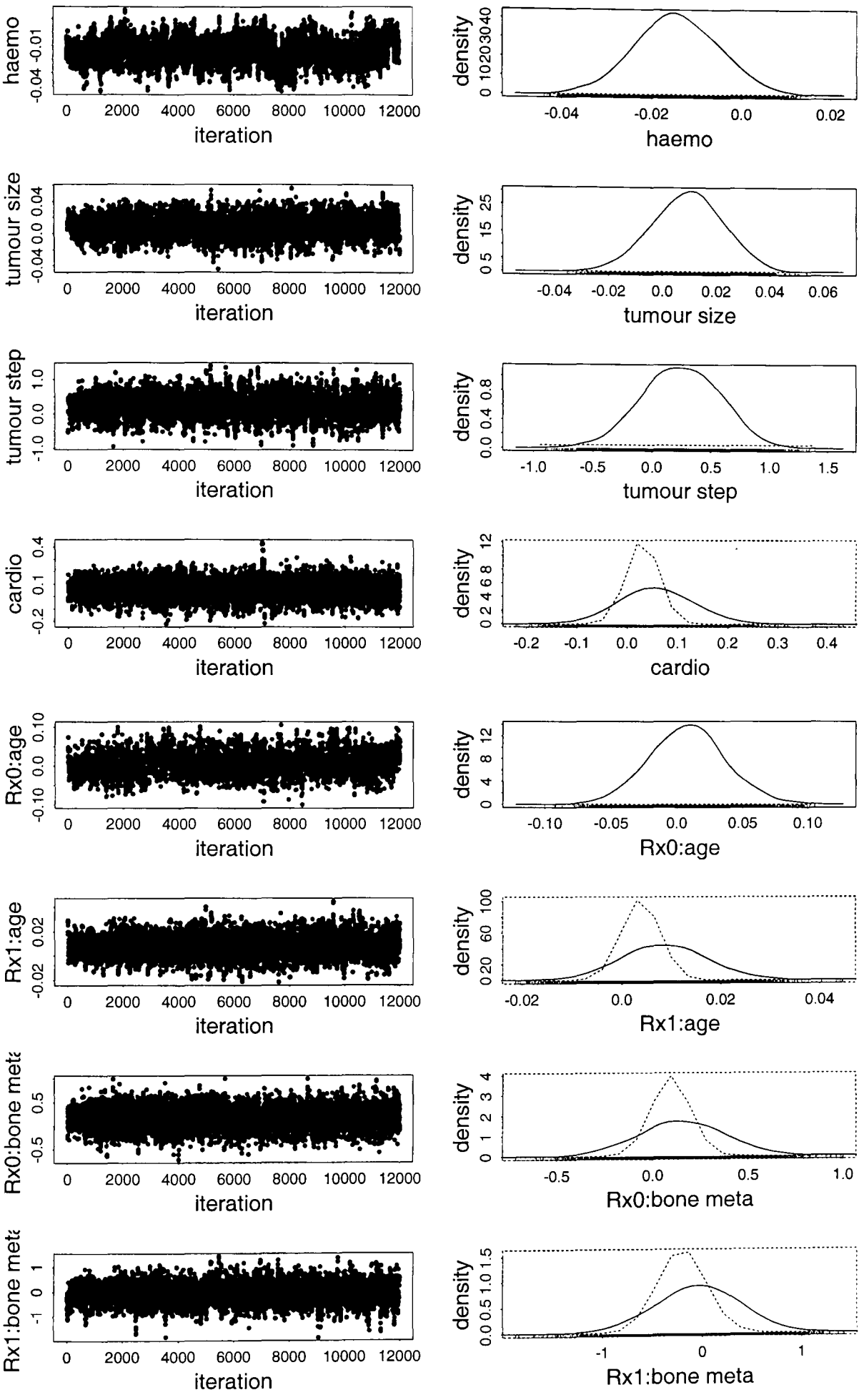


Figure 9.17: Sceptical 'Other' effects posteriors

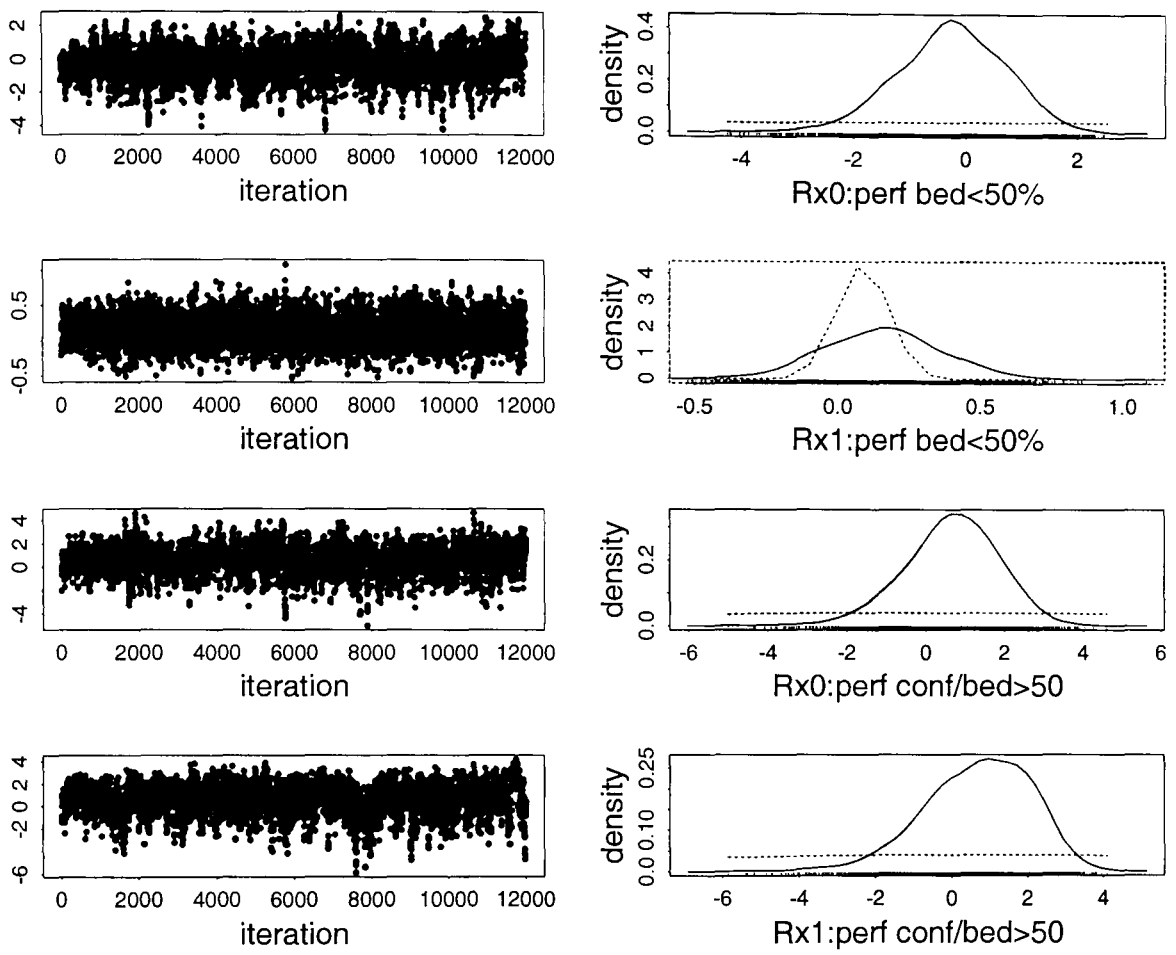


Figure 9.18: Sceptical 'Other' effects posteriors (continued)

The equivalent of figure 9.12, the pointwise confidence intervals for the frailty C.D.F., is shown below in figure 9.19 and they appear to be very similar.

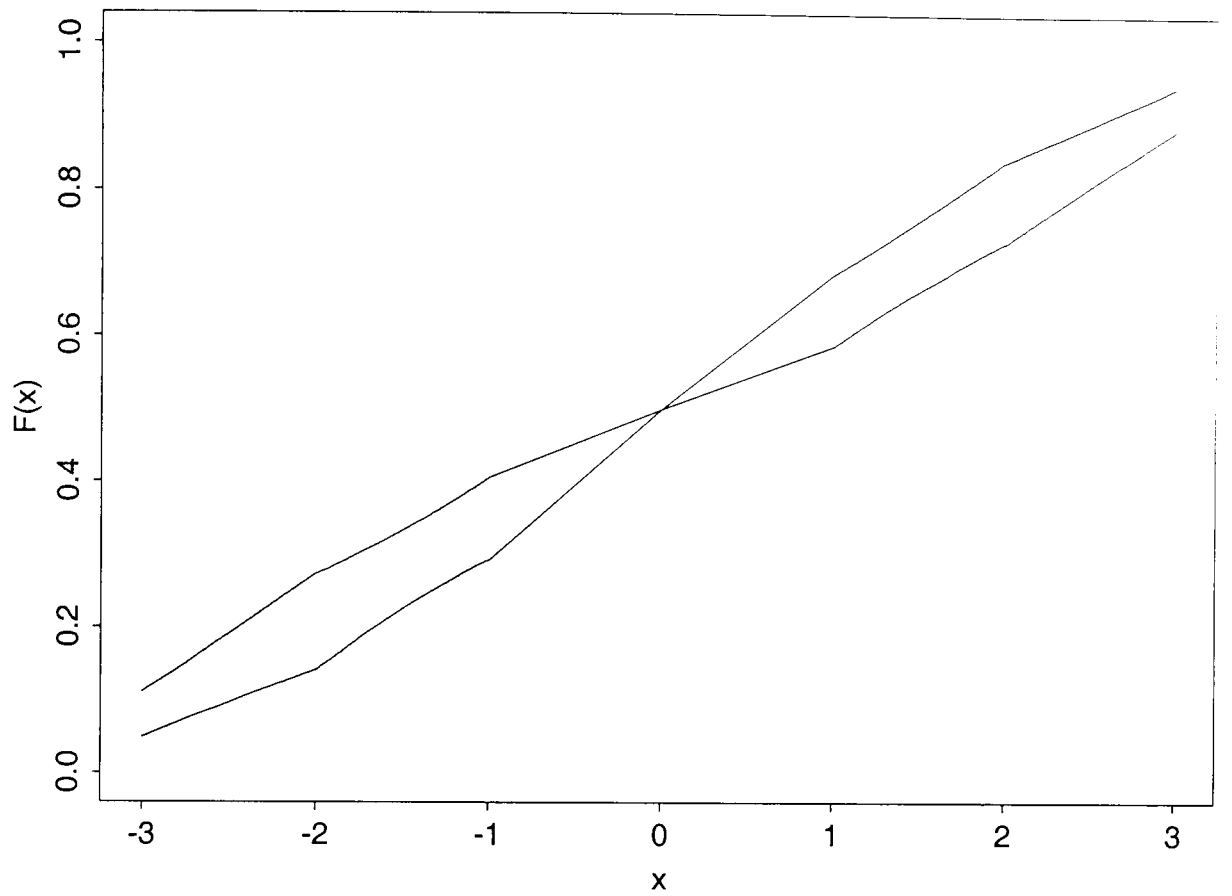


Figure 9.19: 90% pointwise confidence intervals for the C.D.F.

As was intended all of the density estimates now put at least 5% of their mass on the other side of zero from their modulus. For the coefficients with a sceptical prior, the variance of the prior distribution is smaller than the variance of the resulting posterior distribution indicating that the level of *a priori* certainty required to reject the conclusions of the reference prior is higher than the weight of evidence provided by the trial data.

9.5.3 Comparison

To compare the classical and Bayesian analysis (reference prior) the mean estimates of the fixed effects are plotted, along with their 95% credible intervals, against the fixed effects of the gamma frailty model in figure 9.20.

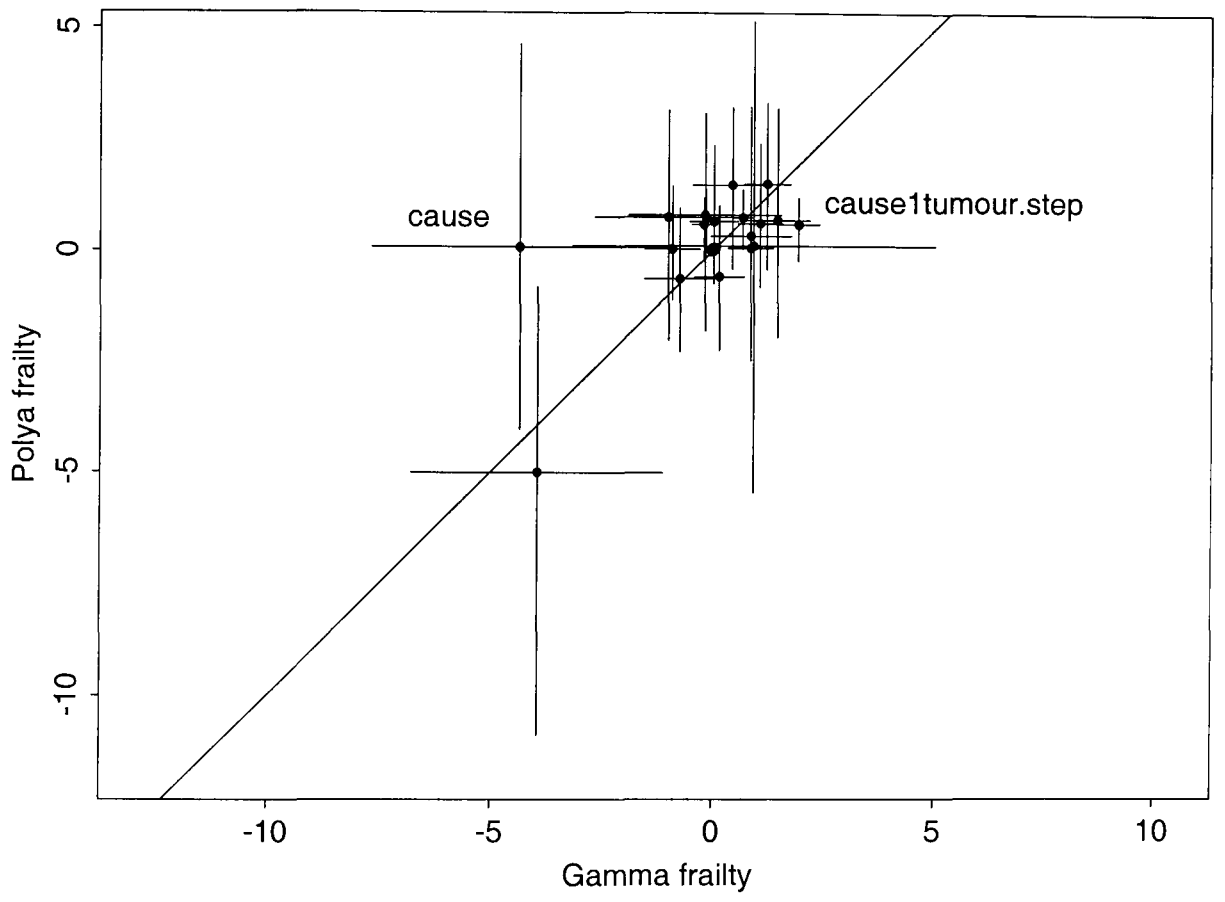


Figure 9.20: Comparison of the fixed effects

There are two covariates whose credible interval does not intersect the line of equality: cause, and prostate:tumour step. However, given that there are 27 estimates, multiple testing comes into play and this should not be considered as a disagreement between the two models. When the influence of individual patients was considered by an

approximate jack-knifing (the *dfbeta* method as discussed in Therneau and Grambsch (2000, pages 153-159)) in the classical analysis there did not appear to be any hugely influential individuals. For Cause the percentage change in the estimate obtained by omitting each individual was between -7.0% and +6.0%. For Prostate:Tumour step the percentage change was between -2.2% and +0.9%.

The main difference between the models is in the frailty distribution. The Pólya tree has a much larger variance that the log-normal model which, in turn, is much larger than the gamma frailty. It is difficult to judge whether this is some shape-driven aspect of the frailty distribution, such a skewness or a heavy-tailed property, as the cumulative density function does not display such features very well. Unfortunately, it is not possible to compare the density function since the posterior Pólya tree density is very spiked with an infinite number of discontinuities. Also it is not at all clear whether the posterior Pólya tree *density* is a consistent estimator. The conclusion of this is that we can only consider expectations, or integrals, with respect to a Pólya tree density. On the positive side, the width of the confidence intervals in figure 9.12 seem to indicate that the estimated distribution is reasonably robust.

It can be observed that, although the posterior distributions for the fixed effects are in agreement, as concerns location, with the classical analyses, the variances are larger in the Bayesian analysis. This may be due to the infinite-parameter distribution draining information from the fixed effects, or there may be some play-off between the random-effects taking a more prominent role in describing the data, and hence requiring more room–or variance–to manoeuvre, which results in a less prominent role for the fixed effects.

9.6 Comparison with existing analysis

The data were originally analysed in Byar and Green (1980), which is a paper that aims to promote the concept of subgroup analysis which optimises treatment. It is undoubtedly aimed at a statistically illiterate and (or) medical audience as the following quotes indicate:

> It is undoubtedly true that most progress in treating patients with cancer has come from the ideas of medical researchers and from the observations of clinicians who are actually treating patients rather than from clever statistical analysis of data collected in the course of treatment.

and

> Non-statisticians in the audience will have to excuse us for a moment while we suggest how tests for treatment-covariate interactions may be carried out.

This data set is used as an illustratory example. Their principal analysis is based on all-causes survival. They do perform regression but rather than using the original scale of the covariates they decide to condense the values into intervals, label the intervals as $0, 1, 2, \ldots$, and then use these labels as a continuous scale where the label=2 can be interpreted as having twice the effect of label=1. The covariates which are considered, due to having a significant univariate effect, are: haemoglobin, performance status, history of cardio vascular disease, stage-grade category, standardised weight, age, tumour size. This is broadly in agreement with our choice of covariates.

The paper does not make clear what univariate model they used. It could be an exponential, Weibull or Cox model. Also it is not clear whether the parameters which are estimated are to be interpreted as an additive or a multiplicative effect on the

hazard function. In all cases a test was performed for an interaction between treatment (the condensed version used here) and the covariate. The final model fitted used the exponential model with additive hazards, where

$$S(t|X) = \exp(-\lambda(X)t), \;\; \lambda(X) = \beta X.$$

The coefficients estimated, and their p-values are given in table 9.9.

variable	coef	p-value
intercept	8.79	<.0001
haemoglobin	8.34	.0067
performance	11.7	.0477
cardio	9.86	<.0001
stage	13.4	<.0001
age	0.804	.7794
tumour size	17.8	.0014
Rx	-3.105	.1707
Rx : stage	-8.45	.0455
Rx : age	9.98	.0143

Table 9.9: Previous analysis

When the patients are stratified according to whether their 4-year survival was >60%, 40% to 60%, 20% to 40%, or <20%, there was good agreement between the actuarial survival curves and the predicted survival curves. They briefly consider the patterns associated with the cause of death, by producing a 3-way table, where patients are classified by cause of death, treatment, and predicted optimal treatment. This, along with the treatment interactions in the model, supports the theory that the treatment helps to

treat prostate cancer but also has some early toxic effects which are counter-productive to cardiovascular disease.

This is in broad agreement with the main effects of our model given in table 9.4. Our model helps to clarify which variables are important for which causes of death. Only age and bone metastases are significant for both. Haemoglobin, tumour size, tumour grade, and the indicator of 'confined to bed >50%' are significant for prostate. These are all non-significant for 'other' causes. Cardiovascular history and the indicator of 'confined to bed <50%' is significant for 'other' causes, which is dominated by cardiovascular death.

9.7 Conclusion

The data set has been re-analysed using a more sophisticated model which may be able to capture dependencies between causes of failure, and also allows the consideration in greater detail of which variables influence which causes of failure. Although there is good agreement about the fixed effects, both between the new models presented and previous analyses. This is not the case for the modeled frailty distribution. An unsolved problem is the extraction of a full log-likelihood from the log-normal, rather than a partial-penalized-profile-log-likelihood. This would allow the formal comparison of the gamma and the log-normal distribution by means of their deviances or A.I.C. The use of the Pólya tree distribution added to the confusion although it does give some indication as to the entire distribution of the frailty distribution, rather than attempting to reduce it to one parameter. Possible developments to the model would be to consider the treatment in its original four stages. A further refinement of the cause is unlikely to reveal anything through low statistical power and a model was considered which grouped all the cancer deaths together, but very little changed.

Chapter 10

Overview and future directions

10.1 Counting Process Applications

The counting process theory which was used in chapter 2 enabled the production of confidence bands on important functions such as the crude incidence function. These bands have been derived in a fairly ad hoc manner with their main raison d'être being the ability to calculate critical values such as $\mathbb{P}(\sup_{t\in[a,b]} |W(t)| < k)$, where W is a Brownian bridge. At present the author is only aware of one free-standing computer package that can calculate such quantities. Although this package is freely down-loadable from the internet (`http://www.nrcan.gc.ca/gsc/mrd/sdalweb/wiener/index.html`), it can only run on MS-DOS and may stop being forward compatible with the current Microsoft environment within a decade. It would be desirable if the calculation of such probabilities and quantiles were to become a standard part of statistical software packages.

An area which was only lightly touched upon in this chapter was k-sample testing for differences between the crude incidence function of subgroups of the sample.

Although the concept of hypothesis testing has become largely discredited within the statistics community, they are widely applied within medical statistics, and hence it is desirable to investigate the power of such tests against various alternative hypotheses. At present the tests which are considered in the literature (Gray 1988, Lunn 1998) are only powerful against an ordered difference between the crude incidence functions: where there exists a sequence of groups, $i_1, \ldots, i_k$ such that

$$Q_{i_1}(t) \leq Q_{i_2}(t) \leq \ldots \leq Q_{i_k}(t),$$

for all values of t. This is because they are all formed by deriving a process $\mathbf{Z}(t)$ which is a vector of the difference between the individual crude incidence functions and a weighted average. This is normally distributed with a variation process $\Sigma(t)$, and a mean of zero when the null hypothesis is true. With this $\mathbf{Z}(t)$ process, a chi-square statistic is formed by taking a time point at the end of the study period, τ, and calculating the statistic $\mathbf{Z}(\tau)^t \Sigma^{-1}(\tau)\mathbf{Z}(\tau)$. However, if the crude incidences are different but cross over, rather than diverging, then this statistic will not have optimal power. A more uniformly powerful test can be inspired by the Kolmogorov-Smirnov test where we take the same process, $\mathbf{Z}(t)$, but consider $\sup_{t\in[0,\tau]} |\mathbf{Z}(t)|$, instead of a chi-square statistic. This is considered in more detail in Andersen et al. (1993, section V.4, pp. 390-7).

10.2 Bounds on the joint survival

Chapter 3 contains the result that we can obtain a different set of bounds on the joint survival function in the case of a 2-sample data set. To calculate these bounds we need to assume that there is a covariate-time transformation $\phi(\mathbf{t})$ which can calculate the joint survival for one group at a fixed point $\mathbf{t}$ by transforming this point and evaluating the joint survival at the transformed point conditional on being in the *other* group. It

is assumed that this transformation is known. In some regions of the latent-failure time space these alternative bounds are tighter than the conventional Peterson bounds.

The proof rests on that proposition that when the region of in the latent-failure space where T_1, say, is the smallest is transformed by ϕ, then the resulting region is bounded by the transform of the original bounds. Given that the transformation is continuous and monotonic this seems highly plausible. This proposition may well be a trivial result to a topologist, but the author is currently unaware of a proof.

Given the fact that in a two-armed data set there are two possible bounds for the joint survival it would be useful to know which bounds are tighter in a particular region of the latent-failure time space. This can be answered by simply calculating the bounds, but when the result is generalised to a k-sample data set this may become computationally infeasible. We have limited results as to which bounds are optimal where, but it would be worthwhile to generalise.

10.3 Covariate-time transformation

The following chapter considers this covariate-time transformation, ϕ, in more detail. It makes the specific assumption that $\phi_i(\mathbf{t}) = \phi_i(t_i)$, which means that the derivative matrix is diagonal. This assumption can be justified if we expect that any dependence structure (as opposed to 'scale' or 'location')in the latent-failure times, which can be described in terms of a copula, is invariant to the covariates. It is unknown what happens when this assumption is relaxed and should be investigated further.

The main result of the chapter is the bounds obtained on this covariate-time transformation in the case of a two-armed data set. Pointwise confidence intervals on these bounds can be formed but confidence bands for a range of time points is more difficult. The problem can be expressed quite generally in that we have two functions,

$G(t)$ and $H(t)$, say, where these are estimated, with random error, by $\widehat{G}$ and $\widehat{H}$. Given the error process for each estimate, what can be said about the error process for $\widehat{G}^{-1}(\widehat{H}(t))$?

10.4 Identifiability

The chapter on identifiability starts with the important result of Heckman and Honoré (1989) and considers whether its assumptions can be relaxed. The assumptions are that the joint survival function is of the form

$$S(\mathbf{t}, z) = K(\Lambda_1(t_1)\phi_1(z), \ldots, \Lambda_k(t_k)\phi_k(z)).$$

This assumes that the covariate-time transformation ϕ has a diagonal time derivative matrix, but also assumes that it is of the form $\Lambda(t)\phi(z)$—proportional hazards. It is considered whether the result of identifiability can still be obtained if the proportional hazards assumption is dropped. The answer is no.

It was surprisingly difficult to come up with this answer and it does raise the question of how wide a set of models the single assumption of a diagonal derivative matrix implies. It may be the case that some quantities derived from the joint distribution are identified. A general sensitivity analysis may be useful. A related question is whether or not the assumptions of Heckman and Honoré (1989) are necessary, as well as sufficient, to allow identifiability. If this were the case it would be surprising as the assumption of proportional hazards is not particularly realistic or suitable for all data. A greater understanding is required of why these assumptions give identifiability.

10.5 Frailty Modeling

Here we have considered the assumption that any dependence in the latent-failure time distribution is because of an unobserved covariate. This is described by assuming a

frailty model where the failure times for each cause are independent conditional on this covariate and that the effect of this covariate is to multiply all the cause-specific hazards by a factor.

One question which has been considered is the sensitivity of the model to the assumed distribution of this frailty variable. This is considered in a practical sense in chapter 9 where a non-parametric tool was used to describe the frailty distribution. The broad conclusion was that the fixed effects were not particularly sensitive to the frailty distribution, but the variance of the frailty was sensitive to the choice of distribution.

Another question would be to ask whether multiplying the cause-specific hazards by the same factor is appropriate. If there were a negative correlation between two latent failure times then it would clearly be inappropriate. A natural extension to the univariate frailty model considered here is a multivariate frailty distribution where each cause-specific hazard is multiplied by a different factor and the aim is to describe the distribution of this multivariate collection of frailties. Questions of practical computation and of identifiability need to be addressed.

Within the frequentist framework, there is a large choice of approximate likelihoods, whether it be quasi-likelihood, penalised-likelihood or partial-likelihood or a combination thereof. Which is optimal in terms of robustness, bias, and variance is unknown. A slight improvement to the speed of the existing algorithms is proposed where the interval bisection algorithm is used.

10.6 Pólya trees

The Bayesian non-parametric tool, the Pólya tree, is examined in chapter 8. We have obtained results on what the limiting distribution of the random density is. This result only applies pointwise, so we can say what the marginal distribution of $f_\infty(y)$ is, but the

distribution of the multivariate random variable $(f_\infty(y_1), \ldots, f_\infty(y_k))$ is an unanswered question, as is the random process for an interval of values of y. What *is* true is that they are not independent.

The result obtained assumed a specific form for the parameters of the prior distribution: $\alpha_n = k2^n$. This form can be generalised to $\alpha_n = ka^n$, for positive k and $a > 1$. What the consequences are for f_∞ are unknown.

A key question is whether the posterior density is consistent. This is important if the aim is to see if a density is multi-modal. If the density is not consistent, or alternatively is not smooth enough, then this will be hard to judge. At the moment, from practical experience, we have to examine the mean density function obtained from a simulation and still attempt to judge questions of modality through a mass of spikes. This is rather like examining a profile of a mountain range and counting how many mountains there are despite seeing a large number of local peaks.

The theory behind Pólya trees could easily be extend to cases where $\Omega \neq \mathbb{R}$. An area for investigation is how to perform marginalisation or conditioning if $\Omega = \mathbb{R}^p$.

The integration with the Pólya tree random measure is also considered in this chapter. The basic tool used is the trapezium rule. This is not the cutting edge of numerical integration, but it does allow tractability. Whether any improvements can be obtained by the use of more sophisticated methods of numerical integration is a good question. The main potential for error occurs when the integrand is large in the tails of the Pólya tree's sample space. There is no sensible answer to this problem other than 'don't do it.' This is no more than statistical common sense saying that it is inadvisable to make inferences about quantities for which there is little data.

Appendix A

Data

A.1 Boag 1949

The data are taken from Boag (1949) which records the survival times, in months, of 121 breast cancer patients from the clinical records of one hospital over the period 1929 to 1938. The causes are: Cancer, Other, Censored.

Cancer

0.3, 5, 5.6, 6.2, 6.3, 6.6, 6.8, 7.5, 8.4, 8.4, 10.3,
11, 11.8, 12.2, 12.3, 13.5, 14.4, 14.4, 14.8, 15.7,
16.2, 16.3, 16.5, 16.8, 17.2, 17.3, 17.5, 17.9,
19.8, 20.4, 20.9, 21, 21, 21.1, 23, 23.6, 24, 24,
27.9, 28.2, 29.1, 30, 31, 31, 32, 35, 35, 38, 39,
40, 40, 41, 41, 42, 44, 46, 48, 48, 51, 51, 52,
54, 56, 60, 78, 78, 80, 84, 87, 89, 90, 97, 98,
100, 114, 123, 161, 174

Other

0.3, 4, 7.4, 15.5, 23.4, 46, 46, 51, 65, 68, 83,
88, 96, 110, 111, 112, 132, 162

Censored

111, 112, 113, 114, 114, 117, 121, 123, 129,
131, 133, 134, 134, 136, 141, 143, 167, 177,
179, 189, 201, 203, 203, 213, 228

A.2 Hoel 1972

The data are taken from Hoel (1972). They arise from a laboratory experiment in which mice were given a radiation dose of 300 rads at 5 to 6 weeks old. They were split into two groups according to the conditions in which they were subsequently kept. There were three recorded causes of death.

	Conventional Lab, n=99
Thymic Lymphoma	156, 189, 191, 198, 200, 207, 220, 235, 245, 250, 256, 261, 265, 266, 280, 343, 356, 383, 403, 414, 428, 432
Reticulum cell sarcoma	317, 318, 399, 495, 525, 536, 549, 552, 554, 557, 558, 571, 586, 594, 596, 605, 612, 621, 628, 631, 636, 643, 647, 648, 649, 661, 663, 666, 670, 695, 697, 700, 705, 712, 713, 738, 748, 753
Other	40, 42, 51, 62, 163, 179, 206, 222, 228, 252, 259, 282, 324, 333, 341, 366, 385, 407, 420, 431, 441, 461, 462, 482, 517, 517, 524, 564, 567, 586, 619, 620, 621, 622, 647, 651, 686, 761, 763
	Germ-free, n=82
Thymic Lymphoma	158, 192, 193, 194, 195, 202, 212, 215, 229, 230, 237, 240, 244, 247, 259, 300, 301, 321, 337, 415, 434, 444, 485, 496, 529, 537, 624, 707, 800
Reticulum cell sarcoma	30, 590, 606, 638, 355, 679, 691, 693, 696, 747, 752, 760, 778, 821, 986
Other	136, 246, 255, 376, 421, 565, 616, 617, 652, 655, 658, 660, 662, 675, 681, 734, 736, 737, 757, 769, 777, 800, 806, 825, 855, 857, 864, 868, 870, 870, 873, 882, 895, 910, 934, 942, 1015, 1019

A.3 Prostate Cancer Data

The data are published in Andrews and Herzberg (1985) and were originally published in Byar and Corle (1977) and Byar and Green (1980) and can be downloaded at `http://lib.stat.cmu.edu/datasets/Andrews/T46.1` . The first five patients from this data set of 506 patients are below in S-plus format.

	stage	Tx	date.month	date.day	date.year	time	status	age	weight.index
1	3	0.2mg	8	10	67	72	alive	75	76
2	3	0.2mg	9	21	67	1	cancer	54	116
3	3	5.0mg	1	12	68	40	cerebo	69	102
4	3	0.2mg	3	18	68	20	cerebo	75	94
5	3	placebo	3	21	68	65	alive	67	99

	performance	cardio	SBP	DBP	ECG	haemo	tumour.size	tumour.grade	acid.phos
1	normal	no	15	9	strain	138	2	8	3
2	normal	no	13	7	block	146	42	NA	7
3	normal	yes	14	8	strain	134	3	9	3
4	bed<50%	yes	14	7	benign	176	4	8	9
5	normal	no	17	10	normal	134	34	8	5

	bone.meta
1	no
2	no
3	no
4	no
5	no

A.4 Input file for the M.C.M.C. programme

For the sake of clarity I have included the first five lines of the input file to the C-programme, `prostpart`, for which the code is in section B.4 . The values have been rounded to three significant figures and the lines have been broken to fit on the page.

```
 72   0   1 0.0000   0.0000   0.0000   6.52e-02 -1.45e-17 -4.38e-02 -5.95e-18
 72   0   1 0.0454  -0.0262  -0.0370   4.16e-17  1.18e-02 -3.42e-17 -4.60e-02
 40   0   2 0.0000   0.0525  -0.0371  -7.09e-04  2.62e-17 -3.93e-02  1.08e-17
 40   1   2 0.0454   0.0263   0.0371  -3.82e-17 -3.54e-03  3.38e-17 -4.21e-02
 20   0   3 0.0000   0.0000   0.0000   8.32e-02 -1.85e-17 -5.06e-02 -7.64e-18

-2.64e-02 -5.31e-18 -4.65e-02 -6.91e-18  4.05e-02  5.91e-16  8.13e-03
 1.58e-17 -2.99e-02 -4.21e-18 -5.01e-02 -2.26e-16  3.68e-02 -3.26e-15
-2.61e-02  9.34e-18  5.12e-02  8.02e-18  9.62e-03  8.70e-16 -3.19e-02
-1.22e-17 -2.99e-02  9.12e-18  4.86e-02  1.54e-16 -6.41e-03  3.55e-16
-3.59e-02 -6.70e-18  3.26e-02 -9.32e-18 -5.16e-02 -2.38e-15 -3.42e-03

 1.01e-17 -1.15e-02 -1.34e-17  1.63e-02  2.73e-17 -9.95e-03 -2.95e-17
 8.64e-03 -8.84e-17 -8.02e-03 -3.94e-16  1.76e-02 -1.77e-17 -8.46e-03
-2.78e-15  1.54e-02  4.56e-16 -2.19e-02 -3.25e-16 -2.29e-03  2.80e-16
-3.22e-02 -9.99e-17  1.36e-02 -4.86e-16 -2.25e-02 -4.50e-18 -3.10e-03
 1.75e-15  1.58e-03 -2.41e-16  2.84e-02  2.74e-16  2.28e-01 -1.75e-16

 1.01e-02 -1.39e-17 -3.80e-03  2.13e-17 -3.49e-03 -1.54e-17
 4.26e-17  1.08e-02 -7.65e-17 -3.99e-03  1.33e-16 -3.02e-03
-1.87e-02  1.59e-16 -8.24e-04 -1.93e-16 -3.53e-03  2.05e-16
-7.90e-17 -1.91e-02 -9.41e-17 -7.16e-04  6.81e-17 -3.80e-03
 5.43e-03 -9.54e-17  2.16e-02  1.16e-16  3.38e-04 -1.49e-16
```

Appendix B

Code

B.1 Crude Incidence estimator

The following is S-plus code (Becker et al. 1988, Chambers and Hastie 1992). The input is a vector of observed times, a vector of causes of failure, and an argument which defines the code for censored values. The output is a list composed of a vector of sorted times, a matrix of the crude incidence function with one column for each cause, and a matrix which estimates the variance of each estimator.

```
CrIn<-function(time, cause, censor = "0")
{
index <- order(time)
time <- time[index]
cause <- cause[index]
Causes <- levels(cause)[levels(cause) != censor]
dN <- 1 * outer(as.character(cause), Causes, "==")
Y <- length(time):1
dLj <- dN/Y
dL <- apply(dLj, 1, sum)
Sminus <- c(1,cumprod(1 - dL)[1:(length(time)-1)])
Q <- apply(dLj * Sminus, 2, cumsum)
COV <- (Sminus^2 * dLj + Q^2 * dL + 2 * Sminus * Q * dLj
- 2 * Q * (Sminus * dLj +Q * dL) + Q^2 * dL)/Y
```

```
COV <- apply(COV, 2, cumsum)
list(time = time, crude.incidence = Q, variance = COV)
}
```

B.2 Cox frailty model with bisection algorithm

This is a set of S-plus functions which use the (old-style) *class* structure of the S-plus language (Venables and Ripley 2000, chapter 4).

```
coxglmm
function(x, ...)
{
UseMethod("coxglmm")
}
```

```
coxglmm.default
function(formula, data, random, subset, start = 7, upper = start, lower = start/
10, verbose = T, method = "bisect", disp = "REML", ...)
{
call <- match.call()
#obtain the fixed effects design matrix
m <- match.call(expand = F)
temp <- c("", "formula", "data", "weights", "subset", "na.action")
m <- m[match(temp, names(m), nomatch = 0)]
m[[1]] <- as.name("model.frame")
m <- eval(m, sys.parent())
Terms <- terms(formula, specials = "strata", data = data)
attr(Terms, "intercept") <- 1
xvars <- as.character(attr(Terms, "variables"))
if((yvar <- attr(Terms, "response")) > 0)
xvars <- xvars[ - yvar]
if(length(xvars) > 0) {
xlevels <- lapply(m[xvars], levels)
xlevels <- xlevels[!sapply(xlevels, is.null)]
if(length(xlevels) == 0)
xlevels <- NULL
}
```

```
else xlevels <- NULL
temp <- untangle.specials(Terms, "strata", 1)
if(length(temp$vars)) {
X <- model.matrix(Terms[ - temp$terms], m)
strata <- as.numeric(strata(m[, temp$vars], shortlabel = T))
}
else {
X <- model.matrix(Terms, m)
strata <- NULL
}
#to remove the intercept but make sure it copes with nested formulae
xint <- match("(Intercept)", dimnames(X)[[2]], nomatch = 0)
if(xint > 0)
X <- X[,  - xint, drop = F]
# extract the responses
Y <- model.extract(m, response)
if(class(Y) != "Surv") {
stop("Error: response must be a Surv object")
}
else {
time <- Y[, 1]
status <- Y[, 2]
}
#   get the random effects matrix
mz <- match.call(expand = F)
mz$formula <- mz$random
temp <- c("", "formula", "data", "weights", "subset", "na.action")
mz <- mz[match(temp, names(mz), nomatch = 0)]
mz[[1]] <- as.name("model.frame")
mz <- eval(mz, sys.parent())
Termsz <- attr(mz, "terms")
attr(Termsz, "intercept") <- 0
Z <- model.matrix(Termsz, mz)
if(method == "bisect") {
fit.upper <- coxglmm.fit(status, time, strata, X, Z, sigma2 =
upper, verbose = T, disp = disp)
fit.lower <- coxglmm.fit(status, time, strata, X, Z, sigma2 =
lower, verbose = T, disp = disp)
if((upper - fit.upper$sigma2) * (lower - fit.lower$sigma2) >
0)
stop("try larger/smaller starting value")
while(upper - lower > 1e-06) {
```

```
bisect <- (fit.lower$sigma2 + fit.upper$sigma2)/2
fit.bisect <- coxglmm.fit(status, time, strata, X,
Z, sigma2 = bisect, verbose = T, disp = disp)
if(bisect - fit.bisect$sigma2 > 0) {
upper <- bisect
fit.upper <- fit.bisect
}
else {
lower <- bisect
fit.lower <- fit.bisect
}
}
fit <- coxglmm.fit(status, time, strata, X, Z, sigma2 = (lower +
upper)/2, verbose = verbose, disp = disp)
}
else {
sigma <- start
sigma.old <- start + 1
while(abs(sigma.old - sigma) > 1e-05) {
fit <- coxglmm.fit(status, time, strata, X, Z, sigma2
 = sigma, verbose = verbose, disp = disp)
sigma.old <- sigma
simga <- fit$sigma2
}
fit <- coxglmm.fit(status, time, strata, X, Z, sigma2 = sigma,
verbose = verbose, disp = disp)
}
fit$fixedterms <- Terms
fit$randomterms <- Termsz
fit$call <- call
fit$x <- X
fit$y <- Y
fit$z <- Z
fit$formula <- call$formula
fit$coefficients <- fit$beta[1:dim(X)[2],  ]
fit$random.effects <- fit$beta[(dim(X)[2] + 1):(dim(X)[2] + dim(Z)[
2]),  ]
fit$n <- dim(X)[1]
fit$var <- solve(fit$Hessian)[1:dim(X)[2], 1:dim(X)[2]]
if(!is.null(xlevels))
attr(fit, "xlevels") <- xlevels
if(!is.null(fit$call$disp) && fit$call$disp == "ML") {
```

```
M <- solve(fit$Hessian[(dim(X)[2] + 1):dim(fit$Hessian)[1],
(dim(X)[2] + 1):dim(fit$Hessian)[1]])
}
else {
M <- (solve(fit$Hessian))[(dim(X)[2] + 1):dim(fit$Hessian)[
1], (dim(X)[2] + 1):dim(fit$Hessian)[1]]
}
r <- eigen(M)$values
fit$varsig <- (2 * (fit$sigma2)^2)/(dim(Z)[2] - (2 * sum(r))/fit$sigma2 +
sum(r^2)/(fit$sigma2)^2)
structure(fit, class = "coxglmm")
}

 coxglmm.fit
function(y, time, strata, X, Z, sigma2 = 7, verbose, disp)
{
#initialisation
if(length(strata) == 0) {
index <- order(time, 1 - y)
newstrats <- length(y)
}
else {
index <- order(strata, time, 1 - y)
newstrats <- table(strata)
}
y <- y[index]
X <- as.matrix(X[index,  ])
Z <- Z[index,  ]
beta <- rep(0, dim(X)[2] + dim(Z)[2])
# M is a block diagonal matrix of lower triangular matrices of 1s
for(i in 1:length(newstrats)) {
if(i == 1)
M <- outer(1:newstrats[1], 1:newstrats[1], ">=")
else {
m <- outer(1:newstrats[i], 1:newstrats[i], ">=")
M <- cbind(rbind(M, matrix(0, ncol = dim(M)[2], nrow =
dim(m)[1])), rbind(matrix(0, ncol = dim(m)[
2], nrow = dim(M)[1]), m))
}
}
#inner loop for coefficients and random effects
iter <- 0
```

```
beta.old <- rep(1, length(beta))
while(max(abs(beta - beta.old)) > 1e-05 & iter < 100) {
eta <- cbind(X, Z) %*% beta
w <- exp(eta)
W <- diag(w, nrow = length(w))
a <- y/(t(M) %*% w)
b <- M %*% a
d <- y - w * b
H <- rbind(t(X), t(Z)) %*% (diag(w * b, nrow = length(b)) -
W %*% M %*% diag(a^2, nrow = length(a)) %*% t(M) %*% W) %*%
cbind(X, Z)
V <- H + diag(rep(c(0, 1/sigma2), c(dim(X)[2], dim(Z)[2])),
nrow = dim(X)[2] + dim(Z)[2])
beta.old <- beta
beta <- beta + solve(V, rbind(t(X), t(Z)) %*% d - rep(c(0,
1/sigma2), c(dim(X)[2], dim(Z)[2])) * beta)
iter <- iter + 1
}
#inner loop for dispersion
iter2 <- 0
sigma2.old <- sigma2 + 1
sigmaouter <- sigma2
while(abs(sigma2 - sigma2.old) > 1e-07 & iter2 < 100) {
V <- H + diag(rep(c(0, 1/sigma2), c(dim(X)[2], dim(Z)[2])),
nrow = dim(X)[2] + dim(Z)[2])
if(disp == "ML") {
v <- sum(diag(solve(V[(dim(X)[2] + 1):(dim(H)[1]),
(dim(X)[2] + 1):(dim(H)[2])])))
}
else {
v <- sum(diag(solve(V)[(dim(X)[2] + 1):(dim(H)[1]),
(dim(X)[2] + 1):(dim(H)[2])]))
}
sigma2.old <- sigma2
sigma2 <- (t(beta[(dim(X)[2] + 1):(dim(H)[1]),  ]) %*% beta[
(dim(X)[2] + 1):(dim(H)[1]),  ] + v)/dim(Z)[2]
iter2 <- iter2 + 1
}
loglik <- t(y) %*% (w - t(M) %*% w) - 1/2 * (t(beta[(dim(X)[2] + 1):
(dim(H)[1]),  ]) %*% beta[(dim(X)[2] + 1):(dim(H)[1]),  ]/
sigma2)
if(verbose)
```

```
cat("\nCoefficients: ", beta[1:dim(X)[2]], "\nSigma^2: ",
sigma2, "\nPQL: ", loglik, "\n")
list(sigma2 = sigma2, beta = beta, Hessian = V, loglik = loglik)
}

print.coxglmm
function(x, ...)
{
cat("Call:\n")
print(x$call)
cat("\nFixed effects:\n")
print(x$coefficients)
cat("\nVariance of random effect:\n")
cat(x$sigma2, "\n")
cat("\nPenalised Quasi Log-Likelihood:\n")
cat(x$loglik, "\n")
cat("\nEstimating variance of the random effects variance:\n")
cat(x$varsig, "\n")
invisible(x)
}

summary.coxglmm
function(x, ...)
{
class(x) <- "coxph"
UseMethod("summary", x, ...)
cat("\nRandom Effects Variance\n")
cat(x$sigma2, "\n")
cat("\nEstimating variance of the random effects variance:\n")
cat(x$varsig, "\n")
cat("\nPenalised Quasi Log-Likelihood:\n")
cat(x$loglik, "\n")
invisible(x)
}

plot.coxglmm
function(x, ...)
{
par.store <- par()
on.exit(par(par.store))
par(ask = T)
base <- baseline(x)
```

```
plot(cumhaz ~ time, base, type = "s", ...)
base$upper <- base$cumhaz + 1.96 * sqrt(base$variance)
base$lower <- base$cumhaz - 1.96 * sqrt(base$variance)
lines(base$time, base$upper, type = "s", lty = 2)
lines(base$time, base$lower, type = "s", lty = 2)
par(ask = T)
plot(exp( - cumhaz) ~ time, base, type = "s", ylim = c(0, 1), ylab =
"survival", ...)
lines(base$time, exp( - base$upper), type = "s", lty = 2)
lines(base$time, exp( - base$lower), type = "s", lty = 2)
}

baseline
function(fit)
{
r <- exp(cbind(fit$x, fit$z) %*% fit$beta)
n <- length(r)
# upper triangular matrix
M <- outer(1:n, 1:n, "<=")
index <- order(fit$y[, 1], 1 - fit$y[, 2])
N <- fit$y[, 2]
cumhaz <- cumsum(N[index]/(M %*% (r[index])))
var <- cumsum(N[index]/(M %*% (r[index]))^2)
time <- fit$y[, 1][index]
index <- (N[index] == 1)
data.frame(time = time[index], cumhaz = cumhaz[index], variance = var[
index])
}
```

B.3 Ammended exisiting frailty code

This is an ammended version of the existing code by T Therneau (Therneau and Grambsch 2000) which can cope with interactions and frailty terms in a coxph formula.

```
coxph2
function(formula = formula(data), data = sys.parent(), weights, subset,
```

```
na.action, init, control, method = c("efron", "breslow", "exact"),
singular.ok = T, robust = F, model = F, x = F, y = T, ...)
{
method <- match.arg(method)
call <- match.call()
m <- match.call(expand = F)
temp <- c("", "formula", "data", "weights", "subset", "na.action")
m <- m[match(temp, names(m), nomatch = 0)]
special <- c("strata", "cluster", "frailty")
Terms <- if(missing(data)) terms(formula, special) else terms(formula,
special, data = data)
m$formula <- Terms
m[[1]] <- as.name("model.frame")
m <- eval(m, sys.parent())
if(missing(control))
control <- coxph.control(...)
Y <- model.extract(m, "response")
if(!inherits(Y, "Surv"))
stop("Response must be a survival object")
weights <- model.extract(m, "weights")
offset <- attr(Terms, "offset")
tt <- length(offset)
offset <- if(tt == 0) rep(0, nrow(Y)) else if(tt == 1)
m[[offset]]
else {
ff <- m[[offset[1]]]
for(i in 2:tt)
ff <- ff + m[[offset[i]]]
ff
}
attr(Terms, "intercept") <- 1
#Cox model always has \Lambda_0
strats <- attr(Terms, "specials")$strata
cluster <- attr(Terms, "specials")$cluster
dropx <- NULL
if(length(cluster)) {
if(missing(robust))
robust <- T
tempc <- untangle.specials(Terms, "cluster", 1:10)
ord <- attr(Terms, "order")[tempc$terms]
if(any(ord > 1))
stop("Cluster can not be used in an interaction")
```

```
cluster <- strata(m[, tempc$vars], shortlabel = T)
#allow multiples
dropx <- tempc$terms
}
if(length(strats)) {
temp <- untangle.specials(Terms, "strata", 1)
dropx <- c(dropx, temp$terms)
if(length(temp$vars) == 1)
strata.keep <- m[[temp$vars]]
else strata.keep <- strata(m[, temp$vars], shortlabel = T)
strats <- as.numeric(strata.keep)
}
if(length(dropx))
X <- model.matrix(Terms[ - dropx], m)[, -1, drop = F]
else X <- model.matrix(Terms, m)[, -1, drop = F]
type <- attr(Y, "type")
if(type != "right" && type != "counting")
stop(paste("Cox model doesn't support \"", type,
"\" survival data", sep = ""))
if(missing(init))
init <- NULL
# Check for penalized terms
pterms <- sapply(m, inherits, "coxph.penalty")
if(any(pterms)) {
pattr <- lapply(m[pterms], attributes)
#
# the 'order' attribute has the same components as 'term.labels'
#   pterms always has 1 more (response), sometimes 2 (offset)
# drop the extra parts from pterms
tempf <- untangle.specials(Terms, "frailty", 1:10)
ord <- attr(Terms, "order")[tempf$terms]
if(any(ord > 1))
stop("Penalty terms cannot be in an interaction")
pcols <- (attr(X, "assign")[-1])[tempf$vars]
#penalized are hard sometimes
if(control$eps.miss) control$eps <- 1e-07
if(control$iter.miss)
control$iter.max <- 20
fit <- coxpenal.fit(X, Y, strats, offset, init = init, control,
weights = weights, method = method, row.names(m), pcols,
pattr)
}
```

```
else {
if(method == "breslow" || method == "efron") {
if(type == "right")
fitter <- get("coxph.fit")
else fitter <- get("agreg.fit")
}
else if(method == "exact")
fitter <- get("agexact.fit")
else stop(paste("Unknown method", method))
fit <- fitter(X, Y, strats, offset, init, control, weights =
weights, method = method, row.names(m))
}
if(is.character(fit)) {
fit <- list(fail = fit)
oldClass(fit) <- "coxph"
}
else {
if(any(is.na(fit$coef))) {
vars <- (1:length(fit$coef))[is.na(fit$coef)]
msg <- paste("X matrix deemed to be singular; variable",
paste(vars, collapse = " "))
if(singular.ok)
warning(msg)
else stop(msg)
}
fit$n <- nrow(Y)
oldClass(fit) <- fit$method[1]
fit$terms <- Terms
fit$assign <- attr(X, "assign")
if(robust) {
fit$naive.var <- fit$var
fit$method <- method
# a little sneaky here: by calling resid before adding the
#   na.action method, I avoid having missings re-inserted
# I also make sure that it doesn't have to reconstruct X and Y
fit2 <- c(fit, list(x = X, y = Y, weights = weights))
if(length(strats))
fit2$strata <- strata.keep
if(length(cluster)) {
temp <- residuals.coxph(fit2, type = "dfbeta",
collapse = cluster, weighted = T)
# get score for null model
```

```
if(is.null(init)) fit2$linear.predictors <- 0 *
fit$linear.predictors else fit2$
linear.predictors <- c(X %*%
init)
temp0 <- residuals.coxph(fit2, type = "score",
collapse = cluster, weighted = T)
}
else {
temp <- residuals.coxph(fit2, type = "dfbeta",
weighted = T)
fit2$linear.predictors <- 0 * fit$
linear.predictors
temp0 <- residuals.coxph(fit2, type = "score",
weighted = T)
}
fit$var <- t(temp) %*% temp
u <- apply(as.matrix(temp0), 2, sum)
fit$rscore <- coxph.wtest(t(temp0) %*% temp0, u,
control$toler.chol)$test
}
#Wald test
if(length(fit$coef) && is.null(fit$wald.test)) {
#not for intercept only models, or if test is already done
nabeta <- !is.na(fit$coef)
if(is.null(init))
temp <- fit$coef[nabeta]
else temp <- (fit$coef - init)[nabeta]
fit$wald.test <- coxph.wtest(fit$var[nabeta, nabeta],
temp, control$toler.chol)$test
}
na.action <- attr(m, "na.action")
if(length(na.action))
fit$na.action <- na.action
if(model)
fit$model <- m
else {
if(x) {
fit$x <- X
if(length(strats))
fit$strata <- strata.keep
}
if(y)
```

```
fit$y <- Y
}
}
if(!is.null(weights) && any(weights != 1))
fit$weights <- weights
fit$formula <- as.vector(attr(Terms, "formula"))
fit$call <- call
fit$method <- method
fit
}
```

B.4 Markov Chain Monte Carlo simulation code

This is the C-code (Kernighan and Ritchie 1978) used for the simulations in chapter 9, in particular the sceptical Bayesian analysis.

To compile the code, on a unix platform, use the command line

```
> gcc -lm prostpart.c -o prostpart
```

where `prostpart.c` is the code below. To use the the code

```
> prostpart infile outfile 1000
```

The programme will perform 1000 iterations reading the data in `infile` and writing the simulated random variables to `outfile`. The input file uses variables which have undergone a linear transformation so that some of the columns are orthogonal; the first five lines of this input file are in section A.4. Also, the programme writes, to the standard interface, simulations from the Pólya tree C.D.F. which can be redirected to another file or programme.

```
#include <stdio.h>
```

```
#include <math.h>
#include <stdlib.h>
#include<plot.h>
#include<time.h>

#define MAXSAMPLE 10000
#define MAXSIMS 10000
#define PI 3.1415926536

float critical[] = {
0, -0.674489750196082, 0.674489750196082, -1.15034938037601,
-0.318639363964375, 0.318639363964375, 1.15034938037601,
-1.53412054435255, -0.887146559018876, -0.48877641111467,
-0.157310684610171, 0.157310684610171, 0.48877641111467,
0.887146559018876, 1.53412054435255, -1.86273186742165,
-1.31801089730354, -1.00999016924958, -0.776421761147928,
-0.579132162255556, -0.402250065321725, -0.237202109328788,
-0.0784124127331122, 0.0784124127331121, 0.237202109328788,
0.402250065321725, 0.579132162255556, 0.776421761147928,
1.00999016924958, 1.31801089730354, 1.86273186742165,
-2.15387469406146, -1.67593972277344, -1.41779713799627,
-1.22985875921659, -1.07751556704028, -0.946781756301046,
-0.830510878205399, -0.724514383492366, -0.626099012346422,
-0.533409706241281, -0.445096524985517, -0.36012989178957,
-0.277690439821577, -0.197099084294312, -0.117769874579095,
-0.0391760855030976, 0.0391760855030977, 0.117769874579095,
0.197099084294312, 0.277690439821577, 0.36012989178957,
0.445096524985517, 0.53340970624128, 0.626099012346421,
0.724514383492366, 0.830510878205399, 0.946781756301046,
1.07751556704028, 1.22985875921659, 1.41779713799627,
1.67593972277344, 2.15387469406146, -2.41755901623651,
-1.9874278859299, -1.76167041036307, -1.60100866488608,
-1.4734675779471, -1.3662038163721, -1.27269864119054,
-1.18916435019934, -1.11319427716093, -1.04315826331845,
-0.977897543940542, -0.916556667533113, -0.858484474141833,
-0.803172565597918, -0.750215375467941, -0.69928330238322,
-0.650104070647995, -0.602449453164424, -0.556125593618691,
-0.510965806738248, -0.46682512285259, -0.4235760842012,
-0.381105454763556, -0.339311606538817, -0.298102412930487,
-0.257393526100938, -0.21710694721013, -0.17716982099174,
-0.137513402144336, -0.0980721524886611, -0.0587829360689431,
-0.0195842852301269, 0.0195842852301269, 0.0587829360689431,
```

```
0.098072152488661, 0.137513402144336, 0.17716982099174,
0.21710694721013, 0.257393526100938, 0.298102412930487,
0.339311606538817, 0.381105454763557, 0.4235760842012,
0.46682512285259, 0.510965806738248, 0.556125593618692,
0.602449453164424, 0.650104070647995, 0.69928330238322,
0.750215375467941, 0.803172565597918, 0.858484474141832,
0.916556667533112, 0.977897543940542, 1.04315826331845,
1.11319427716093, 1.18916435019934, 1.27269864119054,
1.3662038163721, 1.4734675779471, 1.60100866488608,
1.76167041036307, 1.9874278859299, 2.41755901623651,
-2.66006746861747, -2.26622680920966, -2.06352789831625,
-1.92135077429371, -1.80989223848061, -1.71722811750574,
-1.63732538276806, -1.56668858606841, -1.50310294312927,
-1.44507257981808, -1.3915374879959, -1.34171784108025,
-1.29502240670581, -1.25099171546255, -1.20926123170916,
-1.16953661020714, -1.13157655838619, -1.09518065276139,
-1.06018047943536, -1.02643306313791, -0.993815907860883,
-0.962223195295421, -0.931562830007115, -0.9017541138301,
-0.872725894627041, -0.844415077375257, -0.816765415315091,
-0.789726519943266, -0.763253043732571, -0.737304000438654,
-0.71184219593942, -0.686833748574731, -0.662247682488414,
-0.638055580922517, -0.614231289060245, -0.590750658062819,
-0.567591323544569, -0.544732512988176, -0.522154877598002,
-0.499840344883735, -0.477771988903886, -0.455933915613139,
-0.43431116117521, -0.412889601443654, -0.391655871092592,
-0.370597291109629, -0.349701803553895, -0.328957912640491,
-0.308354631344837, -0.287881432831012, -0.267528206101097,
-0.247285215340805, -0.227143062502715, -0.20709265272436,
-0.187125162225721, -0.16723200837085, -0.147404821612355,
-0.12763541906627, -0.107915779489187, -0.0882380194499245,
-0.0685943705051181, -0.0489771572021319, -0.0293787757441571,
-0.00979167316134537, 0.00979167316134536, 0.0293787757441571,
0.048977157202132, 0.0685943705051181, 0.0882380194499244,
0.107915779489187, 0.12763541906627, 0.147404821612355,
0.16723200837085, 0.187125162225721, 0.20709265272436,
0.227143062502715, 0.247285215340805, 0.267528206101097,
0.287881432831012, 0.308354631344837, 0.328957912640491,
0.349701803553895, 0.370597291109629, 0.391655871092592,
0.412889601443654, 0.43431116117521, 0.455933915613139,
0.477771988903886, 0.499840344883735, 0.522154877598002,
0.544732512988176, 0.567591323544569, 0.590750658062819,
0.614231289060245, 0.638055580922517, 0.662247682488414,
```

```
0.686833748574731, 0.711842195939419, 0.737304000438655,
0.76325304373257, 0.789726519943266, 0.816765415315091,
0.844415077375258, 0.87272589462704, 0.9017541138301,
0.931562830007114, 0.962223195295421, 0.993815907860883,
1.02643306313791, 1.06018047943536, 1.09518065276139,
1.13157655838619, 1.16953661020714, 1.20926123170916,
1.25099171546255, 1.29502240670581, 1.34171784108025,
1.3915374879959, 1.44507257981808, 1.50310294312927,
1.56668858606841, 1.63732538276806, 1.71722811750574,
1.80989223848061, 1.92135077429371, 2.06352789831624,
2.26622680920966, 2.66006746861747  }; /*Pr( N(0,1)>2.66)=2^(-8) */


struct tnode {
float crit;
float P;
float level;
struct tnode *left;
struct tnode *right;
};


void metrop( float **param, int j, int m, float ***datapointers,
 struct tnode *polya1, struct tnode *polya0, int sampdim);
float posterior( float **param, int m, float ***datapointers,
 struct tnode *polya1, struct tnode *polya0);
float dnorm( float y1, float y01, float mu1, float mu0,
  float sigma1, float sigma0);
float dgamma( float x1, float x0, float l1, float l0, float a);
float dpois( float y1, float y0, float mu1, float mu2);
float dbinomial( float r1, float r0, float p1, float p0,
 float n);
float dpolya( float b1, float b0, struct tnode *polya1,
 struct tnode *polya0);
struct tnode *addtree( struct tnode *p, float critical,
 float level);
void samppolya(struct tnode *where1, struct tnode *where0,
 struct tnode *root1, struct tnode *root0, float **param,
 float ***datapointers, int m);
float proppolya( struct tnode *where1);
float polyaprior( struct tnode *where1, struct tnode *where0);
float polyaexpect( struct tnode *root, float (*function)(float,
```

```
 float ), float lower, float upper, float arg);
float moment1( float x);
float moment2( float x);
float indicator( float x, float arg);
float factorial( float x);
float proposal( float previous, float range);
void Ppolya(struct tnode *where, float x , int m, float *p);
void tnodecopy( struct tnode *nodeA, struct tnode *nodeB);
float lowB(struct tnode *where);
float uppB(struct tnode *where);
float**transpose (float **input, int ncol, int nrow);
int sortunique( float *time, float *x, float *delta, int m);
float max(float *list, int listlength, float lower);
float scale( float *x, int n);
float mvnormal( float *xnew, float *xold, float *mu, float **sigmainv,
 int dim);

main(int argc, char *argv[])
{


FILE * fdata, *foutput;
float **data, **param, **Y, **dN, ***datapointers,
     *mun[1], *scalers;
float quant[] = {
-3.0, -2.0, -1.0, 0.0, 1.0, 2.0, 3.0  };
float starters[] = {
-38.7, -23.1, 5.68, -13.9, 0.0932, 0.338, -0.0753,
0.47, -0.299, -0.225, 0.421, -1.43, -0.0539, -0.0673, 0.532,
0.0349, 0.602, 0.759, -1.15, -0.0371, -0.00496, -0.00491,
0.332, 0.413, 0., 0., 0.
    /* fixed effects ends, random effects starts */
, -2.67, 0.232, 1.83, -1.65, 1.05, 0.0846, -1.55,
0.135, -1.86, -0.284, -1.11, -0.526, -0.0351, 1.54, 1.07,
-0.0423, 1.48, 1.35, -0.872, 0.437, 0.414, 0.218, 1.28, 0.374,
-0.464, -1.99, 0.554, 0.695, 1.63, 0.277, 0.306, 1.02, 0.845,
-1.48, 1.07, 1.04, 1.2, 1.08, 1.43, -1.76, 0.0118, -1.38,
0.0572, -0.335, 1.6, -1.07, -3.64, 1.47, 0.781, 1.69, 1.73,
-0.436, 1.5, 0.597, 1.81, 0.677, 0.524, 2.06, 1.26, 0.945,
2.32, -0.762, 0.966, 0.776, 0.558, -0.101, 1.37, -2.86, -2.38,
-0.491, -3.17, 0.456, -1.01, 1.95, -1.24, 2.1, 1.33, 1.15,
-0.15, -0.914, -0.403, 1.9, 1.55, -2.12, -1.63, 1.5, 0.746, 2,
```

1.72, 0.743, -0.928, -0.141, -1.31, 0.206, -0.593, 1.87,
-0.265, 0.103, -1.98, 2.22, 0.479, 1.36, -2.64, -3, -3.42,
0.832, 2.02, 0.905, 1.36, -1.91, 2.18, -1.21, -1.4, 1.1, 2.18,
-0.202, -2.15, -0.18, 2.11, 2.13, -2.24, 1.35, -0.947, 1.37,
1.79, 2.22, 2.07, -1.16, -0.177, 1.5, 2.27, 0.224, 0.697, 2.1,
1.38, 1.14, -0.704, 1.24, 0.753, -2.53, 1.06, -2.63, 1.39,
1.95, 0.276, 1.78, 2.11, -0.847, -1.46, 0.359, -0.728, 0.185,
-1.64, 1.84, 0.382, 0.274, -0.46, -2.66, -3.08, 0.0378, -2.44,
1.33, 2.1, -0.379, 0.606, -0.586, 1.72, -0.75, -1.28, 1.43,
1.01, 1.45, 0.107, 2.15, 1.08, -1.73, -1.65, 0.934, -1.46,
0.483, -1.16, -0.562, -2.37, -0.24, 1.04, -1.01, -0.139,
-0.317, -1.43, 2.12, 2.2, 2.32, -0.622, -1.04, -2.7, -1.44,
-0.846, -2.1, -1.11, 0.561, -0.423, 1.87, 1.32, 1.19, 0.586,
2.43, 2.69, -1.85, 1.71, -0.706, -0.49, -0.44, 0.161, -2.89,
0.98, -2.29, -2.15, 0.625, 0.2, 0.604, 2.49, -0.875, -0.718,
-0.379, 0.583, 1.23, -0.777, 1.5, -3.24, 1.24, 0.659, 1.92,
2.32, -1.67, 0.871, 0.21, -1.69, -0.36, -0.939, 0.664, 0.521,
1.39, -0.556, -1.13, -3.06, -0.452, -0.792, -1.87, 2.31,
-2.28, 1.12, -1.06, 1.96, 0.79, 1.62, -1.24, 0.592, 2.05,
-0.261, 0.764, -1.61, 1.54, 0.299, -0.153, -0.485, -1.44,
1.04, 1.9, 0.756, -2.9, 1.38, 1.37, 0.235, -1.79, 1.46, 0.875,
1.88, 1.16, -0.287, 1.11, -0.815, -2.69, -4.12, -2.11, 0.709,
-2.63, -3.11, 0.207, -2.32, -2.87, 1.17, -1.38, -2.59, -0.131,
-2.54, 0.844, 0.381, 1.84, -2.11, -1.83, 0.805, 1.34, -1.83,
-1.35, -2.29, -2.37, 1.74, -2.6, -2.48, -1.23, -1.85, 2.26,
0.0941, 1.25, 1, -1.37, 2.31, -0.4, 1.71, -1.97, 1.16, -0.476,
0.39, 1.53, 0.474, 2.52, 1.91, -0.148, -0.862, 2.09, 0.197,
1.56, 0.463, 1.97, -1.03, 0.534, -1.1, -0.028, -0.627, 0.752,
1.38, 1.17, 1.36, 0.198, 0.19, 0.731, -0.132, -0.02, -0.056,
2.52, 2.1, -0.409, -0.632, 0.569, 2.09, -0.276, 0.659, -2.31,
-1.91, -2.08, 0.208, 0.494, 1.21, 1.7, 1.32, -0.409, -3.09,
-3.43, 0.97, 0.949, 2.52, -0.00847, 1.63, -0.181, -1.96,
-0.993, 0.843, -1.54, 1.68, 1.45, -1.94, 2.39, 0.721, -2.23,
2.08, 0.438, -1.07, 0.139, 1.09, 1.58, -0.399, 1.7, 1.94,
0.761, -0.965, 0.228, -1.91, 1.08, -0.0556, 0.468, -1.57,
0.565, 2.17, 1.81, -0.216, 0.331, 0.706, 2.25, -0.678, 1.36,
1.01, 1.71, -2.81, 0.522, -2.73, 0.0914, -1.36, 2.21, 1.7,
-0.728, -0.579, 0.791, 2.02, 2.23, -2.03, -1.53, 0.224, -0.67,
-1.41, 1.34, 1.74, 0.586, -0.78, -0.756, -0.623, 2.26, 0.775,
2.64, 1.19, 0.487, 1.94, 1.79, 2.1, -1.97, -1.23, 1.67, 0.636,
-0.0533, 1.82, -0.124, 2.08, 0.956, -0.923, 2.07, 1.42, 1.95,
1.48, 0.486, 1.53, -0.0289, -0.525, -1.55, 0.306, -1.27, 0.37,

```
0.927, -0.866, 2.28, 0.271, -1.28, 0.542, 2.04, -1.92, 1.76,
-1.55, -1.18, 0.247, 1.93, 0.982, 0.326, 2.43, 2.42, 0.503,
0.275, 0.00247   };

float muraw[] = {
-38.7442484590995, -23.1378682268593,
5.67649625670294, -13.907972716795, 0.0931783114425791,
0.337834364839979, -0.0752851004003422, 0.470224709915386,
-0.298785038532856, -0.225319467811483, 0.42084060477797,
-1.42775926202799, -0.0538508483055472, -0.0672816365946448,
0.532094280005891, 0.0349276596412233, 0.602066922021037,
0.759001567577583, -1.15488141529265, -0.0370977406594049,
-0.00495849763383467, -0.0049145609567213, 0.332246292917992,
0.412687043722702, 0.0, 0.0, 0.0   };

float sigmainvraw[] = {
0.000565, -0.000326, -0.00046, 5.18e-19, -0.00391,
1.63e-19, -0.00119, 4.51e-19, -0.000867, 3.91e-19,
-0.00064, -7.7e-18, -0.00912, 9.72e-16, -0.000237,
3.53e-17, -0.00169, 1.7e-16, -0.000624, 5.5e-18,
-0.00073, -6.84e-18, -0.000356, 3.43e-17, 9.17e-05,
-5.03e-17, -0.000234, -0.000326, 0.00194, -0.000977,
-0.00214, 0.00225, -2.13e-05, 0.000686, 3.87e-05,
0.0005, 0.000152, 0.000369, 0.00979, 0.00526, -0.0215,
0.000137, -0.000875, 0.000973, -0.00405, 0.00036,
1.39e-05, 0.000421, 9.14e-05, 0.000205, -0.000876,
-5.29e-05, 0.000833, 0.000135, -0.00046, -0.000977,
0.00138, 0.00159, 0.00318, 1.58e-05, 0.000971,
-2.86e-05, 0.000709, -0.000113, 0.000528, -0.00725,
0.00822, 0.0159, -0.000505, 0.000648, 0.00146, 0.003,
0.000416, -1.03e-05, 0.000644, -6.77e-05, 0.000317,
0.000649, -0.00011, -0.000617, 0.000234, 5.18e-19,
-0.00214, 0.00159, 0.0123, -3.78e-06, -0.00747,
9.2e-07, -0.00345, 1.71e-06, -0.00707, 3.67e-06,
-0.0524, 0.000451, 0.0284, -0.000402, 0.00714, 5e-05,
0.0126, -5.31e-05, -0.000438, 2.87e-05, 0.00293,
1.56e-05, 0.0046, -2.04e-05, 0.000484, 2.51e-05,
-0.00391, 0.00225, 0.00318, -3.78e-06, 0.027,
-3.76e-08, 0.00823, 6.83e-08, 0.006, 2.68e-07,
0.00443, 1.73e-05, 0.0631, -3.79e-05, 0.00168,
-1.55e-06, 0.0117, -7.15e-06, 0.00432, 2.45e-08,
0.00505, 1.61e-07, 0.00246, -1.55e-06, -0.000633,
```

1.47e-06, 0.00162, 1.63e-19, -2.13e-05, 1.58e-05,
-0.00747, -3.76e-08, 8.48, 9.15e-09, -3.81, 1.7e-08,
0.268, 3.65e-08, -1.7, 4.49e-06, -0.231, -4e-06,
-1.46, 4.98e-07, -0.471, -5.28e-07, -0.298, 2.86e-07,
-0.459, 1.55e-07, -0.599, -2.03e-07, 0.856, 2.5e-07,
-0.00119, 0.000686, 0.000971, 9.2e-07, 0.00823,
9.15e-09, 0.00251, -1.66e-08, 0.00183, -6.53e-08,
0.00135, -4.21e-06, 0.0192, 9.23e-06, 0.00049,
3.76e-07, 0.00355, 1.74e-06, 0.00131, -5.97e-09,
0.00154, -3.93e-08, 0.000749, 3.77e-07, -0.000194,
-3.58e-07, 0.000494, 4.51e-19, 3.87e-05, -2.86e-05,
-0.00345, 6.83e-08, -3.81, -1.66e-08, 6.33, -3.09e-08,
0.199, -6.62e-08, -0.00206, -8.14e-06, 0.12, 7.25e-06,
-0.105, -9.03e-07, -1.27, 9.58e-07, -0.0119,
-5.18e-07, 0.00783, -2.82e-07, 0.33, 3.68e-07, -0.539,
-4.53e-07, -0.000867, 0.0005, 0.000709, 1.71e-06,
0.006, 1.7e-08, 0.00183, -3.09e-08, 0.00143,
-1.21e-07, 0.000992, -7.82e-06, 0.014, 1.72e-05,
0.000348, 6.99e-07, 0.00258, 3.24e-06, 0.000926,
-1.11e-08, 0.00112, -7.3e-08, 0.000543, 7e-07,
-0.00014, -6.66e-07, 0.000358, 3.91e-19, 0.000152,
-0.000113, -0.00707, 2.68e-07, 0.268, -6.53e-08,
0.199, -1.21e-07, 0.0346, -2.6e-07, -0.0781, -3.2e-05,
-0.00906, 2.85e-05, -0.101, -3.55e-06, -0.12,
3.76e-06, -0.0189, -2.04e-06, -0.0297, -1.11e-06,
-0.0167, 1.44e-06, 0.0151, -1.78e-06, -0.00064,
0.000369, 0.000528, 3.67e-06, 0.00443, 3.65e-08,
0.00135, -6.62e-08, 0.000992, -2.6e-07, 8.44,
-1.68e-05, -0.582, 3.68e-05, -1.34, 1.5e-06, 0.272,
6.94e-06, -0.163, -2.38e-08, -0.673, -1.57e-07, -1.19,
1.5e-06, -0.261, -1.43e-06, 0.108, -7.7e-18, 0.00979,
-0.00725, -0.0524, 1.73e-05, -1.7, -4.21e-06,
-0.00206, -7.82e-06, -0.0781, -1.68e-05, 1.06,
-0.00206, -0.0894, 0.00184, 0.342, -0.000229, 0.256,
0.000243, 0.0807, -0.000131, 0.104, -7.14e-05, 0.0619,
9.31e-05, -0.166, -0.000115, -0.00912, 0.00526,
0.00822, 0.000451, 0.0631, 4.49e-06, 0.0192,
-8.14e-06, 0.014, -3.2e-05, -0.582, -0.00206, 0.194,
0.00452, 0.0932, 0.000184, 0.00881, 0.000853, 0.021,
-2.93e-06, 0.0594, -1.92e-05, 0.0894, 0.000185,
0.0166, -0.000175, -0.0035, 9.72e-16, -0.0215, 0.0159,
0.0284, -3.79e-05, -0.231, 9.23e-06, 0.12, 1.72e-05,

-0.00906, 3.68e-05, -0.0894, 0.00452, 0.274, -0.00403,
0.049, 0.000502, 0.0591, -0.000533, 0.00734, 0.000288,
0.0111, 0.000157, 0.0283, -0.000204, -0.0337,
0.000252, -0.000237, 0.000137, -0.000505, -0.000402,
0.00168, -4e-06, 0.00049, 7.25e-06, 0.000348,
2.85e-05, -1.34, 0.00184, 0.0932, -0.00403, 5.56,
-0.000164, -0.017, -0.00076, 0.483, 2.61e-06, 0.186,
1.71e-05, -0.0492, -0.000164, 0.0907, 0.000156,
-0.368, 3.53e-17, -0.000875, 0.000648, 0.00714,
-1.55e-06, -1.46, 3.76e-07, -0.105, 6.99e-07, -0.101,
1.5e-06, 0.342, 0.000184, 0.049, -0.000164, 0.387,
2.04e-05, 0.334, -2.17e-05, 0.076, 1.17e-05, 0.114,
6.38e-06, 0.099, -8.32e-06, -0.119, 1.03e-05,
-0.00169, 0.000973, 0.00146, 5e-05, 0.0117, 4.98e-07,
0.00355, -9.03e-07, 0.00258, -3.55e-06, 0.272,
-0.000229, 0.00881, 0.000502, -0.017, 2.04e-05, 4.14,
9.46e-05, 0.514, -3.25e-07, -0.263, -2.13e-06, 0.0777,
2.05e-05, -1.25, -1.95e-05, -0.00554, 1.7e-16,
-0.00405, 0.003, 0.0126, -7.15e-06, -0.471, 1.74e-06,
-1.27, 3.24e-06, -0.12, 6.94e-06, 0.256, 0.000853,
0.0591, -0.00076, 0.334, 9.46e-05, 1.9, -0.0001,
0.0795, 5.43e-05, 0.026, 2.95e-05, 0.0449, -3.85e-05,
-0.347, 4.75e-05, -0.000624, 0.00036, 0.000416,
-5.31e-05, 0.00432, -5.28e-07, 0.00131, 9.58e-07,
0.000926, 3.76e-06, -0.163, 0.000243, 0.021,
-0.000533, 0.483, -2.17e-05, 0.514, -0.0001, 0.625,
3.45e-07, -0.00154, 2.27e-06, -0.00539, -2.17e-05,
-0.137, 2.07e-05, -0.162, 5.5e-18, 1.39e-05,
-1.03e-05, -0.000438, 2.45e-08, -0.298, -5.97e-09,
-0.0119, -1.11e-08, -0.0189, -2.38e-08, 0.0807,
-2.93e-06, 0.00734, 2.61e-06, 0.076, -3.25e-07,
0.0795, 3.45e-07, 0.0161, -1.86e-07, 0.0216,
-1.01e-07, 0.0198, 1.32e-07, -0.0291, -1.63e-07,
-0.00073, 0.000421, 0.000644, 2.87e-05, 0.00505,
2.86e-07, 0.00154, -5.18e-07, 0.00112, -2.04e-06,
-0.673, -0.000131, 0.0594, 0.000288, 0.186, 1.17e-05,
-0.263, 5.43e-05, -0.00154, -1.86e-07, 0.0711,
-1.23e-06, 0.0847, 1.18e-05, 0.0951, -1.12e-05,
-0.0151, -6.84e-18, 9.14e-05, -6.77e-05, 0.00293,
1.61e-07, -0.459, -3.93e-08, 0.00783, -7.3e-08,
-0.0297, -1.57e-07, 0.104, -1.92e-05, 0.0111,
1.71e-05, 0.114, -2.13e-06, 0.026, 2.27e-06, 0.0216,

```
-1.23e-06, 0.038, -6.66e-07, 0.0303, 8.69e-07,
-0.0222, -1.07e-06, -0.000356, 0.000205, 0.000317,
1.56e-05, 0.00246, 1.55e-07, 0.000749, -2.82e-07,
0.000543, -1.11e-06, -1.19, -7.14e-05, 0.0894,
0.000157, -0.0492, 6.38e-06, 0.0777, 2.95e-05,
-0.00539, -1.01e-07, 0.0847, -6.66e-07, 8.54,
6.39e-06, 0.096, -6.07e-06, 0.56, 3.43e-17, -0.000876,
0.000649, 0.0046, -1.55e-06, -0.599, 3.77e-07, 0.33,
7e-07, -0.0167, 1.5e-06, 0.0619, 0.000185, 0.0283,
-0.000164, 0.099, 2.05e-05, 0.0449, -2.17e-05, 0.0198,
1.18e-05, 0.0303, 6.39e-06, 5.25, -8.33e-06, -0.0825,
1.03e-05, 9.17e-05, -5.29e-05, -0.00011, -2.04e-05,
-0.000633, -2.03e-07, -0.000194, 3.68e-07, -0.00014,
1.44e-06, -0.261, 9.31e-05, 0.0166, -0.000204, 0.0907,
-8.32e-06, -1.25, -3.85e-05, -0.137, 1.32e-07, 0.0951,
8.69e-07, 0.096, -8.33e-06, 0.387, 7.92e-06, 3.88e-05,
-5.03e-17, 0.000833, -0.000617, 0.000484, 1.47e-06,
0.856, -3.58e-07, -0.539, -6.66e-07, 0.0151,
-1.43e-06, -0.166, -0.000175, -0.0337, 0.000156,
-0.119, -1.95e-05, -0.347, 2.07e-05, -0.0291,
-1.12e-05, -0.0222, -6.07e-06, -0.0825, 7.92e-06,
0.183, -9.77e-06, -0.000234, 0.000135, 0.000234,
2.51e-05, 0.00162, 2.5e-07, 0.000494, -4.53e-07,
0.000358, -1.78e-06, 0.108, -0.000115, -0.0035,
0.000252, -0.368, 1.03e-05, -0.00554, 4.75e-05,
-0.162, -1.63e-07, -0.0151, -1.07e-06, 0.56, 1.03e-05,
3.88e-05, -9.77e-06, 0.0954   };


float prob[7];
char c[10];
struct tnode *polya1 = NULL;
struct tnode *polya0 = NULL;
int m = 0, i, j, k, n, test, munique;

/*EDIT*/
int sampdim = 1000, datadim = 30;  /*EDIT for (
    upper bounds on) sampdim-dim sample space, data
    dim-dim data-space */

param = (float **) malloc( (unsigned) 2 * sizeof(float
    *));
```

```
data =  (float **) malloc( (unsigned) MAXSAMPLE
    *sizeof(float * ));
/*set the random seed */
srand(26);

for ( i = 0; i < 255; i++) {
polya1 = addtree(polya1, critical[i], 1.0);
polya0 = addtree(polya0, critical[i], 1.0);
}



fdata = fopen(argv[1], "r");
while (test != EOF) {
*(data + m) = (float *) malloc( (unsigned) (datadim
    + 1) * sizeof(float));
/* the (datadim+x) allows for x extra v
    ariables to be calculated) */
for ( j = 0; j < datadim; j++) {
test = fscanf(fdata, "%f", *(data
    + m) + j);
}
m++;
}

/*TRANSFORM DATA HERE */
data = transpose( data, datadim + 1, m);
/*CHECK that you have the corect columns for ti
    mes, and deltas */
munique = sortunique( data[0], data[30], data[1],
     m);
Y = (float **) malloc( (unsigned) munique * sizeof( float
    *));
dN = (float **) malloc( (unsigned) munique * sizeof( float
    *));
for ( j = 0; j < munique; j++) {
Y[j] = (float *) malloc( (unsigned) m *
    sizeof(float));
dN[j] = (float *) malloc( (unsigned) m
    *sizeof(float));
for ( i = 0; i < m; i++) {
```

```
Y[j][i] = ( data[0][i] >= data[30][j] );
dN[j][i] = ( fabs(data[30][j] -
    data[0][i]) < 0.0001 );
}
}


/* CHECK to cope with data in different sized a
    rrays we have pointers to the different arrays.
     Ammend as  neccessary */
datapointers = (float ***) malloc( (unsigned) 6
    *sizeof( float * *));
datapointers[0] = data;
datapointers[1] = Y;
datapointers[2] = dN;
mun[0] = (float *)malloc( (unsigned) 3 * sizeof(float));
mun[0][0] = (float) munique;
/* determines the number of unobserved frailtie
    s */
mun[0][1] = max( data[2], m, 0.0);
mun[0][2] = 1; /* use to indicate to Posterior
    that this is the first iteration */
datapointers[3] = mun;
sampdim = 27 + (int) max( data[2], m, 0.0);
datapointers[5] = (float **) malloc( (unsigned) 27
    *sizeof(float *));
for ( i = 0; i < 27; i++) {
datapointers[5][i] = (float *) malloc( (unsigned) 27
    *sizeof(float ));
for ( j = 0; j < 27; j++) {
datapointers[5][i][j] = sigmainvraw[27*i+j];
}
}

datapointers[4] = (float **) malloc( (unsigned) 1
    *sizeof(float *));
datapointers[4][0] = muraw;

for ( i = 0; i < 2; i++) {
param[i] = (float *) malloc( (unsigned) sampdim
    *sizeof(float));
for ( j = 0; j < sampdim; j++) {
```

```
param[i][j] = starters[j];   /*E
    DIT starting values */
}
}

/*sampling proper */

if ( argc == 3) {
n = 10000;
} else {
n = (int) atof( argv[3]);
}
/*command line or default sample size */

foutput = fopen(argv[2], "w");
for ( i = 0; i < n ; ++i) {
/* sample from the polya tree */
samppolya(polya1, polya0,  polya1, polya0,
     param , datapointers, m) ;
for (j = 0; j < sampdim; ++j) {
metrop(param, j, m, datapointers,
     polya1, polya0 , sampdim);
fprintf( foutput, "%f\t", param[1][j]);
}

for (k = 0; k < 7; ++k) {
prob[k] = polyaexpect(polya1, indicator,
     -100, 100, quant[k]);
printf( "%f\t", prob[k]);
}
printf("\n");
fprintf(foutput,  "\n");
}
}


float posterior(float **param, int m, float ***datapointers,
 struct tnode *polya1, struct tnode *polya0)
{

/*EDIT compute the product of likelihood ratio
     and prior ratio p(new)/p(old) */
```

```
float *beta[2], *frail[2], *HR[2], **data, **Y,
     **dN, eta[2], d ;
double sum[2], update;
int dimfrail;

int i, j, k, munique;
data = datapointers[0];
Y = datapointers[1];
dN = datapointers[2];
munique = (int) datapointers[3][0][0];
dimfrail = (int) datapointers[3][0][1];




for ( i = 0; i < 2; i++) {
HR[i] = ( float *) malloc( (unsigned) m
    *sizeof(float));
frail[i] = param[i] + 27;
beta[i] = param[i];
}

update = log( polyaprior(polya1, polya0) );
update += mvnormal( beta[1], beta[0], datapointers[4][0],
     datapointers[5], 27);
for (j = 0; j < dimfrail; j++) {
update += log(dpolya(frail[1][j], frail[0][j],
     polya1, polya0));

}
/*the prior ratio */
for ( i = 0; i < m; i++) {
eta[1] = eta[0] = 0.0;
for ( k = 0; k < 27; k++) {
eta[1] += beta[1][k] * data[k+3][i];
eta[0] += beta[0][k] * data[k+3][i];
}

HR[1][i] = exp(eta[1] + frail[1][(int) data[2][i]-1]);
HR[0][i] = exp(eta[0] + frail[0][(int) data[2][i]-1]);
}
```

```
for ( j = 0; j < munique; j++) {
/* d copes with tied failure times */

sum[1] = sum[0] = d = 0;

for ( i = 0; i < m; i++) {
if ( Y[j][i] ) {
sum[1] += HR[1][i];
sum[0] += HR[0][i];
}
if ( dN[j][i] ) {
update += log( HR[1][i])
    -log( HR[0][i]) ;
d++;
}
}
update += d * (log(sum[0]) - log(sum[1]) );
/* printf( "update  j=%i , %f", j, upda
    te);  */
}
for ( i = 0; i < 2; i++) {
free(HR[i]);
}
/*  printf("log-lik= %f \n", update);     */
return (update > 1000) ? 1 : update;
}


void metrop(float **param, int j, int m, float ***datapointers,
 struct tnode *polya1, struct tnode *polya0, int sampdim)
{
int i;
float p;
double postratio;
/* range=(float *) malloc( (unsigned) sampdim*s
    izeof(float));*/
float range[] = {
20, 16.3, 10.5, 16.1, 14.1, 4.28, 14.7, 4.7, 14.1,
 14.7, 4.3, 9.58, 13.3, 12.6, 4.9, 12, 5.84, 8.19, 10.9, 15.6, 14.7,
 14.5, 4.46, 5.22, 12.6, 13.9, 12.6
```

```
 /* fixed effects ends here */
, 8.65, 8.6, 6.55, 6.55, 8.65, 6.55, 6.55, 8.65, 8.6,
8.75, 8.65, 8.7, 8.6, 6.65, 6.5, 7, 8.6, 6.55, 6.55, 8.7, 8.6, 6.75,
8.85, 8.65, 6.75, 6.6, 8.6, 8.6, 9.3, 6.55, 6.6, 6.5, 6.5, 7.1, 6.5,
8.6, 6.6, 6.5, 6.55, 6.7, 6.5, 6.6, 6.75, 6.5, 8.95, 6.6, 6.75, 6.65,
6.95, 6.55, 8.65, 8.65, 8.6, 6.7, 8.75, 6.55, 6.6, 6.5, 6.6, 6.5,
6.55, 6.55, 6.55, 6.6, 6.6, 6.5, 6.5, 8.7, 6.7, 6.55, 6.55, 6.5, 6.5,
6.55, 8.65, 6.55, 8.7, 6.75, 8.6, 6.5, 8.65, 8.6, 8.65, 6.5, 8.65,
6.6, 6.8, 8.8, 8.8, 6.55, 6.6, 6.55, 6.55, 7.65, 6.5, 8.65, 8.8,
6.55, 6.5, 6.8, 6.65, 8.6, 6.6, 8.6, 6.5, 6.5, 8.6, 6.5, 8.6, 8.6,
6.55, 6.5, 6.5, 6.5, 8.7, 8.6, 6.55, 6.5, 8.6, 8.6, 6.5, 6.55, 6.55,
6.75, 6.55, 6.6, 8.95, 6.55, 8.75, 6.5, 6.5, 6.85, 6.5, 6.5, 8.7,
6.85, 8.6, 8.65, 6.6, 6.6, 8.65, 6.5, 6.55, 6.55, 8.65, 8.9, 8.8,
8.6, 6.6, 8.7, 6.5, 6.5, 6.65, 6.55, 6.95, 6.5, 6.9, 8.7, 6.55, 6.55,
6.55, 6.6, 6.5, 6.55, 8.7, 8.65, 6.75, 8.65, 6.75, 6.65, 6.95, 9.1,
7.1, 6.55, 8.6, 6.55, 8.75, 7.1, 6.5, 6.5, 6.5, 6.55, 8.6, 8.65, 8.6,
8.6, 8.8, 8.9, 6.6, 8.7, 6.5, 6.5, 6.55, 8.6, 6.5, 6.5, 6.6, 6.6,
6.85, 6.65, 7.35, 7.45, 8.65, 6.65, 8.7, 8.85, 7.15, 6.6, 6.55, 6.5,
7.65, 8.65, 8.65, 8.6, 6.5, 6.6, 6.55, 8.9, 6.5, 6.8, 6.55, 6.5,
8.65, 6.55, 6.65, 6.7, 6.6, 6.65, 6.55, 6.55, 6.55, 7.2, 7, 8.75,
6.7, 6.75, 8.65, 6.5, 8.75, 6.55, 8.65, 6.5, 6.55, 6.6, 9.9, 8.65,
6.5, 6.65, 6.55, 8.7, 6.55, 6.95, 6.7, 6.65, 6.8, 6.5, 6.5, 7.4,
9.05, 6.55, 6.55, 6.6, 6.7, 6.6, 6.55, 6.5, 6.55, 6.65, 6.75, 8.65,
7.25, 8.9, 7.55, 6.6, 8.85, 8.65, 6.6, 8.6, 8.65, 6.75, 6.6, 8.65,
7.3, 8.65, 6.55, 6.55, 6.5, 8.7, 6.75, 6.55, 6.55, 6.65, 6.6, 8.65,
8.6, 6.55, 8.65, 8.65, 8.6, 8.65, 6.5, 6.55, 6.75, 6.55, 8.65, 6.5,
6.55, 6.5, 8.65, 6.55, 8.65, 6.55, 6.5, 6.6, 8.6, 6.5, 6.55, 6.55,
8.65, 6.5, 6.55, 6.55, 6.55, 6.5, 8.65, 6.55, 8.65, 8.6, 6.65, 6.55,
6.5, 6.6, 6.6, 8.6, 6.8, 6.6, 6.6, 8.6, 8.6, 6.5, 6.5, 8.65, 6.6,
6.55, 6.5, 6.7, 6.55, 9.1, 8.7, 8.9, 6.6, 6.6, 6.55, 6.55, 6.55,
6.65, 8.7, 8.75, 6.6, 6.6, 6.5, 6.8, 6.55, 6.6, 7.25, 6.7, 6.7, 8.7,
6.5, 6.55, 8.7, 6.5, 6.55, 7.3, 6.5, 6.55, 7, 6.6, 6.55, 6.5, 6.8,
6.5, 6.55, 6.7, 7.1, 6.65, 8.75, 6.55, 8.75, 6.55, 8.65, 6.55, 6.5,
6.55, 8.65, 6.7, 8.6, 6.55, 6.5, 8.6, 8.6, 6.55, 6.55, 8.75, 6.65,
8.8, 6.85, 8.7, 6.5, 6.55, 8.65, 6.75, 8.6, 6.5, 8.6, 6.5, 8.7, 8.65,
6.55, 6.75, 8.65, 6.55, 6.5, 6.55, 8.65, 8.6, 8.6, 6.5, 6.55, 6.5,
6.5, 8.6, 6.55, 6.55, 6.5, 8.7, 8.65, 6.5, 6.55, 6.6, 6.5, 6.65, 6.5,
7.15, 6.55, 6.5, 6.55, 6.55, 6.5, 8.6, 6.6, 8.6, 8.6, 8.6, 6.6, 6.8,
6.6, 6.7, 7.15, 6.6, 6.65, 6.6, 8.6, 8.6, 8.7, 6.5, 8.6, 6.7, 7.3,
8.65, 6.5, 8.65, 6.55, 8.6, 8.75, 8.6, 6.55, 6.55, 8.6, 6.5, 6.5,
6.6, 8.65, 6.65, 6.65   };
```

```
/*EDIT as appropriate to change range of random
     walk proposal distributions*/
param[0][j] = param[1][j];
param[1][j] = proposal( param[1][j], range[j]);
postratio = posterior(param, m, datapointers, polya1,
     polya0 );
p = (rand() + 1) / 32767.0;
if (log(p) > postratio) {
param[1][j] = param[0][j];
}

}


float proposal( float previous, float range)
{
float p;
p = range * (rand() / 32767.0 - 0.5);
return previous + p;
}
```

```
void samppolya( struct tnode *where1, struct tnode *where0,
 struct tnode *root1, struct tnode *root0, float **param,
 float ***datapointers, int m)
{
float ratio, p;
if ( where1 != NULL ) {
where0->P = where1->P;
where1->P = proppolya( where1 );
ratio = posterior( param, m, datapointers,
     root1, root0);
p = (rand() + 1) / 32767.0;
if (log(p) > ratio ) {
where1->P = where0->P;
}
samppolya( where1->left, where0->left,
    root1, root0, param, datapointers, m);
samppolya( where1->right, where0->right,
     root1, root0, param, datapointers, m);
}
```

```
}


float proppolya( struct tnode *where1)
{
float p, range;
p = rand() / 32767.0 - 0.5;
/* EDIT tinker with the sampling range */
range = 2.0 / sqrt( 0.01 * pow( 2, where1->level
    -1));
p = where1->P + p * range;
if ( (p > 0.0) && (p < 1.0)  && (where1->level >
    1.0) /* && (where1->level <5.0) */
    )
return p;
else
return where1->P;
}


float polyaprior( struct tnode *where1, struct tnode *where0)
{
float p;
if ( where1 == NULL)
return 1.0;
else {
p = pow( (where1->P) / (where0->P) * (1
    -where1->P) / (1 - where0->P), 0.01 * pow( 2,
     where1->level - 1) - 1 ) ;
return  polyaprior( where1->left, where0->left )
    *polyaprior(where1->right, where0->right)
    *p ;
/*EDIT change the hyper parameters for
    the polya tree */
}
}



/* ratios of standard densities */
```

```
float dnorm(float y1, float y0, float mu1, float mu0, float sigma1,
 float sigma0)
{
float answer;
answer = sigma0 / sigma1 * exp( -( (y1 - mu1) *
    (y1 - mu1) / sigma1 / sigma1 - (y0 - mu0) * (y0
    -mu0) / sigma0 / sigma0) / 2.0);
return answer;
}


float factorial( float x)
{
if ( x == 0)
return 1.0;
else
return  x * factorial(x - 1);
}


float dbinomial(float r1, float r0, float p1, float p0,
 float n)
{
float update;
if ( p1 < 0 ||  p1 > 1 || r1 < 0 || r1 > n || p0
    < 0 ||  p0 > 1 || r0 < 0 || r0 > n)
return 0.0;
else {
update =  pow(p1, r1) / pow(p0, r0);
update *= pow( 1 - p1, n - r1) / pow( 1
    -p0, n - r0);
update *= factorial( r0) / factorial(r1);
update *= factorial(n - r0) / factorial( n
    -r1);
return update;
}
}


float dgamma( float x1, float x0, float l1, float l0, float a)
{
if ( x1 > 0.0  && x0 > 0.0)
```

```
return pow(x1 * l1 / x0 / l0, a) * x0 /
    x1 * exp(-l1 * x1 + l0 * x0);
else
return 0.0;
}


float dpois( float y1, float y0, float mu1, float mu0)
{
return exp(-(mu1 - mu0)) * pow( mu1 / mu0, y1)
    *pow( mu0, y1 - y0) * factorial(y0) / factorial(y1);
}


float dpolya( float b1, float b0, struct tnode *polya1,
 struct tnode *polya0)
{
float update = 1.0, upper[2] = {
100.0, 100.0  },
lower[2] = {
-100.0, -100.0  };

struct tnode *where1 = polya1;
struct tnode *where0 = polya0;
/* EDIT the highest level of the polya tree (9-
    1)  */
while ( where1 != NULL && where1->level < 10.0 ) {
if ( b1 <= where1->crit) {
update *= where1->P;
upper[1] = where1->crit;
where1 = where1->left;

} else {
update *= (1 - where1->P);
lower[1] = where1->crit;
where1 = where1->right;
}
if ( b0 <= where0->crit) {
update /= where0->P;
upper[0] = where0->crit;
where0 = where0->left;
```

```
} else {
update /= (1 - where0->P);
lower[0] = where0->crit;
where0 = where0->right;

}
}
update *= (upper[0] - lower[0]) / (upper[1] - lower[1]);
if ( upper[1] == 100.0 || lower[1] == -100.0)
update *= (upper[1] - lower[1]) * exp(-b1
    *b1 / 2) / sqrt(2 * PI) * pow(2.0, (where1
    == NULL) ? 8.0 : (where1->level - 1.0) );
if ( upper[0] == 100.0 || lower[0] == -100.0)
update /= (upper[0] - lower[0]) * exp(-b0
    *b0 / 2) / sqrt(2 * PI) * pow(2.0, (where0
    == NULL) ? 8.0 : (where0->level - 1.0));
return update;
}


struct tnode *addtree( struct tnode *p, float critical,
 float level)
{
if ( p == NULL) {
p = (struct tnode *) malloc( sizeof(struct tnode ));
p->crit = critical;
p->P = 0.5;
p->left = p->right = NULL;
p->level = level;
} else {
level++;
if (    critical <= p->crit)
p->left = addtree( p->left, critical,
     level);
else
p->right = addtree( p->right, critical,
     level);
}
return p;

}
```

```
float polyaexpect( struct tnode *root, float (*function)(float,
 float), float lower, float upper, float arg)
{
if ( root->left == NULL)
return (function(upper, arg) + function(lower,
     arg)) / 2;
else
return (root->P) * polyaexpect( root->left,
     function, lower, root->crit, arg) + (1
    -root->P) * polyaexpect( root->right, function,
     root->crit, upper, arg);
}


float moment1( float x)
{
x = 1.0 * x;
return x;
}


float moment2( float x)
{
x = 1.0 * x;
return x * x;
}


float indicator( float x, float arg)
{
return (x <= arg) ? 1.0 : 0.0;
}


void Ppolya( struct tnode *where, float x, int m, float *p)
{
if ( where != NULL) {
if ( x <= where->crit) {
p[m] = where->P;
Ppolya( where->left, x, m + 1,
    p);
```

```
} else {
p[m] = 1 - where->P;
Ppolya( where->right, x, m + 1,
     p);
}
}
}


float lowB( struct tnode *where)
{
float p = 1.0;
struct tnode *now = where;
while (now != NULL) {
p *= now->P;
now = now->left;
}
return  - 0.01159805342 * 2 / p + 2.66006746861747;
}


float uppB( struct tnode *where)
{
float p = 1.0;
struct tnode *now = where;
while (now != NULL) {
p *= (1 - now->P);
now = now->right;
}
return  0.01159805342 * 2 / p - 2.66006746861747;
}


float**transpose (float **input, int ncol, int nrow)
{
int c, r;
float **output;
output = (float **) malloc( (unsigned) ncol * sizeof( float
    *));
for ( c = 0; c < ncol; c++) {
output[c] = (float *) malloc( (unsigned) nrow
    *sizeof( float));
```

```
for ( r = 0; r < nrow; r++) {
output[c][r] = input[r][c];
}
}
for ( r = 0; r < nrow; r++)
free( input[r]);
free( input);
return output;
}


float max( float *list, int listlength, float lower)
{
int i;
for ( i = 0; i < listlength; i++) {
if (list[i] > lower)
lower = list[i];
}
return lower;
}


int sortunique( float *time, float *x, float *delta, int m)
{
int i, j, k, n = 0;
for ( i = 0; i < m; i++) {
if ( delta[i] > 0.5) {
for ( j = 0; j < n || n == 0; j++) {
if (time[i] < x[j]) {
for (k = n; k >
    j; k--) {
/* code folded from here */
x[k] =
    x[k-1];
/* unfolding */
}
x[j] = time[i];
n++;
break;
} else if (time[i] == x[j]) {
break;
} else {
```

```
x[n] = time[i];
if ( j == n - 1
    || n == 0 )
/* code folded from here */
n++;
/* unfolding */
}
}
}
}
return n;
}


float scale(float *x, int n)
{
float mu = 0.0, oldmu = 0.0, s = 0.0;
int i;
for ( i = 0; i < n; i++) {
mu = (i * oldmu + x[i]) / (i + 1.0);
if ( i > 0) {
s = (i - 1) * s / ((float) i)  +
    ( x[i] - oldmu) * (x[i] - oldmu)
    / (i + 1.0);
}
oldmu = mu;
}
if (s > 0) {
for ( i = 0; i < n; i++) {
x[i] = (x[i] - mu) / sqrt(s);
}
}
return sqrt(s);
}


float mvnormal( float *xnew, float *xold, float *mu, float **sigmainv,
 int dim)
{
int i, j;
float sum = 0.0;
for ( i = 0; i < dim; i++) {
```

```
for (j = 0; j < dim; j++) {
sum += -(( xnew[i] - mu[i]) * ( sigmainv[i][j])
    *(xnew[j] - mu[j]) - ( xold[i]
    -mu[i]) * ( sigmainv[i][j]) * (xold[j]
    -mu[j])) / 2.0 ;
}
}
return sum;
}
```

B.5 Code for displaying the Pólya C.D.F.s

This code is written in Python (Lutz and Ascher (1999), `http://www.python.org`) and is used to display the simulation of the C.D.F.'s from a Pólya tree. It is suitable for the output of the `prostpart.c` programme. This can be applied directly to the output, where the file `test.py` is listed below.

```
> prostpart infile outfile 1000| python test.py -3 3.0001 7
```

For practical purposes, on the unix platform used by the author, the buffer management held up the display for minutes at a time and it was more effective to save the output and view it at a later time

```
> prostpart infile outfile 1000 > out
> cat out | python test.py -3 3.0001 7
```

The programme takes its input as a sequence of text lines which give the values of a C.D.F. at a fixed sequence of values. These values are described by the three numerical arguments: -3, 3.0001 and 7. These are the minimum value, the maximum value, and

the number of values per line. The code attempts to label these values to two significant figures, but a bug means that instead of -3,3,7 you have to enter -3,3.00001,7. A window appears which has four buttons: go/pause which starts and stops the animation; keep/discard which indicates whether each C.D.F. should be kept or erased; print which activates a window that can print to a file the graph currently displayed; quit which exits the programme. It requires the Python package to be installed on your platform.

```
from Tkinter import *
import sys, string, math
from tkFileDialog import asksaveasfilename

class Inter(Frame):

def __init__(self, Min=-10, Max=10, Num=4, master=None):
self.Min=Min*1.0
self.Max=Max*1.0
self.Num=Num
Frame.__init__(self,master)
self.createWidgets()
Grid.config(self)
self.update_idletasks()
self.evolve()

def createWidgets(self):
self.xw=350
self.yw=216
self.bb=40
self.iteration=0
self.numbervar=StringVar()
self.numbervar.set( "Iteration Number %d" %self.iteration)
self.NUMBER=Label( self, textvar=self.numbervar)
self.NUMBER.grid(row=0, column=1)
self.QUIT=Button( self, text='quit', command=master.destroy)
self.QUIT.grid(row=1,column=0)
self.PRINT=Button(self, text='print', command=self.postscript_print)
self.PRINT.grid(row=2, column=0)
self.legend=['go', 'pause']
self.flag=0
self.h=StringVar(); self.h.set(self.legend[self.flag])
```

```
self.HOLD=Button(self, textvar=self.h ,command=self.hold)
self.HOLD.grid(row=3, column=0)
self.recleg=['keep','discard']
self.rec=0
self.r=StringVar();self.r.set(self.recleg[self.rec])
self.REC=Button(self, textvar=self.r, command=self.record)
self.REC.grid(row=4, column=0)

self.draw=Canvas(self, width= self.xw+2*self.bb, height=self.yw+2*self.bb)
self.draw.line=self.draw.create_line(0,0,1,1)
self.axes()
self.draw.grid(row=1, rowspan=4,column=1)


def evolve(self):
self.update_idletasks()
if self.flag:
item=sys.stdin.readline()
try:
probs=map(float,string.split(item))[:(self.Num+1)]
coord=()
for i in range(self.Num):
coord=coord+( self.bb+ i *self.xw/(self.Num-1),self.bb+self.yw*(1-probs[i]))
if self.rec:
self.draw.itemconfig(self.draw.line, fill="grey")
else:
self.draw.delete( self.draw.line)
self.draw.line=self.draw.create_line(coord, fill="red",width=1)
self.after(1, self.evolve)
self.iteration=self.iteration+1
self.numbervar.set("Iteration number %d" %self.iteration)
except IndexError:
self.after( 1, self.evolve)
else:
self.after(1, self.evolve)


def axes(self):
self.draw.create_rectangle(self.bb-1, self.bb-1,
self.xw+self.bb+1, self.yw+self.bb+1, width=3 )
ylabs=(0.0, 0.1,0.25, 0.5, 0.75,0.9, 1)
for y in ylabs:
```

```
self.draw.create_text(self.bb-20,
self.bb+self.yw*(1-y),text=y)
self.draw.create_line(self.bb,self.bb+self.yw*(1-y),
self.bb -10 ,  self.bb+self.yw*(1-y)  , width=3  )
self.draw.create_line(self.bb,self.bb+self.yw*(1-y),
 self.bb +self.xw ,self.bb+self.yw*(1-y) , width=1)
xlab=range(self.Num)
for i in range(self.Num):
xlab[i]=self.Min + i*(self.Max - self.Min)/(self.Num -1)
k=int(math.log10(1.0*math.fabs(xlab[i])))
xlab[i]=math.floor(
xlab[i]*math.pow(10,k+2)+0.5)/math.pow(10,k+2)
# supposed to round to 3 significant figures
self.draw.create_text(self.bb+i*self.xw/(self.Num -1),self.bb-20, text=xlab[i])
self.draw.create_line(
self.bb+i*self.xw/(self.Num -1),self.bb -10,
self.bb+i*self.xw/(self.Num-1), self.bb, width=3)



def postscript_print(self):
fname=asksaveasfilename( defaultextension=".ps", title="File to hold PostScript")
if fname:
self.draw.postscript(file=fname)

def hold(self):
self.flag=1-self.flag
self.h.set(self.legend[self.flag])

def record(self):
self.rec=1-self.rec
self.r.set(self.recleg[self.rec])

if __name__ == "__main__":
master=Tk()
master.protocol("WM_DELETE_WINDOW", master.destroy)
if len(sys.argv)==4:
Min=string.atof(sys.argv[1])
Max=string.atof(sys.argv[2])
Num=string.atoi(sys.argv[3])
test=Inter(Min=Min, Max=Max, Num=Num)
else:
```

```
test=Inter()
test.mainloop()
```

Bibliography

Aalen, O. O.: 1976, Non-parametric inference in connection with multiple decrement models, *Scandanavian Journal of Statistics* **3**, 15–27.

Aalen, O. O.: 1978, Non-parametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.

Aalen, O. O.: 1980, A model for non-parametric regression analysis of counting processes, *in* W. Klonechi, A. Kozek and J. Rosiński (eds), *Mathematical Statistics and Probability*, Vol. 2 of *Lecture Notes in Statistics*, Springer, New York, pp. 1–25.

Abbring, J. H. and van den Berg, G. J.: 2003, The identifiability of the mixed proportional hazards competing risks model, *Journal of the Royal Statistical Society, B* **65**, 701–710.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N.: 1993, *Statistical Models Based on Counting Processes*, Springer, New York.

Andrews, D. F. and Herzberg, A. M.: 1985, *Data: a collection of problems from many fields for the student and research worker*, Springer, New York.

Barndorff-Nielsen, O. E. and Cox, D. R.: 1994, *Inference and Asymptotics*, Chapman and Hall, London.

Barron, A., Schervish, M. J. and Wasserman, L.: 1999, The consistency of posterior distributions in nonparametric problems, *Annals of Statistics* **27**, 536–61.

Becker, R. A., Chambers, J. M. and Wilks, A. R.: 1988, *The New S Language*, Chapman and Hall, London.

Bennet, S.: 1983, Analysis of survival data by the proportional odds model, *Statistics in Medicine* **2**, 273–277.

Best, N. G., Cowles, M. K. and Vines, S. K.: 1995, *CODA Manual version 0.30*, MRC Biostatistics Unit, Cambridge, UK.

Billingsley, P.: 1995, *Probability and Measure*, 3rd edn, Wiley, New York.

Billingsley, P.: 1999, *Convergence of Probability Measures*, 2 edn, Wiley, New York.

Boag, J. W.: 1949, Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society* **B11**, 15–44.

Brauer, F. and Nohel, J. A.: 1967, *Ordinary Differential Equations: A first course*, W.A. Benjamin Inc., New York, pp. 332–5.

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

Byar, D. P. and Corle, D. K.: 1977, Selecting optimal treatment in clinical trials using covariate information, *Chronic Diseases* **30**, 445–69.

Byar, D. P. and Green, S. B.: 1980, The choice of treatment for cancer patients based on covariate information: application to prostate cancer, *Bull. Cancer* **67**, 477–88.

Chambers, J. M. and Hastie, T.: 1992, *Statistical Models in S*, Chapman and Hall, London.

Chung, C. F.: 1987, Wiener pack: a subroutine package for computing probabilities associated with Wiener and Brownian bridge processes, *Technical Report 87-12*, Geological Survey of Canada.

Cox, D. R.: 1972, Regression models and lifetables (with discussion), *Journal of the Royal Statistics Society-B* **34**, 187–220.

Cox, D. R.: 1975, Partial likelihood, *Biometrika* **62**, 269–76.

Cox, D. R. and Oakes, D.: 1984, *Analysis of survival data*, Vol. 21 of *Monographs on statistics and applied probability*, Chapman and Hall, London.

Crowder, M.: 2001, *Classical Competing Risks*, Chapman & Hall / CRC, Boca Raton.

De Finetti, B.: 1937/1964, La prévision: ses lois logiques, ses sources subjectives, *Annals d'Institut Henri Poincaré* **7**, 1–68. Reprinted as (De Finetti 1980).

De Finetti, B.: 1980, Foresight: its logical laws, its subjective sources, *in* H. E. Kyburg and H. E. Smokler (eds), *Studies in Subjective Probability*, Dover, New York, pp. 93–158.

de Voogt, H. J., Studer, U., Schröder, F. H., Klijn, J. G., de Pauw, M., Sylvester, R. et al.: 1998, Maximum androgen blockade using LHRH agonist buserelin in combination with short-term (two weeks) or long-term (continuous) cyperoterone acetate is not superior to standard androgen deprivation in the treatment of advanced prostate cancer, *European Urology* **33**, 152–8.

Ferguson, T. S.: 1974, Prior distributions on a space of probability measures, *Annals of Statistics* **2**, 615–29.

Fleming, T. R. and Harrington, D. P.: 1991, *Counting Processes and Survival Analysis*, Wiley.

Genest, C. and MacKay, J.: 1986, Copules archimédiennes et familles de lois bi-dimensionnelles dont les marges sont données, *Canadian Journal of Statistics* **14**, 145–159.

Gill, R. D.: 1980, Censoring and stochastic integrals, *Mathematical Centre Tracts 124*, Mathematisch Centrum, Amsterdam.

Gill, R. D.: 1989, Non- and semi-paremetric maximum likelihood estimators and the von-Mises method (part 1), *Scandinavian Journal of Statistics* **16**, 92–128.

Gill, R. D. and Johansen, S.: 1990, A survey of product-integration with a view towards application in survival analysis, *Annals of Statistics* **18**, 1501–1555.

Gray, R. J.: 1988, A class of k-sample test for comparing the cumulative incidence of a competing risk, *The Annals of Statistics* **16**, 1141–54.

Hall, W. Y. and Wellner, J. A.: 1980, Confidence bands for a survival curve from censored data, *Biometrika* **67**, 133–42.

Heckman, J. J. and Honoré, B. E.: 1989, The identifiability of the competing risks model, *Biometrika* **76**, 325–330.

Hoel, D. G.: 1972, A representation of mortality data by competing risks, *Biometrics* **28**, 475–488.

Ihaka, R. and Gentleman, R.: 1996, R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**(3), 299–314.

Jacobsen, M.: 1982, *Statistical Analysis of Counting Processes*, Vol. 12 of *Lecture Notes in Statistics*, Springer-Verlag, New York.

Kalbfleisch, J. D. and Prentice, R. L.: 2002, *The statistical analysis of failure time data*, Wiley series in probability and statistics, 2 edn, Wiley, New York.

Kaplan, E. L. and Meier, P.: 1958, Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481, 562–563.

Kernighan, B. W. and Ritchie, D. M.: 1978, *The C programming language*, Prentice-Hall, London.

Klein, J. P.: 1991, Small sample moments of some estimators of the variance of the Kaplan-Meier and Nelson-Aalen estimators, *Scandinavian Journal of Statistics* **18**, 333–40.

Lavine, M.: 1992, Some aspects of Pólya tree distributions for statistical modeling, *Annals of Statistics* **20**, 1222–35.

Lavine, M.: 1994, More aspects of Pólya tree distributions for statistical modeling, *Annals of Statistics* **22**, 1161–76.

Lenglart, E.: 1977, Relation de domination entre deux procesus, *Annals d'Institut Henri Poincaré* **13**, 171–179.

Lin, D. Y.: 1997, Non-parametric inference for cumulative incidence functions in competing risks studies, *Statistics in Medicine* **16**, 901–910.

Lunn, M.: 1998, Applying k-sample tests to conditional probabilities for competing risks in a clinical trial, *Biometrics* **54**, 1662–72.

Lunn, M. and McNeil, D.: 1995, Applying Cox regression to competing risks, *Biometrics* **51**, 524–32.

Lutz, M. and Ascher, D.: 1999, *Learning Python*, O'Reilly & Associates Inc., Sebastopol, USA.

Mauldin, R. D., Sudderth, W. D. and Williams, S. C.: 1992, Pólya trees and random distributions, *Annals of Statistics* **20**, 1203–21.

McCulloch, C. E. and Searle, S. R.: 2001, *Generalized, linear, and mixed models*, Wiley, New York.

Miller, R. and Siegmund, D.: 1982, Maximally selected chi square statistics, *Biometrics* **38**, 1011–1016.

Nair, V. N.: 1984, Confidence bands for survival function with censored data: A comparative study, *Technometrics* **26**, 265–75.

Nelsen, R. B.: 1998, *An introduction to copulas*, Vol. 139 of *Lecture notes in statistics*, Springer, New York.

Nelson, W.: 1969, Hazard plotting for incomplete failure data, *Journal of Qualitative Technology* **1**, 27–52.

Pepe, M. S. and Mori, M.: 1993, Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data, *Statistics in Medicine* **12**, 737–51.

Peterson, A. V.: 1976, Bounds for a joint distribution with fixed sub-distribution functions: Application to competing risks, *Proceedings of the National Academy of Science, USA* **73**, 11–13.

Prentice, R. L. and Kalbfleisch, J. D.: 1978, The analysis of failure times in the presence of competing risks, *Biometrics* **34**, 541–54.

Raftery, A. E. and Lewis, S. M.: 1996, Implementing MCMC, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in practice*, Chapman and Hall, pp. 115–30.

Rebolledo, R.: 1980, Central limit theorems for local martingales, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **51**, 269–286.

Revuz, D. and Yor, M.: 1999, *Continuous Martingales and Brownian Motion*, Vol. 293 of *A series of comprehensive studies in mathematics*, 3 edn, Springer-Verlag, New York.

Schall, R.: 1991, Estimation in generalized linear models with random effects, *Biometrika* **78**, 719–27.

Schervish, M. J.: 1995, *Theory of statistics*, Springer, New York.

Spiegelhalter, D. J., Myles, J. P., Jones, D. R. and Abrams, K. R.: 2000, Bayesian methods in health technology assessment: a review, *Health Technology Assessment* **4**(38), 11–14.

Sylvester, R. J., Denis, L., de Voogt, H. et al.: 1998, The importance of prognostic factors in the interprestation of two EORTC metastatic prostate cancer trials, *European Urology* **33**, 134–43.

Therneau, T. M. and Grambsch, P. M.: 2000, *Modeling survival data: extending the Cox model*, Springer-Verlag, New York.

Tsiatis, A. A.: 1975, A nonidentifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Science, USA* **72**, 20–22.

Venables, W. N. and Ripley, B. D.: 2000, *S Progamming*, Springer, New York.

Walker, S. G. and Mallick, B. K.: 1997, Hierarchical generalised linear models and frailty models with Bayesian nonparametric mixing, *Journal of the Royal Statistical Society, B* **59**, 845–60.

Wilkinson, G. N. and Rogers, C. E.: 1973, Symbolic description of factorial models for analysis of variance, *Applied Statistics* **22**, 392–99.

Wolfinger, R.: 1993, Laplace's approximation for nonlinear mixed models, *Biometrika* **80**, 791–5.