WARWICK
THE UNIVERSITY OF WARWICK

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:
http://wrap.warwick.ac.uk/88814

Copyright and reuse:
This thesis is made available online and is protected by original copyright.
Please scroll down to view the document itself.
Please refer to the repository record for this item for information to help you to cite it.
Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications

DYNAMIC ANALYSIS

OF

SURVIVAL MODELS

AND

RELATED PROCESSES

Dani Gamerman

Thesis submitted for the degree of Ph.D. in Statistics

DEPARTMENT OF STATISTICS

UNIVERSITY OF WARWICK

MAY 1987

To Life

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to Mike West. Apart from being a highly competent supervisor, he was always available for discussions and often had a good idea bubbling out of his mind. He must be credited for devising the original idea for this thesis and is directly responsible for many developments thereafter.

To Bluma, for being such a loving and caring person. She was always there in the bad and the good times making everything that much nicer. As if that was not enough, she also helped by typing most of the manuscript of this thesis.

It is also important to acknowledge the members and students of the Department of Statistics for providing a good environment for the development of research, in particular Andy Pole and Ewart Shaw for the provision of most of the graphical facilities used in the microcomputer software described in the last chapter, and CAPES for the financial support. I would also like to thank Lita for the secretarial assistance and Rob McMahon for the numerous useful tips on *xroff*.

Finally, many thanks to all my friends and relatives for showing goodwill towards the success of my work even with little or no knowledge of what was it about.

Table of Contents

Summary vi

1. Introduction 1

1.1 Outline of the thesis 3

1.2 Notation and terminology 5

1.3 Basic results and definitions 7

2. Dynamic models 10

2.1 Dynamic linear models 11

2.2 Dynamic generalised linear and non-linear models 16

2.3 Dynamic models in continuous time 27

3. Analysis for survival data 34

3.1 Introduction 34

3.2 Temporal factorisation of the likelihood 38

3.3 Piecewise exponential distributions 43

3.4 Dynamic survival models and parametric updating 45

3.5 Model criticism 49

3.6 Prediction of future occurrences 52

4. Numerical applications 62

4.1 The two-sample problem 62

4.2 Time-dependent covariates 77

4.3 An econometric application 82

Appendix 4.1 Dynamic linear growth model 94

5. Special cases 98

5.1 Analysis of random samples 98

5.2 Life tables 111

5.3 Competing risks 120

6. Analysis of point processes 128

6.1 Introduction 129

6.2 The no-covariates case: discrete and continuous estimation 133

6.3 Marked point processes 138

6.4 Special reliability models 141

7. Generalisations and further topics 150

7.1 Multivariate point processes and multivariate failure data 150

7.2 Survival analysis in continuous time 153

7.3 Model assumptions and inference procedures 155

7.4 Robustness and analysis of residuals 157

7.5 Prior assesssments 158

7.6 Simulation 159

8. Microcomputer software for the analysis 161

8.1 SURVIVAL functions 161

8.2 SURVIVAL in practice 169

SUMMARY

This thesis presents new methods of analysis of survival data based on a Dynamic Bayesian approach. The models allow the parameters to change with time. The analysis is tractable and emphasises predictive aspects of the models. The survival problems covered include linear and non-linear regression, analysis of random samples, time-dependent covariates, life tables and competing risks. The analysis is also extended to a number of point processes. Numerical applications are provided and the microcomputer software to perform them is described.

CHAPTER ONE

INTRODUCTION

This thesis is concerned with the analysis and application of survival models. Survival models are mainly used in the study of survival times of patients subject to some treatment or combination of treatments. The aims are to understand the mechanisms governing people's ability to survive and make predictions for future patients. Although the terminology derives from medicine, the models are applicable to a number of different areas, such as reliability, demography and economics. Such areas of application are also explored in this thesis. The medical related terminology is retained however. One reason for this is that dynamic models are badly needed in medical studies where research seems to be moving towards time varying models faster than in other areas. It is also important to keep the same name for the same mathematical quantity in order to fix ideas and for clarity. It must be stressed, however that the methods on this thesis apply equally well to the analysis of failure time data, duration data or life tables.

Survival data often comes from a medical experiment to study a certain disease where a number of patients are subjected to a number of conditions (some of which are controlled by design). These patients are observed until one of the following three events occurs:

(a) the person dies from the disease studied by the experiment;

(b) the person dies from another cause;

(c) the person is withdrawn from the experiment.

Only in the first case is the survival time fully observed. As far as the experiment is concerned, the last two cases provide the same incomplete information: the patient survived the disease up to a certain point in time. There are a number of reasons for an event of type (c). The patient could emigrate, the experiment could have a limited duration after which the patient is still alive, and others. This incomplete information is a very common

feature of survival data, referred to as censoring. In many survival models and in this thesis, it is assumed that censoring occurs randomly. This is a strong, important but reasonable assumption. It implies that the time the person emigrates has no relation to the survival time or that the moment he/she dies of a car crash or heart attack provides no information as to his/her state on a cancer treatment.

The statistical problems considered in this thesis concern regression models. An experiment is carried out and, along with the (censored and uncensored) survival times, a number of additional pieces of information are recorded for each patient. These are usually patient attributes such as sex, age, and so forth or some variables controlled by the experiment such as type of treatment, dosage of drugs, etc. The aim of the study is to make inference about the effects of these factors and then predict the survival time for future patients given their attributes and the conditions under which they undergo the experiment.

The models of this thesis provide a dynamic approach to the study of survival data. Most models currently used and implemented in computer packages are based on static models. These assume that the effects regression variables have on the survival time are constant in time. Although this is reasonable in some applications, it is too restrictive in general and, in certain cases, it has been observed that factors have a distinctly dynamic effect on the outcome of the experiment. This is particularly evident in some of the examples provided in this thesis.

It will be shown that the adoption of dynamic models naturally leads to the consideration of survival data from a different perspective. This is a temporal perspective used also in time series where information is always associated with time. This approach is particularly appropriate in the study of survival models where the data is always a temporal measurement and should be treated accordingly.

The statistical framework under which the derivations are made is Bayesian throughout the thesis. This not only represents my philosophical position but seems to be

the only sensible way to carry out a dynamic inference procedure. In particular, linear Bayesian methods are used in connection with conjugate prior for canonical parameters. This procedure is implemented in West, Harrison & Migon (1985) providing a flexible yet workable method of inference avoiding the repeated use of high-dimensional numerical integrations. Other methods however are also discussed. One particular feature of the dynamic Bayesian approach is the routine handling of tied and censored observations. This is in contrast to the unnatural requirements for extra computations and modifications in other, standard models.

In developing the methods, particular attention was given to the development of an applicable methodology without oversimplification. It is for this reason that much of the material of the thesis comes from applications to real data. Considerable effort was also put into the development of widely applicable computer software. It is hoped that this class will include non-statisticians with data in their hands, models in their minds and nothing to link those two.

1.1 - Outline of the thesis

In the rest of this introductory chapter, a brief description of the material of the following chapters is given. A number of basic results from probability and statistics needed in the thesis are also provided.

The second chapter presents the dynamic models already available that are relevant for this thesis. It starts with the class of DLM (dynamic linear models) proposed by Harrison & Stevens (1976) to make forecasts of normal time series. The dynamic GLM (generalised linear models) are then presented as a generalisation of the DLM for non-normal time series. The basic linear models of this thesis are later shown to be an extension of those models. The analysis for non-linear models and for models in continuous time are also given as they are needed for a precise construction of the models for survival data.

Chapter 3 introduces the dynamic models for survival data and presents the related inference procedures. In associating a temporal factorisation of the likelihood with piecewise exponential distributions, the models are almost completely constructed. It only leaves a system equation specifying the evolution of the parameters to be set. After doing that, the models can be put into the inferential structure of the dynamic GLM to derive the required components of the analysis. Particularly useful are the on-line (given present information) and smoothed (given total information) distributions of the parameters. Also important are the predictive likelihood to assess the fit of the model entertained and predictive distributions for future individuals.

Chapter 4 provides three applications of the models to real data. The first one shows how useful dynamic models can be in detecting change in the effect of the covariates. The second application shows how to adapt the analysis to time-dependent covariates whilst presenting a comparison with classical (static) models. Finally, an application from economics is given with discussion of the discount approach to evolution modelling and switching models when the set of relevant covariates changes with time.

Chapter 5 deals with less orthodox uses of the regression models. Initially, the analysis of censored random samples is given. Some comparisons to the classical Kaplan & Meier (1958) estimator and data analyses are also presented. Then, the analysis of life tables with systematic exploration of different forms for the baseline hazard is discussed. Finally, the analysis is extended to competing risk models where the cause of death is relevant.

Chapter 6 presents the extension of the models to the analysis to point process observations. The analysis is presented first for discrete observation points and then for continuous observation points. Marked point processes are also analysed, and analyses for some particular point process models detailed. These models have all been proposed to the study of software reliability and are shown to lead to tractable analysis with the dynamic models proposed in this thesis.

Chapter 7 gives a brief account of some of the generalisations envisaged and some pointers to further research. It starts with the analysis of multivariate point processes and multivariate failure data. Then the analysis of survival data is extended to continuous observation points as previously done with point processes. A number of points like specific model assumptions, robustness, forms of evolution, prior elicitation, analysis of residuals and simulation are discussed.

The final chapter presents the development of a microcomputer package of the regression models. The programmes are written in APL and have been structured in a user-friendly way so that a very basic knowledge of the models is sufficient to enable the use of the package. Particular attention was given to data management and graphics output as well as the statistical analysis.

1.2 - Notation and Terminology

The material of this thesis comes mainly from three different areas: survival data analysis, Bayesian forecasting and control theory. It would be too much to expect all those areas to have the same notation for the same quantities; there is even notational disagreement within each of these areas. The strategy adopted was to use the above order of the areas to provide the choice of notation in doubtful cases. The main references setting the notation in the respective areas are:

(a) the 1972 paper by Cox in the Journal of Royal Statistic Society, Series B, volume 34, pages 187-220 on regression models for survival data;

(b) the 1985 paper by West, Harrison & Migon in the Journal of the American Statistical Association, volume 80, pages 73-97 on Bayesian forecasting for dynamic GLM;

(c) Maybeck's (1979) 'Stochastic models, estimation and control' book on basic control theory.

The only exception on the above rule is the adoption of Y to denote failure times instead of T. This is because the letter T is left to represent the observation times in this time series approach. Some commonly used quantities, to be explicitly defined in context later

are the following: $\mathbf{z}$ denotes a row vector of covariates, β denotes the a column vector of parameters modelling the effect of the covariates and λ denotes hazard function. In fact, all vectors are column vectors with the exception of $\mathbf{z}$. Prior mean and variances for β are denoted by $\mathbf{a}$ and $\mathbf{R}$ and posterior mean and variances by $\mathbf{m}$ and $\mathbf{C}$ respectively. The system evolution matrix is denoted by $\mathbf{G}$ and the system error and its variance by $\mathbf{w}$ and $\mathbf{W}$. Finally, the remaining notation choices come from control theory.

Whenever necessary, random variables are denoted by capital letters and their observed values by the corresponding lower case letters. Distributions can always be defined by their probability density functions f over the Lebesgue measure, and are assumed to be at least left continuous. Distribution functions are denoted by F. Two very important quantities are the survival function S, given by

$$S(y) = Pr(Y > y) = 1 - F(y) \tag{1.1}$$

and the hazard function λ given by

$$\lambda(y) = \lim_{\Delta \to 0^+} Pr(y - \Delta < Y \leq y \mid Y > y - \Delta) \tag{1.2}$$

with the slightly unusual definition accounting for the left-continuity.

The notation $X \sim F$ stands for 'the random variable X has distribution F'. So that $X \sim N[\mu, \sigma^2]$ means that X has normal distribution with density

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} \cdot (x-\mu)^2 \right\} \quad , x \in \mathbf{R}$$

and $X \sim G(\alpha, \gamma)$ means that X has Gamma distribution with density

$$f(x) = \frac{\gamma^{\alpha}}{\Gamma(\alpha)} \cdot x^{\alpha-1} \exp(-\gamma x) \quad , x > 0$$

The notation $\mathbf{X} \sim [\mathbf{m}, \mathbf{C}]$ is also used to denote that $E[\mathbf{X}] = \mathbf{m}$ and $V[\mathbf{X}] = \mathbf{C}$.

In order to distinguish between the various time instances used in this thesis, the following distinction is made for a function g of time: if g assumes values at times $t = 1, 2, \ldots$, then a general value of g is denoted by g_t; if it assumes values at times $t_1, t_2, \ldots$ then it is

denoted by g_i; and if it assumes values at times in $[0,\infty]$ then it is denoted by $g(t)$.

The following notation is also used:

- D denotes data or given information;
- $\biguplus$ denotes a union of exclusive sets;
- $\prod_{i\in I} a_i = 1$ and $\sum_{i\in I} b_i = 0$ for sequences of real numbers $\{a_i\}$ and $\{b_i\}$ if $I = \emptyset$;
- $\prod_{i\in I} A_i = A_{i_n} \cdots A_{i_2} A_{i_1}$, for A_j $p \times p$ matrices, $j = 1,\ldots,n$ and $I = \{i_1, \ldots, i_n\}$;
- $[x]$ denotes the integer part of x;
- $\Delta \to 0^+$ denotes $\Delta \to 0$ and $\Delta > 0$;
- $o(s)$ denotes a function satisfying $\lim_{s\to 0} \frac{o(s)}{s} = 0$

1.3 - Basic results and definitions

The most important ones relate λ, S and f. From (1.1), it is clear that

$$f(y) = -\dot{S}(y) = -\frac{dS(y)}{dy} \tag{1.3}$$

From (1.2),

$$f(y) = \lambda(y)\, S(y) \tag{1.4}$$

In general, the variables studied are time variables so that $\lambda(y) = 0$, $y < 0$. Hence replacing (1.4) in (1.3) and solving the equation for S, gives

$$S(y) = \exp\left[\int_0^y -\lambda(u)du\right] \tag{1.5}$$

As anticipated in the beginning of this chapter, survival variables are typically censored variables. This implies that their contribution to the likelihood is either through f (in case they are uncensored) or through S if censored. To avoid this essentially unnecessary complication, one can use the following

Definition 1.1. Let $\mathbf{Y} = (Y_1, \ldots, Y_m)$ have joint density $f(\cdot)$. Define

$$p(y_1, \ldots, y_m) = \lim_{\rho \to 0} \frac{Pr(y_i \in \Delta Y_i, i \in F;\ Y_j > y_j, j \in C)}{\prod_{i \in F} \Delta_i}$$

with $F =$ set of labels of individuals dying

$C =$ set of labels of individuals being censored

$\Delta Y_i =$ suitably small interval of length Δ_i containing y_i

$\rho = \max_{1 \le i \le m} \Delta_i$

With the above definition, every contribution to the likelihood is through p rather than f and S. The distinction between censored and uncensored contributions is made with the occurrence (or death) indicator χ. It assumes the value 1 for uncensored observations and 0, otherwise. So, if $m = 1$ in the definition, it reduces to

$$p(y) = \begin{cases} f(y) = \lambda(y) S(y) & , \chi = 1 \\ S(y) & , \chi = 0 \end{cases}$$

The occurrence indicator can be used to summarise the above equation via

$$p(y) = [\lambda(y)]^{\chi} S(y) \tag{1.6}$$

Also, if $F = \{1,...,m\}$ and $C = \varnothing$, then $p = f$ and if $C = \{1,...,m\}$ and $F = \varnothing$, then $p = S$.

An important result related to the function p is

Theorem 1.1. If $\mathbf{Y} = (Y_1, \ldots, Y_m)$ has density f then

$$p(y_1, \ldots, y_m) = \prod_{i=1}^{m} p(y_i \mid y_{i-1}, \ldots, y_1) \tag{1.7}$$

This is in fact a basic probability result but is quite useful in the sequel. It suffices to observe that:

(i) if $F = \{1,...,m\}$ and $C = \varnothing$ then (1.7) reduces to $f(y_1,...,y_m) = \prod_{i=1}^{m} f(y_i \mid y_{i-1},...,y_1)$;

(ii) if $C = \{1,...,m\}$ and $F = \varnothing$ then (1.7) reduces to $S(y_1,...,y_m) = \prod_{i=1}^{m} S(y_i \mid y_{i-1},...,y_1)$.

These results are elementary and Theorem 1.1 is only extending to the cases in between.

A final important result for the Bayesian analyses of this thesis states that if $\lambda \sim G(\alpha, \gamma)$ and the likelihood for λ is of the form

$$\lambda^{a} \exp(-b\lambda)$$

for some a, b then the posterior for λ is $G(\alpha + a, \gamma + b)$. The above likelihood appears from Exponential, Poisson and Gamma distributions for the observation and these are extensively used throughout this thesis.

References

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc., Ser. B,* **34**, 187-220.

Harrison, P. J. & Stevens, C. F. (1976). Bayesian forecasting (with discussion). *J. R. Statist. Soc., Ser. B,* **38**, 205-247.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Ass.,* **53**, 457-481.

Maybeck, P. S. (1979). *Stochastic Models, Estimation and Control, Vol. 1.* New York: Wiley.

West, M., Harrison, P. J. & Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Amer. Statist. Ass.,* **80**, 73-97.

CHAPTER TWO

DYNAMIC MODELS

In this chapter, the models upon which the dynamic approach is based are presented. They consist of a wide class of models that has been increasing substantially in the past two decades. Dynamic models as opposed to static ones are getting recognised as a more sensible way of modelling phenomena in different fields of science. It should be emphasised that the important feature of a dynamic model is not only the ability to accept changes in the underlying structure of a process but to ascertain the way in which those changes are taking place. By merely acknowledging changes, one is simply assuming a superposition of static models and the fundamental relation between past and present is lost. It is the specification of this relation that renders meaning to a model. To assume complete knowledge of those relations, however, is to pretend having complete control of the process which is rarely achieved. It seems more appropriate to specify them in stochastic terms under which the modeller attempts to describe a process as accurately as he possibly can, that is, based on all the information available to him. In this chapter, and indeed in the whole of this thesis, this is how a dynamic model is understood.

In the rest of this chapter a number of ways that dynamic models have been proposed are presented as a build-up process to the analysis developed in the forthcoming chapters. All have the basic structure of being defined at two different but complementary levels: the system and the observational levels. The first level provides the dynamic structure of the model by specifying the way the modeller believes the process under study to behave. It deals with the parameters of the model and therefore is not observable. It is the second, observational level that will provide the model with the information from observed data. This level is needed for in most models the knowledge of the process is limited and it provides an important ingredient of statistical inference, observed data. The relation between the two levels is established essentially via Bayes' theorem although

some other tools like limiting results for the continuous time case and linear Bayesian methods are also used.

The first section of this chapter presents the dynamic linear models (DLM) proposed by Harrison & Stevens (1976). These assume normality on both levels and represent a good starting point to the understanding of dynamic models and how they operate. Although normality is not assumed on either levels in subsequent chapters, the DLM possess many of the features of later models and a number of points can be discussed at this early stage. The second section deals with the extensions of the DLM developed recently at Warwick (Migon & Harrison, 1985; West, Harrison & Migon, 1985; West & Harrison, 1986). The dynamic analysis of generalised linear models (GLM) where the observations are no longer normally distributed and of non-linear models (NLM) is presented. Another important extension is towards the analysis in continuous time. This is presented in the final section first for discrete observations and continuous systems and then for continuous observations. The former is needed for a correct understanding of the analysis of unequally spaced observations and the latter is needed in the final chapters where data is supposed to arrive continuously in time. Although this material can be found in a number of control theory textbooks (Wong, 1971; McGarty, 1974; Maybeck, 1979) it has not yet been extensively studied within the areas of time series and Bayesian statistics. I hope to help in bridging this gap.

2.1 - Dynamic Linear Models

These models were formally proposed by Harrison & Stevens (1976) for the analysis of time series where the r-dimensional observation vector is obtained at equally spaced time intervals. It is completely characterised by

$$\textit{Observation equation} : \mathbf{y}_t = \mathbf{Z}_t \boldsymbol{\beta}_t + \mathbf{v}_t \quad , \mathbf{v}_t \sim N(\mathbf{0},\mathbf{V}_t) \tag{2.1}$$

$$\textit{System equation} \quad : \boldsymbol{\beta}_t = \mathbf{G}_t \boldsymbol{\beta}_{t-1} + \mathbf{w}_t \quad , \mathbf{w}_t \sim N(\mathbf{0},\mathbf{W}_t) \tag{2.2}$$

where $\mathbf{Z}_t$ is a $r \times p$ matrix of covariates at time t and $\boldsymbol{\beta}_t$ is the p-dimensional regression parameter modelling their effects. The observation equation has the form of a linear

regression model on $\mathbf{Z}_t$ but for the dependence on time of the regression parameter $\boldsymbol{\beta}_t$. The system equation tells how the system (represented by the p-dimensional parameter $\boldsymbol{\beta}$) is thought to behave. The matrices $\mathbf{G}_t$, t 1,2,.. describe the deterministic part of the system while the system disturbances $\mathbf{w}_t$, t 1,2,.... add extra uncertainty to the system dynamics as it moves one step ahead in time. Both the system disturbances and observation errors are assumed to be independent random variables. The $r \times r$ and $p \times p$ covariance matrices $\mathbf{V}_t$, $t = 1,2,....$ and $\mathbf{W}_t$, t 1,2,...., respectively, are all known. It is of primary interest to obtain, at each time t, the distribution of $\boldsymbol{\beta}_t$ conditional on all relevant information prior to t (D_{t-1}) in order to make predictions, and conditional on all information available at t ($D_t = \mathbf{y}_t , D_{t-1}$) to proceed in the sequential analysis. Assuming the initial condition $[\boldsymbol{\beta}_0 \mid D_0] \sim N(\mathbf{m}_0, \mathbf{C}_0)$, those distributions are shown to be normal for all t (Harrison & Stevens, 1976). It is interesting to go into some details of this inference process to settle notation and to gain some insight into the forthcoming generalisations. If $[\boldsymbol{\beta}_{t-1} \mid D_{t-1}] \sim N(\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$ then the system evolution (2.2) is used to set $[\boldsymbol{\beta}_t \mid D_{t-1}] \sim N(\mathbf{a}_t, \mathbf{R}_t)$ where

$$\mathbf{a}_t = \mathbf{G}_t \, \mathbf{m}_{t-1} \tag{2.3a}$$

$$\mathbf{R}_t = \mathbf{G}_t \, \mathbf{C}_{t-1} \, \mathbf{G}_t^T + \mathbf{W}_t \tag{2.3b}$$

By use of Bayes' theorem, the distribution for $\boldsymbol{\beta}_t$ is updated after observing $\mathbf{y}_t$ to $[\boldsymbol{\beta}_t \mid D_t] \sim N(\mathbf{m}_t, \mathbf{C}_t)$ where

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{S}_t (\mathbf{Q}_t + \mathbf{V}_t)^{-1} (\mathbf{y}_t - \mathbf{f}_t) \tag{2.4a}$$

$$\mathbf{C}_t = \mathbf{R}_t - \mathbf{S}_t (\mathbf{Q}_t + \mathbf{V}_t)^{-1} \mathbf{S}_t^T \tag{2.4b}$$

with

$$\mathbf{f}_t = \mathbf{Z}_t \, \mathbf{a}_t \tag{2.5a}$$

$$\mathbf{S}_t = \mathbf{R}_t \, \mathbf{Z}_t^T \tag{2.5b}$$

$$\mathbf{Q}_t = \mathbf{Z}_t \, \mathbf{R}_t \, \mathbf{Z}_t^T \tag{2.5c}$$

So, it is the system evolution equation (2.2) that sets the pattern of the underlying

process. The role of the observation level is to correct, or rather redirect (see 2.4), the direction the process takes. Equations (2.4) are often referred to as the (discrete time) Kalman filter.

In most time series applications, the on-line distributions of $[\beta_t \mid D_t]$ contains all the information that is needed. Based on this, one can predict future observations conditional on the total information at any one time, D_t. It so happens that, in some cases, one can afford (or even have) to wait for some time after t in order to obtain more information. The most informative distribution for β_t is no longer the one conditional on D_t but that conditional on D_N, for some integer $N>t$. This procedure of using not only information up to and including time t but also information beyond time t in the distribution of β_t is called smoothing. Among the various possible types of smoothing, the most important ones are the fixed-interval, fixed-point and fixed-lag smoothings (Maybeck, 1982). In the first one, N is fixed and t is allowed to vary from 1, 2, ... to N, in the second, t is fixed and N varies from t+1, t+2, ... and in the last one, k=N-t is fixed and t (N) varies from 1 (k+1), 2 (k+2), ... There are practical situations leading to all those three cases but I shall be concerned here only with the first one. It is used in the case when there is limited data running out at time N and one wants to get the distribution for the β's given the total information collected. So, from here on, smoothing stands for fixed-interval smoothing. West (1982) showed that the above distributions can be recursively obtained as

$$[\beta_t \mid D_N] \sim N(\mathbf{m}_t^N, \mathbf{C}_t^N) \quad , t = 1, 2, ..., N-1 \tag{2.6}$$

with

$$\mathbf{m}_t^N = \mathbf{m}_t + \mathbf{C}_t \mathbf{G}_{t+1}^T \mathbf{R}_{t+1}^{-1} (\mathbf{m}_{t+1}^N - \mathbf{a}_{t+1}) \tag{2.7a}$$

$$\mathbf{C}_t^N = \mathbf{C}_t - \mathbf{C}_t \mathbf{G}_{t+1}^T \mathbf{R}_{t+1}^{-1} (\mathbf{R}_{t+1} - \mathbf{C}_{t+1}^N) \mathbf{R}_{t+1}^{-1} \mathbf{G}_{t+1} \mathbf{C}_t \tag{2.7b}$$

The recursions are initialised at $t=N$, with $[\beta_N \mid D_N] \sim N(\mathbf{m}_N^N, \mathbf{C}_N^N)$ where $\mathbf{m}_N^N = \mathbf{m}_N$ and $\mathbf{C}_N^N = \mathbf{C}_N$. An insight into the proof of this recursive algorithm can be gained in the next section when the result is extended for the dynamic GLM and NLM.

The analysis for regression and other linear models is easily accommodated within the structure of DLM. If N independent observations are made, the above models fit into the (2.1)-(2.2) representation by taking $\beta_1 = \beta_2 = \cdots = \beta_N = \beta$, that is, $\mathbf{G}_1 = \mathbf{G}_2 = \cdots = \mathbf{G}_N = \mathbf{I}$ and $\mathbf{W}_1 = \mathbf{W}_2 = \cdots = \mathbf{W}_N = 0$. The order given to the observations is irrelevant due to the sequential nature of Bayes' theorem (Box & Tiao, 1973) and $[\beta | D_N] \sim [\beta_t | D_N] \sim [\beta_N | D_N] \sim N(\mathbf{m}_N, \mathbf{C}_N)$, for t= 1,...,N. This result is confirmed in (2.6) by observing that $\mathbf{m}_t^N = \mathbf{m}_N$ and $\mathbf{C}_t^N = \mathbf{C}_N$ in (2.7) since substitution of the above conditions in (2.7) lead to $\mathbf{m}_t^N = \mathbf{m}_{t+1}^N$ and $\mathbf{C}_t^N = \mathbf{C}_{t+1}^N$, t= 1, 2, ..., N-1. So, the static case can be analysed with the structure of dynamic models by specialisation of the system evolution equation. Although this result seems somewhat obvious, it is going to be useful later on in comparing more complex models with their static counterparts.

The analysis of unequally spaced observations can be dealt with by assuming for instance that after time t the next observation comes at time t+k, k= 2, 3, ... In this case, there is no correction for updating from t until t+k and repeated use of (2.2) for this time interval gives

$$\beta_{t+k} = \left[\prod_{i=1}^{k} \mathbf{G}_{t+i}\right]\beta_t + \sum_{i=1}^{k}\left[\prod_{j=1}^{k-i} \mathbf{G}_{t+i-j}\right]\mathbf{w}_{t+i} \qquad (2.8)$$

and the distribution for $[\beta_{t+k} | D_t]$ can be obtained. Standard analysis can be resumed and proceeded as before. It is not clear, however, what should be done were k not an integer. It is the need for this finer structure for the system that leads to the developments in continuous time of section 3.

The specification of the values $\mathbf{W}_t$, $t = 1, 2, \ldots$ is crucial for the practical implementation of the model. Apart from the fact that they are symmetric positive semidefinite (there is no restriction about correlation between the components of $\mathbf{w}_t$) matrices, they still require the specification of $p(p+1)/2$ quantities at each time t from the modeller. Recently, Ameen & Harrison (1985) suggested the use of a diagonal discount matrix $\mathbf{B}_t$ in order to set $\mathbf{R}_t$ as

$$\mathbf{B}_t^{-\frac{1}{2}} \mathbf{G}_t \mathbf{C}_{t-1} \mathbf{G}_t^T \mathbf{B}_t^{-\frac{1}{2}} \tag{2.9}$$

Each diagonal element of $\mathbf{B}_t$ is a number in (0,1] so that variances are increased according to the intuitively appealing notion of the amount of discount deemed necessary in the passage of information for the parameters. This approach has been modified to discounting by blocks of parameters by Ameen & Harrison (1985). It involves the specification of groups of parameters subject to common discounts applied directly to the their respective covariance matrix. This approach is now being routinely used in Bayesian forecasting (see Harrison & West, 1986) and is more fully described in chapter 7. The discount approach reduces drastically the number of quantities to be specified at each time to p which can be a bonus specially if p is large. Those advantages are partly lost when observations come frequently in unequally spaced intervals. In this case, one is interested in the specification of a discount matrix for the whole period between successive observations. The interval lengths are frequently different and when they assume a wide range of values, the comparison of the discounts used is by no means an easy task. To make this point clear, suppose the system is constant so that $\mathbf{G}_t = \mathbf{I}$ and $\mathbf{W}_t = \mathbf{W}$ for all t and that, as before, the first observation made after t is at t+k ,k= 2, 3, ... The use of the standard DLM leads, by direct substitution in (2.8), to

$$\boldsymbol{\beta}_{t+k} = \boldsymbol{\beta}_t + \sum_{i=1}^{k} \mathbf{w}_{t+i} \tag{2.10}$$

and due to the constancy assumption

$$V(\boldsymbol{\beta}_{t+k} \mid D_t) = V(\boldsymbol{\beta}_t \mid D_t) + k\,\mathbf{W} \tag{2.11}$$

On the other hand, repeated use of (2.9) with a constant discount **B** leads to

$$V(\boldsymbol{\beta}_{t+k} \mid D_t) = \mathbf{B}^{-k/2}\, V(\boldsymbol{\beta}_t \mid D_t)\, \mathbf{B}^{-k/2} \tag{2.12}$$

Comparison of (2.11) with (2.12) shows that standard DLM lead to a variance growing linearly with time interval but the discount variance grows exponentially with it. Again, as in the DLM, the discount approach does not state how to evolve if the time interval is not an integer. The difference is that there is no structure to date formally developed for discount models in continuous time as there is for the DLM. So, extra care must be

exercised when using discount ideas in the environment of different time intervals.

In standard control theory models, the system disturbances $\mathbf{w}_t, t = 1,2, \cdots$ are often replaced by $\Gamma_t \mathbf{u}_t, t = 1,2, \cdots$ where $\mathbf{u}_t, t = 1,2, \cdots$ are the q-dimensional system disturbances and $\Gamma_t, t = 1,2, \cdots$ are $p \times q$ known matrices (Jazwinski, 1970). If $\mathbf{u}_t$ is a vector of independent disturbances, this representation can be used to obtain the spectral decomposition

$$\mathbf{W}_t = \Gamma_t V(\mathbf{u}_t) \Gamma_t^T = \sum_{i=1}^{q} \sigma_{it} \sigma_{it}^T \tag{2.13}$$

for suitable values of $\sigma_{it}, i = 1, \ldots, p$ and of the $p \times p$ orthogonal matrix Γ_t (Rao, 1973, pg. 40). The theory of principal components analysis could be called in to select a number of vectors σ_{it} that best summarise the (essentially subjective) information in $\mathbf{W}_t$. From a practical point of view, the selection of the first principal component seems adequate in many applications. It implies that the system is affected by a single disturbance u_t, each parameter by a different amount determined by the corresponding value in Γ_t. This approach leaves the practitioner with the specification of only p quantities at each time. It has none of the problems mentioned before and can be used in continuous time modelling. $\mathbf{W}_t$ becomes a singular matrix but this is not a problem for the analysis. In fact, this is the system disturbance structure adopted in the numerical applications although any other specification of $\mathbf{W}_t, t = 1,2, \cdots$, or indeed of $\mathbf{B}_t, t = 1,2, \cdots$, that copes with unequal time intervals can be used as well.

2.2 - Dynamic Generalised Linear and Non-linear Models

The class of generalised linear models was introduced by Nelder & Wedderburn (1972) to analyse regression problems for observations coming from a distribution of the exponential family. The linearity was preserved by explicit assumption of the dependence of the distribution of each scalar observation y_t, on the form $\eta_t = \mathbf{z}_t \beta$, where $\mathbf{z}_t$ is a p-dimensional vector of covariates associated with y_t, t= 1, 2, ... and β is the regression parameter. If the distribution of y_t belongs to the exponential family, its density can be

written as

$$f(y_t \mid \lambda_t) = \exp\{\phi^{-1} y_t \lambda_t - a(\lambda_t)\} b(\phi, \lambda_t) \qquad (2.14)$$

where ϕ is known and λ_t is the canonical parameter (McCullagh & Nelder, 1983). The model can then be completed by establishing the link $g(\lambda_t)=\eta_t$ for all t. The dynamic model is defined by letting $\boldsymbol{\beta}$ depend on t through a stochastic structure similar to (2.2). The Bayesian approach of West, Harrison & Migon (1985) assumes that structure without normality. This leaves the process only partially specified through its means and variances but gives enough flexibility to allow for a conjugate prior distribution for λ_t. This readily provides updating equations for λ_t and a predictive distribution for y_{t+1}. λ_t is related to η_t, however, through the useful concept of guide relationship. The model can be summarised as follows:

(i) *observation equation*: equation (2.14)

(ii) *system equation* : $\boldsymbol{\beta}_t = \mathbf{G}_t \boldsymbol{\beta}_{t-1} + \mathbf{w}_t \quad , \mathbf{w}_t \sim (\mathbf{0}, \mathbf{W}_t)$

(discount ideas can be alternatively used to set $\mathbf{R}_t$ as in (2.9))

(iii) ***prior distributions*** : $(\boldsymbol{\beta}_t \mid D_{t-1}) \sim (\mathbf{a}_t, \mathbf{R}_t)$ with unspecified density and $\lambda_t \mid D_{t-1}$ has conjugate distribution with link $g(\lambda_t)=\eta_t$ implying $E[g(\lambda_t) \mid D_{t-1}] = f_t$, $Cov[\boldsymbol{\beta}_t, g(\lambda_t) \mid D_{t-1}] = \mathbf{s}_t$, $V[g(\lambda_t) \mid D_{t-1}] = q_t$ where $f_t = \mathbf{z}_t \mathbf{a}_t$, $\mathbf{s}_t = \mathbf{R}_t \mathbf{z}_t^T$ and $q_t = \mathbf{z}_t \mathbf{s}_t$.

The posterior distribution of $\lambda_t \mid D_t$ stays within the same conjugate family with parameters changed by Bayes' theorem with $E[g(\lambda_t) \mid D_t] = g_t$ and $V[g(\lambda_t) \mid D_t] = p_t$. The posterior distribution for $\boldsymbol{\beta}_t$, $(\boldsymbol{\beta}_t \mid D_t) \sim (\mathbf{m}_t, \mathbf{C}_t)$ with

$$\mathbf{m}_t = \mathbf{a}_t + \frac{\mathbf{s}_t}{q_t}(g_t - f_t) \qquad (2.15a)$$

$$\mathbf{C}_t = \mathbf{R}_t - \frac{\mathbf{s}_t \mathbf{s}_t^T}{q_t}\left(1 - \frac{p_t}{q_t}\right) \qquad (2.15b)$$

is obtained by use of linear Bayesian methods on the joint (partially specified) distribution

$$\left[\begin{pmatrix} \boldsymbol{\beta}_t \\ g(\lambda_t) \end{pmatrix} \Big| D_{t-1} \right] \sim \left[\begin{pmatrix} \mathbf{a}_t \\ f_t \end{pmatrix}, \begin{pmatrix} \mathbf{R}_t & \mathbf{s}_t \\ \mathbf{s}_t^T & q_t \end{pmatrix} \right]$$

The details from the calculations above can be found in Migon (1984) and West, Harrison & Migon (1985). It can be seen from (iii) that the guide relation is used to construct the conjugate prior for λ_t and also appears in the Kalman filter-like updating equations for β_t. In the normal case, taking $\eta_t = \lambda_t$ as the guide relationship leads to the same results as in the DLM, that is, (2.15) coincides with (2.4) (see West, Harrison & Migon, 1985). In non-normal situations, the link provides a guide to the choice of the prior distribution of λ_t, and the covariance term $\mathbf{s}_t$ relating λ_t to β_t. As with other Bayesian models, the choice of prior is essentially left to the modeller, subject only to the requirements that the mean and variance are approximately as implied by the guide. In general, numerical considerations may lead to slightly modified values for the prior moments. The choice of a conjugate analysis is ,as usual, useful computationally but is not necessary.

Example 1. *Exponential distribution - Logarithmic link*

This is in fact an important example because exponential-like distributions and likelihoods are to appear throughout this thesis. In this case,

$$f(y_t \mid \lambda_t) = \lambda_t \exp\{-\lambda_t y_t\}$$

and the conjugate prior is $G(\alpha_t, \gamma_t)$ for some α_t, γ_t. The prior for β_t is $[\beta_t \mid D_{t-1}] \sim [\mathbf{a}_t, \mathbf{R}_t]$ and the link for $\eta_t = \mathbf{z}_t \beta_t$ is $\log \lambda_t \simeq \eta_t$. From the properties of the Gamma distribution,

$$E[\log \lambda_t \mid D_{t-1}] = \psi(\alpha_t) - \log \gamma_t \tag{2.16a}$$

$$V[\log \lambda_t \mid D_{t-1}] = \psi(\alpha_t) \tag{2.16b}$$

where $\psi(x) = d \log \Gamma(x)/dx$ is usually called the digamma function and as supplementary descriptions of the distribution of $\log \lambda_t$, its mode is $\log(\alpha_t/\gamma_t)$ and its curvature at the mode is α_t^{-1}. Equality between the moments of $\log \lambda_t$ and η_t implies that α_t and γ_t are the solution of

$$\psi(\alpha_t) - \log \gamma_t = f_t = \mathbf{z}_t \mathbf{a}_t \tag{2.17a}$$

$$\psi(\alpha_t) = q_t = \mathbf{z}_t \mathbf{s}_t \tag{2.17b}$$

This means laborious evaluations, even for computers with built-in gamma and

digamma functions, for a relation that is only meant as a guide in constructing the prior. West, Harrison & Migon (1985) suggest the use of mode and curvature at mode (to replace mean and variance) as general measures of location and spread of the distribution of $\log \lambda_t$ giving

$$f_t = \log(\alpha_t / \gamma_t) \tag{2.18a}$$

$$q_t = \alpha_t^{-1} \tag{2.18b}$$

or the inverse relation

$$\alpha_t = q_t^{-1} \tag{2.19a}$$

$$\gamma_t = q_t^{-1} \exp(-f_t) \tag{2.19b}$$

Another justification for the relations (2.18)-(2.19) in terms of non-linear models is given in example 2.

After the prior $\lambda_t \mid D_{t-1} \sim G(\alpha_t, \gamma_t)$ is constructed with (2.19), it can be updated to $\lambda_t \mid D_t \sim G(\alpha_t + 1, \gamma_t + y_t)$ via Bayes' theorem with the likelihood for λ_t from y_t. The posterior moments of $\log \lambda_t$ are $g_t = \psi(\alpha_t + 1) - \log(\gamma_t + y_t)$ and $p_t = \psi(\alpha_t + 1)$. Use of the recursive relations

$$\psi(x+1) = \psi(x) + x^{-1} \quad , x > 0 \tag{2.20}$$

$$\psi(x+1) = \psi(x) - x^{-2} \quad , x > 0 \tag{2.21}$$

along with f_t and q_t as in (2.18) and the above values of g_t and p_t in (2.15) gives

$$\mathbf{m}_t = \mathbf{a}_t + \frac{\mathbf{s}_t}{q_t}\left[q_t + \log(q_t) + \psi(q_t^{-1}) - \log(1 + q_t y_t \exp(f_t))\right] \tag{2.22a}$$

$$\mathbf{C}_t = \mathbf{R}_t - \left[1 + \frac{1}{q_t} - \frac{1}{q_t^2}\psi\left(\frac{1}{q_t}\right)\right]\mathbf{s}_t \mathbf{s}_t^T \tag{2.22b}$$

The same considerations leading to (2.18) can be invoked to approximate

$$g_t = \log\left(\frac{\alpha_t - 1}{\gamma_t + y_t}\right) \tag{2.23a}$$

$$p_t = (\alpha_t + 1)^{-1} \tag{2.23b}$$

implying that $[\boldsymbol{\beta}_t \mid D_t] \sim \mathbf{m}_t, \mathbf{C}_t$ with

$$\mathbf{m}_t = \mathbf{a}_t + \frac{\mathbf{s}_t}{q_t} \log \left(\frac{1 - q_t}{1 - q_t\, y_t \exp(f_t)} \right) \tag{2.24a}$$

$$\mathbf{C}_t = \mathbf{R}_t - \frac{\mathbf{s}_t \mathbf{s}_t^T}{1 - q_t} \tag{2.24b}$$

If Y_t is a censored observation at y_t, it provides the likelihood $\exp(-\lambda_t y_t)$ which gives the posterior $[\lambda_t \mid D_t] \sim G(\alpha_t, \gamma_t + y_t)$. In this case, $g_t = \psi(\alpha_t) - \log(\gamma_t + y_t)$, $p_t = \psi(\alpha_t) - q_t$ and $[\boldsymbol{\beta}_t \mid D_t] \sim [\mathbf{m}_t, \mathbf{C}_t]$ with

$$\mathbf{m}_t = \mathbf{a}_t - \frac{\mathbf{s}_t}{q_t} \log \left[1 - q_t\, y_t \exp(f_t) \right] \tag{2.25a}$$

$$\mathbf{C}_t = \mathbf{R}_t \tag{2.25b}$$

In this case, the approximating values (2.23) to g_t and p_t lead to the same expressions for $\mathbf{m}_t$ and $\mathbf{C}_t$.

The prediction for a new observation Y_{t+1} with covariates $\mathbf{z}_{t+1}$ is accomplished by constructing $[\lambda_{t+1} \mid D_t] \sim G(\alpha_{t+1}, \gamma_{t+1})$ using $g(\lambda_{t+1}) \simeq \eta_{t+1}$ and $[\boldsymbol{\beta}_{t+1} \mid D_t] \sim [a_{t+1}, R_{t+1}]$. The predictive distribution for Y_{t+1} is

$$\begin{aligned} p(y_{t+1} \mid D_t) &= \int_0^\infty p(y_{t+1} \mid \lambda_{t+1}, D_t)\, f(\lambda_{t+1} \mid D_t)\, d\lambda_{t+1} \\ &= \int_0^\infty \lambda_{t+1}^{y_{t+1}} e^{-\lambda_{t+1} y_{t+1}} \frac{\gamma_{t+1}^{\alpha_{t+1}}}{\Gamma(\alpha_{t+1})} \lambda_{t+1}^{\alpha_{t+1}-1} e^{-\gamma_{t+1}\lambda_{t+1}}\, d\lambda_{t+1} \\ &= \left[\frac{\alpha_{t+1}}{\gamma_{t+1} + y_{t+1}} \right]^{y_{t+1}} \left[1 + \frac{y_{t+1}}{\gamma_{t+1}} \right]^{-\alpha_{t+1}}, \quad y_{t+1} > 0. \end{aligned} \tag{2.27}$$

So that, for instance, its density is

$$f(y_{t+1} \mid D_t) = \frac{\alpha_{t+1}}{\gamma_{t+1} + y_{t+1}} \left[1 + \frac{y_{t+1}}{\gamma_{t+1}} \right]^{-\alpha_{t+1}}, \quad y_{t+1} > 0. \tag{2.28}$$

The study of non-linear models in the context of Bayesian forecasting was intro-

duced by Migon (1984). It deals mainly with two types of non-linearity: non-linear system evolution when the system equation has the linear term $\mathbf{G}_t\boldsymbol{\beta}_{t-1}$ replaced by a general transfer function $\mathbf{G}_t(\boldsymbol{\beta}_{t-1})$ and when η_t cannot be written as $\mathbf{z}_t\boldsymbol{\beta}_t$ and is replaced by a non-linear relation $\mathbf{z}_t(\boldsymbol{\beta}_t)$. There are other types of non-linearity, e.g. when the disturbance term in the system equation has the form $\Pi(\boldsymbol{\beta}_t)\mathbf{w}_t$ depending on the state of the system at that time, but those are not considered here. These models can be analysed as the linear ones with the following changes:

(i') $[\boldsymbol{\beta}_t \mid D_t] \sim \mathbf{a}_t^*, \mathbf{R}_t^*$ with $\mathbf{a}_t^* = E[\mathbf{G}_t(\boldsymbol{\beta}_{t-1}) \mid D_{t-1}]$ and $\mathbf{R}_t^* = V[\mathbf{G}_t(\boldsymbol{\beta}_{t-1}) \mid D_{t-1}]$;

(ii') $E[g(\lambda_t) \mid D_{t-1}] \sim f_t^*$, $Cov[\boldsymbol{\beta}_t, g(\lambda_t) \mid D_{t-1}] \simeq \mathbf{s}_t^*$ and $V[g(\lambda_t) \mid D_{t-1}] \sim q_t^*$ with $f_t^* = E[\mathbf{z}_t(\boldsymbol{\beta}_t) \mid D_{t-1}]$, $\mathbf{s}_t^* = Cov[\boldsymbol{\beta}_t, \mathbf{z}_t(\boldsymbol{\beta}_t) \mid D_{t-1}]$ and $q_t^* = V[\mathbf{z}_t(\boldsymbol{\beta}_t) \mid D_{t-1}]$.

(iii') $[\boldsymbol{\beta}_t \mid D_t] \sim \mathbf{m}_t^*, \mathbf{C}_t^*$ with $\mathbf{m}_t^*$ and $\mathbf{C}_t^*$ as in (2.15) with starred elements.

Those are just about the only changes needed and the posterior distributions for λ_t and $\boldsymbol{\beta}_t$ are obtained as in the linear case, via Bayes' theorem and linear Bayesian methods respectively. The problem lies, of course, in the determination of $\mathbf{a}_t^*$, $\mathbf{R}_t^*$, f_t^*, $\mathbf{s}_t^*$ and q_t^*. The two main streams developed to tackle this problem are to assume further, as in the DLM, that the system process is normal (Migon & Harrison, 1985) or to replace the non-linear relations by first order (linear) and second order approximations for $\boldsymbol{\beta}$ around its mean (West & Harrison ,1986). This second approach is more general and is generally presented in the engineering context (Jazwinski, 1970; Maybeck, 1982) while the additional assumption of normality seems strong and unjustified. Linear approximation gives

$$\mathbf{a}_t^* = \mathbf{G}_t(\mathbf{m}_{t-1}^*) \tag{2.29a}$$

$$\mathbf{R}_t^* = \mathbf{H}_t \mathbf{C}_{t-1}^* \mathbf{H}_t^T + \mathbf{W}_t \tag{2.29b}$$

$$f_t^* = \mathbf{z}_t(\mathbf{a}_t^*) \tag{2.29c}$$

$$\mathbf{s}_t^* = \mathbf{R}_t^* \mathbf{h}_t \tag{2.29d}$$

$$q_t^* = \mathbf{h}_t^T \mathbf{s}_t^* \tag{2.29e}$$

with

$$\mathbf{H}_t \quad \frac{d\,\mathbf{G}_t(\mathbf{m}_{t-1}^*)}{d\,\boldsymbol{\beta}_{t-1}} \tag{2.30a}$$

$$\mathbf{h}_t \quad \frac{d\,\mathbf{z}_t(\mathbf{a}_t^*)}{d\,\boldsymbol{\beta}_t} \tag{2.30b}$$

Second order approximations can be used to correct $\mathbf{a}_t^*$ and f_t^* further to

$$\mathbf{a}_t^* = \mathbf{G}_t(\mathbf{m}_{t-1}^*) + \mathbf{b}_y \tag{2.31a}$$

$$f_t^* = \mathbf{z}_t(\mathbf{a}_t^*) + b_z \tag{2.31b}$$

where $\mathbf{b}_y$ is the p-dimensional vector with k^{th} component given as

$$\frac{1}{2}\,tr\left[\frac{d^2\,\mathbf{g}_{kt}(\mathbf{m}_{t-1}^*)}{d\,\boldsymbol{\beta}_{t-1}\,d\,\boldsymbol{\beta}_{t-1}^T}\,\mathbf{C}_{t-1}^*\right] \quad \text{and} \quad b_z = \frac{1}{2}\,tr\left[\frac{d^2\,\mathbf{z}_t(\mathbf{a}_t^*)}{d\,\boldsymbol{\beta}_t\,d\,\boldsymbol{\beta}_t^T}\,\mathbf{R}_t\right]$$

where $\mathbf{g}_{kt}$ is the k^{th} component of the function $\mathbf{G}_t$ and $tr[\mathbf{X}]$ is the trace of the matrix $\mathbf{X}$. There are also some expressions for $\mathbf{R}_t^*$, $\mathbf{s}_t^*$ and q_t^* based on second order approximations (Jazwinski, chapter 9, 1970) but the values given in (2.29) are the only ones considered here.

Example 2. *Exponential distribution - Logarithmic link revisited*

One can analyse this model via the non-linear link $\lambda_t = \exp(\mathbf{z}_t\boldsymbol{\beta}_t)$. The system equation is still linear so that $\mathbf{a}_t^* = \mathbf{a}_t$ and $\mathbf{R}_t^* = \mathbf{R}_t$. The above link gives $f_t^* = \exp(f_t)$, $\mathbf{s}_t^* = f_t^*\mathbf{s}_t$ and $q_t^* = (f_t^*)^2 q_t$ where f_t, $\mathbf{s}_t$ and q_t are as obtained in example 1. Since the function g is now the identity, one can take $[\lambda_t \mid D_{t-1}] \sim G(\alpha_t, \gamma_t)$ with $E[\lambda_t \mid D_{t-1}] = f_t^*$ and $V[\lambda_t \mid D_{t-1}] = q_t^*$. This gives

$$\frac{\alpha_t}{\gamma_t} = \exp(f_t) \quad \text{and} \quad \frac{\alpha_t}{\gamma_t^2} = q_t \exp(2f_t)$$

implying that (α_t, γ_t) are related to (f_t, q_t) as in (2.18)-(2.19). The above relations give $[\boldsymbol{\beta}_t \mid D_t] \sim [\mathbf{m}_t^*, \mathbf{C}_t^*]$ with

$$\mathbf{m}_t^* = \mathbf{a}_t + \frac{\mathbf{s}_t}{q_t}\left[\frac{1+q_t}{1+q_t\,y_t\exp(f_t)} - 1\right] \tag{2.32a}$$

$$\mathbf{C}_t^* = \mathbf{R}_t - \frac{\mathbf{s}_t\,\mathbf{s}_t^T}{q_t}\left[1 - \frac{1+q_t}{(1+q_t\,y_t\exp(f_t))^2}\right] \tag{2.32b}$$

It is interesting to observe that the correction introduced in the mean of β_t once y_t is observed (given by $\mathbf{m}_t^* - \mathbf{a}_t$) is a linear approximation of the one obtained in example 1 after using approximations (2.23) (see equation (2.24a)). The second order approximation (2.31b) gives $f_t^* = \exp(f_t)[1 + (q_t/2)]$ and use of normality gives $f_t^* = \exp[f_t + (q_t/2)]$.

The inference for linear and non-linear models (of the types shown here) is, in fact, similar. The difference lies in the determination of moments for some parametric functions which is more difficult and is rarely achieved without some approximation or additional assumption. Static models can be analysed within the structure of dynamic GLM and NLM by explicitly assuming no evolution in the system equation, that is, $\mathbf{I} = \mathbf{G}_1 = \mathbf{G}_2 = \cdots$ and $0 = \mathbf{W}_1 = \mathbf{W}_2 = \cdots$ as in the DLM. An additional complication in the passage from DLM to dynamic GLM appears here. The parameters can still be updated by the ingenious use of linear Bayesian methods but their posterior distributions are no longer independent of the order the observations are analysed. West, Harrison & Migon (1985) provide an example in which it is demonstrated that the order given to the observations has little effect. Since the models in this thesis deal with many observations at a time, the static analysis can be used to process them one by one. This point, however, needs more attention and I will return to it in the next chapter.

The final point in this section is about smoothing. This procedure plays a vital role in the predictive ability of the models to be proposed in the next chapters. The following theorem summarises the result needed.

Theorem 2.1 Under the Dynamic GLM structure presented in this section, that is (i)-(iii), the smoothed means and variances for the parameters conditional on information D_N is given by

$$[\beta_t \mid D_N] \sim [\mathbf{m}_t^N, \mathbf{C}_t^N] \quad , t = 1, 2, \ldots, N \tag{2.33}$$

with $\mathbf{m}_t^N$ and $\mathbf{C}_t^N$ as in (2.7) and $\mathbf{m}_N^N = \mathbf{M}_N$ and $\mathbf{C}_N^N = \mathbf{C}_N$.

Proof: a) Let us prove first for t=N-1

$$f(\beta_N,\beta_{N-1}|D_N) \alpha f(y_N|\beta_N,\beta_{N-1},D_{N-1})\, f(\beta_N,\beta_{N-1}|D_{N-1}) \tag{2.34}$$

via Bayes' theorem applied to the prior for $(\beta_N,\beta_{N-1}|D_{N-1})$ and observation y_N. The distribution of y_N depends on λ_N which is directly related to β_N, so

$$f(y_N|\beta_N,\beta_{N-1},D_{N-1}) \quad f(y_N|\beta_N,D_{N-1}) \tag{2.35}$$

(In fact, given β_N, the distribution of y_N does not depend on D_{N-1} either in the context of GLM but D_{N-1} is left in the conditioning part with view to further generalisations.)

Also,

$$f(\beta_{t+1},\beta_t|D_t) = f(\beta_t|\beta_{t+1},D_t)\, f(\beta_{t+1}|D_t) \quad , t=1,2,....,N-1 \tag{2.36}$$

Expressions (2.35) and (2.36) for t=N-1 replaced in (2.34) give

$$\begin{aligned} f(\beta_N,\beta_{N-1}|D_N) &\alpha \; f(y_N|\beta_N,D_{N-1})\, f(\beta_N|D_{N-1})]\, f(\beta_{N-1}|\beta_N,D_{N-1}) \\ &\alpha\; f(\beta_N|D_{N-1},y_N)\, f(\beta_{N-1}|\beta_N,D_{N-1}) \end{aligned} \tag{2.37}$$

by use of Bayes' theorem to the densities in the brackets. The expression in the right hand side (RHS) is a product of densities that integrates to 1 with respect to (β_N,β_{N-1}). (2.37) is therefore an equality and after rejoining $D_N=\{D_{N-1},y_N\}$ can be rewritten as

$$f(\beta_N,\beta_{N-1}|D_N) = f(\beta_{N-1}|\beta_N,D_{N-1})\, f(\beta_N|D_N) \tag{2.38}$$

This implies that

$$E[\beta_{N-1}|D_N] = E[E[\beta_{N-1}|\beta_N,D_{N-1}]|D_N] \tag{2.39a}$$

$$V[\beta_{N-1}|D_N] = E[V[\beta_{N-1}|\beta_N,D_{N-1}]|D_N] + V[E[\beta_{N-1}|\beta_N,D_{N-1}]|D_N] \tag{2.39b}$$

where the inner expectations in the RHS of (2.39) are with respect to the distribution of $[\beta_{N-1}|\beta_N,D_{N-1}]$ and the outer ones with respect to the distribution of $[\beta_N|D_N]$. The moments of the joint distribution of $[\beta_{N-1},\beta_N|D_{N-1}]$ are given by (ii) as

$$\left[\begin{pmatrix}\beta_{N-1}\\ \beta_N\end{pmatrix}\Big| D_{N-1}\right] \sim \left[\begin{pmatrix}\mathbf{m}_{N-1}\\ \mathbf{a}_N\end{pmatrix}, \begin{pmatrix}\mathbf{C}_{N-1} & \mathbf{C}_{N-1}\mathbf{G}_N^T\\ \mathbf{G}_N\mathbf{C}_{N-1} & \mathbf{R}_N\end{pmatrix}\right] \tag{2.40}$$

and use of linear Bayesian methods implies

$$E[\boldsymbol{\beta}_{N-1}|\boldsymbol{\beta}_N, D_{N-1}] = \mathbf{m}_{N-1} + \mathbf{C}_{N-1}\mathbf{G}_N^T\mathbf{R}_N^{-1}(\boldsymbol{\beta}_N - \mathbf{a}_N) \tag{2.41a}$$

$$V[\boldsymbol{\beta}_{N-1}|\boldsymbol{\beta}_N, D_{N-1}] = \mathbf{C}_{N-1} - \mathbf{C}_{N-1}\mathbf{G}_N^T\mathbf{R}_N^{-1}\mathbf{G}_N\mathbf{C}_{N-1} \tag{2.41b}$$

Using the fact that $E[\boldsymbol{\beta}_N|D_N] = \mathbf{m}_N = \mathbf{m}_N^N$ and $V[\boldsymbol{\beta}_N|D_N] = \mathbf{C}_N = \mathbf{C}_N^N$, insertion of (2.41) in (2.39) gives

$$\mathbf{m}_{N-1}^N = E[\boldsymbol{\beta}_{N-1}|D_N] = \mathbf{m}_{N-1} - \mathbf{C}_{N-1}\mathbf{G}_N^T\mathbf{R}_N^{-1}(\mathbf{m}_N^N - \mathbf{a}_N) \tag{2.42a}$$

$$\mathbf{C}_{N-1}^N = V[\boldsymbol{\beta}_{N-1}|D_N] = \mathbf{C}_{N-1} - \mathbf{C}_{N-1}\mathbf{G}_N^T\mathbf{R}_N^{-1}(\mathbf{R}_N - \mathbf{C}_N^N)\mathbf{R}_N^{-1}\mathbf{G}_N\mathbf{C}_{N-1} \tag{2.42b}$$

b) Suppose now that t= 1, 2, ... N-2

Using the ideas in (a), one can write

$$f(\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t|D_N) \propto f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N|\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t, D_t)\, f(\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t|D_t) \tag{2.43}$$

The same arguments leading to (2.35) can be used to get

$$f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N|\boldsymbol{\beta}_N, \ldots, \boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t, D_t) = f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N|\boldsymbol{\beta}_N, \ldots, \boldsymbol{\beta}_{t+1}, D_t) \tag{2.44}$$

From the system evolution equation,

$$f(\boldsymbol{\beta}_N, \ldots, \boldsymbol{\beta}_{t+2}|\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t, D_t) = f(\boldsymbol{\beta}_N, \ldots, \boldsymbol{\beta}_{t+2}|\boldsymbol{\beta}_{t+1}, D_t) \tag{2.45}$$

because given $\boldsymbol{\beta}_{t+1}$ one obtains all subsequent parameters without reference to $\boldsymbol{\beta}_t$.

Also,

$$f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N, \boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t|D_t) = f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N|\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t, D_t)\, f(\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t|D_t) \tag{2.46}$$

The LHS of (2.46) is

$$\int f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N, \boldsymbol{\beta}_N, \ldots, \boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t|D_t)\, d\boldsymbol{\beta}_N \cdots d\boldsymbol{\beta}_{t+2} =$$

$$= \int f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N|\boldsymbol{\beta}_N, \ldots, \boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t, D_t)\, f(\boldsymbol{\beta}_N, \ldots, \boldsymbol{\beta}_{t+2}|\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t, D_t) \cdot f(\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t|D_t)\, d\boldsymbol{\beta}_N \cdots d\boldsymbol{\beta}_{t+2}$$

$$= f(\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t|D_t) \int f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N|\boldsymbol{\beta}_N, \ldots, \boldsymbol{\beta}_{t+2}, \boldsymbol{\beta}_{t+1}, D_t) \cdot f(\boldsymbol{\beta}_N, \ldots, \boldsymbol{\beta}_{t+2}|\boldsymbol{\beta}_{t+1}, D_t)\, d\boldsymbol{\beta}_N \cdots d\boldsymbol{\beta}_{t+2}$$

$$= f(\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t|D_t) \int f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N, \boldsymbol{\beta}_N, \ldots, \boldsymbol{\beta}_{t+2}|\boldsymbol{\beta}_{t+1}, D_t)\, d\boldsymbol{\beta}_N \cdots d\boldsymbol{\beta}_{t+2}$$

$$= f(\boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t|D_t)\, f(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_N|\boldsymbol{\beta}_{t+1}, D_t) \tag{2.47}$$

Equating (2.47) to the RHS of (2.46) gives

$$f(y_{t+1}, \ldots, y_N \mid \beta_{t+1}, \beta_t, D_t) = f(y_{t+1}, \ldots, y_N \mid \beta_{t+1}, D_t) \tag{2.48}$$

Using (2.36) and (2.48) gives

$$f(\beta_{t+1}, \beta_t \mid D_N) \propto f(y_{t+1}, \ldots, y_N \mid \beta_{t+1}, D_t)\, f(\beta_{t+1} \mid D_t)\, f(\beta_t \mid \beta_{t+1}, D_t) \tag{2.49}$$

and finally

$$f(\beta_{t+1}, \beta_t \mid D_N) = f(\beta_t \mid \beta_{t+1}, D_t)\, f(\beta_{t+1} \mid D_N) \tag{2.50}$$

where (2.50) is obtained from (2.49) as (2.38) from (2.37). The procedure is then exactly as in (a) with t replacing N-1 and t+1 replacing N in (2.39)-(2.42) giving $\mathbf{m}_t^N$ and $\mathbf{C}_t^N$ as in (2.7).

Corollary 2.1 Under a dynamic NLM with non-linear transfer function $\mathbf{G}_t$ and/or non-linear regression function $\mathbf{s}_t$, t= 1, 2, ..., N, the smoothed parametric distributions are

$$[\beta_t \mid D_N] \sim [\mathbf{m}_t^N, \mathbf{C}_t^N]$$

with

$$\mathbf{m}_t^N = \mathbf{m}_t{}^* + \mathbf{C}_t{}^* \mathbf{H}_{t+1}^T (\mathbf{R}_{t+1}{}^*)^{-1} (\mathbf{m}_{t+1}^N - \mathbf{a}_{t+1}{}^*) \tag{2.51a}$$

$$\mathbf{C}_t^N = \mathbf{C}_t{}^* - \mathbf{C}_t{}^* \mathbf{H}_{t+1}^T (\mathbf{R}_{t+1}{}^*)^{-1} (\mathbf{R}_{t+1}{}^* - \mathbf{C}_{t+1}^N)(\mathbf{R}_{t+1}{}^*)^{-1} \mathbf{H}_{t+1} \mathbf{C}_t{}^* \tag{2.51b}$$

where $\mathbf{H}_t$ is given by (2.30a) and $\mathbf{M}_N^N = \mathbf{m}_N{}^*$ and $\mathbf{C}_N^N = \mathbf{C}_N{}^*$.

Proof: the proof follows exactly along the same lines of the proof of theorem 2.1 with the sole replacement of $(\mathbf{m}_{N-1}, \mathbf{C}_{N-1}, \mathbf{G}_N, \mathbf{a}_N, \mathbf{R}_N)$ by $(\mathbf{m}_{N-1}{}^*, \mathbf{C}_{N-1}{}^*, \mathbf{H}_N, \mathbf{a}_N{}^*, \mathbf{R}_N{}^*)$ in (2.40)-(2.42).

In fact, the smoothed parametric moments can be written under the general formula

$$E[\beta_t \mid D_N] = E[\beta_t \mid D_t] + Cov[(\beta_t, \beta_{t+1}) \mid D_t]\, V^{-1}[\beta_{t+1} \mid D_t] \cdot \{ E[\beta_{t+1} \mid D_N] - E[\beta_{t+1} \mid D_t] \} \tag{2.52a}$$

$$V[\beta_t \mid D_N] = V[\beta_t \mid D_t] + Cov[(\beta_t, \beta_{t+1}) \mid D_t]\, V^{-1}[\beta_{t+1} \mid D_t] \cdot \{ V[\beta_{t+1} \mid D_t] - V[\beta_{t+1} \mid D_N] \}\, V^{-1}[\beta_{t+1} \mid D_t]\, Cov^T[(\beta_t, \beta_{t+1}) \mid D_t] \tag{2.52b}$$

A linear transfer function $\mathbf{G}_t$ applied to (2.52) reproduces (2.7) while a linear approximation of the sort (2.29) to a non-linear function $\mathbf{G}_t$ gives (2.51). One of the crucial points in the proof is to show that given the parameter for a certain time, the distribution of future occurrences does not depend on any other previous parameter. Although the result seems

intuitively to hold because of the way the system equation was constructed, it takes some lines to be proven. This result is a consequence of the Markovian structure of the model and is also responsible for the extension of these smoothing results to survival models.

2.3 - Dynamic Models in Continuous Time

The initial reason to extend the dynamic models to continuous time is to have the ability to analyse processes with observations coming at arbitrary points. In doing this, the system is also ready to deal with analysis of processes where observations are made at every time. The models presented here are a natural extension of the DLM presented in section 2.1. The extension of the material of section 2.2 is presented in chapters 6 and 7. The presentation of this section is mainly based in Jazwinski (1970) and Maybeck (1979). The necessary extensions are presented here with a view to applications to GLM, so that, the normality (or Gaussianity, as it is commonly referred to by engineers) assumption is dropped whenever it is not essential for further developments.

A p-dimensional stochastic process $\{\mathbf{x}(t), t \in T\}$ is said to have independent increments if for all finite sets $\{t_i, i = 1, ..., n\} \subset T$, the random vectors $\mathbf{x}(t_2)-\mathbf{x}(t_1)$, $\mathbf{x}(t_3)-\mathbf{x}(t_2)$,..., $\mathbf{x}(t_n)-\mathbf{x}(t_{n-1})$ are independent. If, in addition, $E[\mathbf{x}(t_i) - \mathbf{x}(t_{i-1})] = 0$, i= 1,..., n-1, it is said to have orthogonal increments (Doob, 1953, pg. 99) where $t_0 = \min T$ is generally taken as 0. Orthogonal independent increments (OII) processes provide a suitable structure to model system disturbances in continuous time. A Brownian motion (BM) process can be defined as an OII process such that for every $t \in T$, $\mathbf{x}(t)$ is normally distributed. In general, those processes have a diffusion $\mathbf{Q}$ associated with them so that for every $s < t \in T$

$$E[|\mathbf{x}(t) - \mathbf{x}(s)|^2] = \mathbf{Q}(t - s) \tag{2.53}$$

A related process is the white (W) process. $\{\mathbf{w}(t), t \in T\}$ is white if for all finite sets $\{t_i, i = 1,..., n\} \subset T$, $\mathbf{w}(t_1)$, ..., $\mathbf{w}(t_n)$ are independent random variables. If, in addition they are assumed to be normally distributed, the process is called a white Gaussian (WG) process. It should be stressed that all properties of OII and W processes are properties of BM

and WG processes, respectively. In order to highlight the relation between OII and W processes and to prepare for the stochastic calculus needed in the sequel, a suitable notion of continuity for random processes is needed. A stochastic process $\{\mathbf{u}(t), t \in T\}$ is mean square continuous if, for all t,

$$\lim_{h \to 0} E[\,|\mathbf{u}(t+h) - \mathbf{u}(t)|^2\,] = 0 \tag{2.54}$$

and the notation is $l.i.m._{h \to 0}\, \mathbf{u}(t+h) = \mathbf{u}(t)$. $\mathbf{u}(t)$ is said to be the limit in the mean of $\mathbf{u}(t-h)$ as $h \to 0$. With this notion, stochastic processes can be differentiated and integrated. The first result that can be obtained is that OII processes are (mean square) continuous but not differentiable. Following (2.54), a process $\{\mathbf{u}(t), t \in T\}$ is differentiable if

$$l.i.m._{h \to 0} \frac{\mathbf{u}(t+h) - \mathbf{u}(t)}{h}$$

exists, and in this case, is denoted by $\dot{\mathbf{u}}(t)$. From the properties of OII processes and from (2.53), if $\{\mathbf{u}(t), t \in T\}$ is such a process, then

$$E[\,\mathbf{u}(t+h) - \mathbf{u}(t)\,] = 0 \tag{2.55a}$$

$$E[\,|\mathbf{u}(t+h) - \mathbf{u}(t)|^2\,] = h\,\mathbf{Q} \tag{2.55b}$$

From (2.55) it is clear that $l.i.m._{h \to 0}\, \mathbf{u}(t+h) = \mathbf{u}(t)$ but

$$E\left[\frac{|\mathbf{u}(t+h) - \mathbf{u}(t)|^2}{h}\right] = \frac{\mathbf{Q}}{h} \tag{2.56}$$

and as $h \to 0$, this goes to infinity. So, OII processes have unbounded variation but can be formally related to W processes via

$$\mathbf{w}(t) = \frac{d\,\mathbf{u}(t)}{dt} \quad or \quad \mathbf{u}(t) = \int_0^t \mathbf{w}(s)\,ds \tag{2.57}$$

for a process $\mathbf{w}(t)$ with mean function $\mathbf{m}_{\mathbf{w}}(t) = E[\,\mathbf{w}(t)\,] = 0$ and covariance function $\mathbf{P}_{\mathbf{w}}(t,s) = Cov[\,\mathbf{w}(t), \mathbf{w}(s)\,] = \mathbf{Q}\,\delta(t-s)$ where δ is the Dirac delta function. This characterisation can be extended by allowing for time-varying diffusions $\mathbf{Q}(t)$ so that (2.53) is now

$$E[\,|\mathbf{x}(t) - \mathbf{x}(s)|^2\,] = \int_s^t \mathbf{Q}(u)\,du \tag{2.58}$$

The corresponding formal derivative W (WG) process $\mathbf{w}(t)$ will now have $\mathbf{P}_{\mathbf{w}}(t,s) = \mathbf{Q}(t)\delta(t-s)$.

Stochastic integration of forms like

$$\int_a^b \Pi(t)\, d\mathbf{u}(t) \tag{2.59}$$

where $\mathbf{u}(t)$ is a stochastic process can be accomplished with mean square limits. The case when $\Pi(t)$ is deterministic was introduced by Wiener (1930) and is presented here. This was later extended by Itô (1944) to the case of random functions $\Pi(t)$ but is beyond the needs of this thesis. In both cases, however, (2.59) is defined as

$$l.i.m._{\rho \to 0} \sum_{i=0}^{N-1} \Pi(t_i)\,[\mathbf{u}(t_{i+1}) - \mathbf{u}(t_i)] \tag{2.60}$$

where $\rho = \max_i (t_i - t_{i-1})$ and $a = t_0 < t_1 < t_2 < \cdots < t_N = b$. A full account of these stochastic integrals can be found in Maybeck (1979).

The system equation can now be set in a stochastic differential form

$$d\boldsymbol{\beta}(t) = \mathbf{F}(t)\boldsymbol{\beta}(t)\, dt + \Pi(t)\, d\mathbf{u}(t) \tag{2.61}$$

where $\{\boldsymbol{\beta}(t), t \geq 0\}$ is a p-dimensional stochastic process representing the system, $\{\mathbf{u}(t), t \geq 0\}$ is a q-dimensional OII process representing the system disturbances and Γ and $\mathbf{F}$ are $p \times q$ and $p \times p$ matrix functions, respectively. This can be made formally equivalent to

$$\dot{\boldsymbol{\beta}}(t) = \mathbf{F}(t)\boldsymbol{\beta}(t) + \Pi(t)\mathbf{w}(t) \tag{2.62}$$

where $\mathbf{w}(t)$ is the corresponding W process. The solution of (2.61) is

$$\boldsymbol{\beta}(t) = \mathbf{G}(t,0)\boldsymbol{\beta}(0) + \int_0^t \mathbf{G}(t,s)\Pi(s)\, d\mathbf{u}(s) \tag{2.63}$$

with $\mathbf{G}$ satisfying

$$\frac{d\mathbf{G}(t,s)}{dt} = \mathbf{F}(t)\mathbf{G}(t,s)$$

and the initial condition $\mathbf{G}(0,0) = \mathbf{I}$. Setting the initial condition of the process to

$$\beta(0) \sim \mathrm{m}(0), \mathrm{C}(0) \tag{2.64}$$

independent of $\mathrm{u}(t)$, the process is completely defined and its moments can be obtained from (2.60) and (2.63). If Gaussianity is assumed for $\mathrm{u}(t)$ and for $\beta(0)$, then $\beta(t)$ is also Gaussian. The transition matrix $\mathbf{G}$ satisfies $\mathbf{G}(c,b)\mathbf{G}(b,a) = \mathbf{G}(c,a)$, $0 \leq a < b < c$, so that after defining $0 \leq t_1 < t_2 < \cdots < t_N$, the system can be recursively obtained at these points as

$$\beta(t_i) = \mathbf{G}(t_i, t_{i-1})\beta(t_{i-1}) + \int_{t_{i-1}}^{t_i} \mathbf{G}(t_i, s)\Gamma(s)\, d\mathrm{u}(s) \tag{2.65}$$

This is a continuous version of (2.8). When $\{t_i, i = 1,2,\ldots,N\}$ represent different arbitrary observation points, it is useful to write $\mathbf{G}(t_i,s) = \mathbf{G}_i(t_i - s)$ since, in most applications the transition over an interval is only dependent on the length of that interval. In the case of discrete observations, the identification $\beta(t_i) = \beta_i$, i= 0, 1, ..., N is made since those are the only system points that matter. The moments of β_i can be recursively obtained from (2.60), (2.65) and the initial condition (2.64) as

$$E[\beta_i] = \mathbf{G}_i(b_i)\, E[\beta_{i-1}] \tag{2.66a}$$

$$V[\beta_i] = \mathbf{G}_i(b_i)\, V[\beta_{i-1}]\, \mathbf{G}_i^T(b_i) + \int_{t_{i-1}}^{t_i} \mathbf{G}_i(t_i - s)\Gamma(s)\mathbf{Q}(s)\Gamma^T(s)\mathbf{G}_i(t_i - s)\, ds \tag{2.66b}$$

where all the moments are conditional on the same information and $b_i = t_i - t_{i-1}$ is the length of the interval between observation points, i= 1, ..., N. Again, if $\mathrm{u}(t)$ and β_0 are Gaussian, β_i is also Gaussian with moments given by (2.66). The generalisation of the DLM for observation points $0 \leq t_1 < t_2 < \cdots$ is characterised by

(i) *Observation equation* : $\mathbf{y}_i = \mathbf{Z}_i\beta_i + \mathbf{v}_i$, $\mathbf{v}_i \sim N(0, \mathbf{V}_i)$

where $\mathbf{y}_i$ is an r-dimensional observation made at time t_i and $\beta_i = \beta(t_i)$.

(ii) *System equation* : $d\beta(t) = \mathbf{F}(t)\beta(t)\, dt + \Gamma(t)\, d\mathrm{u}(t)$

where $\mathbf{F}(t)$ and $\Gamma(t)$ are as in (2.61) and $\mathrm{u}(t)$ is a BM process with time varying diffusion $\mathbf{Q}$.

For the analysis of this model, the differential equation in (ii) can be specialised to the difference system equation

$$\boldsymbol{\beta}_i = \mathbf{G}_i(b_i)\boldsymbol{\beta}_{i-1} + \mathbf{w}_i \quad , \mathbf{w}_i \sim N(0, \mathbf{W}_i) \tag{2.67}$$

for i= 1, 2, ..., where $\mathbf{w}_i$ and $\mathbf{W}_i$ are respectively

$$\int_{t_{i-1}}^{t_i} \mathbf{G}_i(t_i - s)\boldsymbol{\Pi}(s)\, d\mathbf{u}(s) \quad \text{and} \quad \int_{t_{i-1}}^{t_i} \mathbf{G}_i(t_i - s)\boldsymbol{\Pi}(s)\mathbf{Q}(s)\boldsymbol{\Gamma}^T(s)\mathbf{G}_i^T(t_i - s)\, ds$$

The analysis then follows exactly as in section 2.1 with $\mathbf{W}_i$ as above and $\mathbf{G}_i(b_i)$ replacing $\mathbf{G}_i$ as a reminder that interval lengths are no longer unitary. The same comments apply for the analysis of dynamic GLM with arbitrary observation points without the normality on the system equation. If interval lengths are small and/or the system is time invariant ($\mathbf{G}_i = \mathbf{I}$ or $\mathbf{F}(t)=0$) or slowly varying ($\mathbf{G}_i \approx \mathbf{I}$), $\mathbf{W}_i$ can be approximated to

$$b_i\, \boldsymbol{\Pi}(t_i)\mathbf{Q}(t_i)\boldsymbol{\Gamma}^T(t_i)$$

When taking the system disturbance as scalar for applications, this gives

$$\mathbf{W}_i = b_i\, \boldsymbol{\sigma}_i \boldsymbol{\sigma}_i^T \tag{2.68}$$

for $\boldsymbol{\sigma}_i = \boldsymbol{\Pi}(t_i)/\sqrt{\mathbf{Q}(t_i)}$, i= 1, 2, ...

These models can now be extended to the case when observations are made continuously in time. The observation equation is

$$\mathbf{y}(t) = \mathbf{Z}(t)\boldsymbol{\beta}(t) + \mathbf{v}(t)\,, \; t \geq 0 \tag{2.69}$$

where $\mathbf{v}(t)$ is a WG process such that $E[\mathbf{v}(t)] = 0$ and $E[\mathbf{v}(t)\mathbf{v}^T(s)] = \mathbf{V}(t)\delta(t-s)$ and, as before, to give mathematical meaning to (2.69), the formal relations $\mathbf{y}(t) = \dot{\mathbf{x}}(t)$ and $\mathbf{v}(t) = \dot{\nu}(t)$ lead to

$$d\mathbf{x}(t) = \mathbf{Z}(t)\boldsymbol{\beta}(t)\, dt + d\nu(t)$$

The system equation remains unchanged and, defining $D(t) = \{D_0, \mathbf{y}(s), s \in [0,t]\}$, leads to

$$[\boldsymbol{\beta}(t) \mid D(t)] \sim N[\mathbf{m}(t), \mathbf{C}(t)] \quad , t \geq 0 \tag{2.70}$$

where $\mathbf{m}(t)$ and $\mathbf{C}(t)$ satisfy the differential equations

$$\dot{\mathbf{m}}(t) = \mathbf{F}(t)\mathbf{m}(t) + \mathbf{C}(t)\mathbf{Z}^T(t)\mathbf{V}^{-1}(t)[\mathbf{y}(t) - \mathbf{Z}(t)\mathbf{m}(t)] \tag{2.71a}$$

$$\dot{\mathbf{C}}(t) = \mathbf{F}(t)\mathbf{C}(t) + \mathbf{C}(t)\mathbf{F}^T(t) + \boldsymbol{\Gamma}(t)\mathbf{Q}(t)\boldsymbol{\Gamma}^T(t) - \mathbf{C}(t)\mathbf{Z}^T(t)\mathbf{V}^{-1}(t)\mathbf{Z}(t)\mathbf{C}(t) \qquad (2.71b)$$

Equations (2.71) are the continuous time Kalman filter and (2.71b) is often called a Riccati equation (Kalman & Bucy, 1961). They can also be obtained as a limit result when the time interval between successive observations goes to 0. A similar derivation is carried out in later chapters for non-normal observation equations.

References

Ameen, J. R. M. and Harrison, P. J. (1985). Normal discount Bayesian models. In *Bayesian Statistics 2*, (eds: J. M. Bernardo et al.). Amsterdam: North-Holland.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical analysis*. Reading: Addison-Wesley.

Doob, J. L. (1953). *Stochastic Processes*. New York: John Wiley.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting (with discussion). *J. R. Statist. Soc., Ser. B*, **38** , 205-247.

Harrison, P. J. & West, M. (1986). Bayesian forecasting in practice. Warwick Bayesian Statistics Study Year report no. 13.

Itô, K. (1944). Stochastic integral. *Proc. Imp. Acad. Tokyo,* **20**, 519-524.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Trans. ASME J. Basic Eng.*, **83**, 95-108.

Maybeck, P. S. (1979). *Stochastic Models, Estimation and Control, Vol. 1*. New York: Academic Press.

-------(1982). *Stochastic Models, Estimation and Control, Vol. 2*. New York: Academic Press.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. London: Chapman & Hall.

McGarty, T. P. (1974). *Stochastic Systems and State Estimation*. New York: John Wiley.

Migon, H. S. (1984). An approach to non-linear Bayesian forecasting problems with applications. Unpublished Ph.D. thesis, Univ. of Warwick.

Migon, H. S. and Harrison, P. J. (1985). An application of non-linear Bayesian forecasting to television advertising. In *Bayesian Statistics 2*. (eds: J. M. Bernardo et al.). Amsterdam: North-Holland.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalised linear models. *J. R. Statist. Soc., Ser. A*, **135**, 370-384.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications (Second edition)*. New York: John Wiley.

West, M. (1982). Aspects of recursive Bayesian estimation. Unpublished Ph.D. thesis, Univ. of Nottingham.

West, M. and Harrison, P. J. (1986). Monitoring and adaptation in Bayesian forecasting models. *J. Amer. Statist. Ass.*, **81**, 741-750.

West, M., Harrison, P. J. and Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Amer. Statist. Ass.*, **80**, 73-97.

Wiener, N. (1930). Generalized harmonic analyses. *Acta Math.*, **55**, 117-258.

Wong, E. (1971). *Stochastic Processes in Information and Dynamical Systems*. New York: McGraw-Hill.

CHAPTER THREE

ANALYSIS FOR SURVIVAL DATA

This chapter introduces the survival models and the methods of analysis. Almost all the work done in this area uses the hazard function as a basis for modelling. This thesis is no exception in its recognition of the intuitive appeal of the hazard in terms of risk and this is fully exploited here. The main difference lies in the use of dynamic models as opposed to the commonly used static models in the context of survival analysis. This is discussed among other things in the introduction. The use of dynamic ideas implies an association of the information obtained with some sort of time ordering. This is accomplished via an essential yet simple result factorising the likelihood introduced in section 3.2. The suitable and flexible class of piecewise exponential (PE) distributions is presented in the next section. After that, the models are ready to be put into the structure of dynamic GLM and the parameters can be estimated. The tools for checking the adequacy of a given model and predicting the performances of given individuals are provided in the final sections. Throughout this chapter, it will be assumed that the covariates are constant in time. The case when the covariates depend on time is dealt with in the next chapter.

3.1 - Introduction

As anticipated in chapter 1, most survival models have hazard function of the form

$$\lambda(t;\mathbf{z}) = \lambda_0(t)\, g(\mathbf{z}) \tag{3.1}$$

where $\mathbf{z}$ is a vector of covariates. These are called multiplicative, proportional hazards or static models. The name multiplicative stems from the form of the hazard with two multiplicative factors: one relating to the general failure pattern and the other one carrying the effect of the covariates. If two individuals have covariates $\mathbf{z}_1$ and $\mathbf{z}_2$, respectively, writing $c = g(\mathbf{z}_1)/g(\mathbf{z}_2)$ gives

$$\lambda(t;\mathbf{z}_1) = c\ \lambda(t;\mathbf{z}_2) \tag{3.2}$$

This is the reason for the proportional hazards name; the hazard functions of different individuals are always proportional as functions of time. One of the consequences of these models can be seen after replacing (3.2) in (1.5). The survival functions for two individuals relate via

$$S_2(t) = [S_1(t)]^c \tag{3.3}$$

and, since survival functions are non-increasing, they never cross. The covariates have their effects completely separated from time and, once their values are obtained, the characteristics of the different individuals are unchanged. These models assume a static effect of the covariates on the failure pattern of the individuals.

These models were very important in developing the analysis of survival models but practitioners began to question the assumptions as their applications started to point otherwise. Specifically in medicine, where survival models are extensively used in cancer treatment, there is strong evidence that the effects some covariates have in the survival of a patient do change with time. Some prognostic factors like tumour size may have a decreasing effect in explaining death by cancer and some treatments have a definitely changing effect as time passes. In economics, when studying duration of unemployment spell, it is reasonable to expect some social factors to be more relevant in early stages and others in later ones. Although this may seems obvious to the layman, it is not until recently that statisticians began to address to this question and build more appropriate models.

To understand some of the work that has been done, one has to go back to proportional hazards models. The function g is generally taken as

$$g(\mathbf{z}) = \exp(\mathbf{z}\boldsymbol{\beta}) \tag{3.4}$$

where $\mathbf{z} = (z_1\, z_2 \cdots z_p)$ is the vector of covariates and $\boldsymbol{\beta}$ is the p-dimensional parameter modelling their effects. This is a natural way to ensure that $g \geq 0$. The models proposed fall broadly into two categories: complete ignorance and complete knowledge about the way the effects change with time.

In the first category, are the models proposed by Anderson & Senthilselvan (1982) and later extended by Gore et al. (1984). The original idea was to divide the time scale R^+ in two meaningful intervals and assume a different value for β for each of these intervals. Gore et al. proposed the division of time scale in k intervals each with a different value of β. The estimation can be carried out separately within each interval with the use of the factorisation ideas presented in the next section. The important point is that the β's of different intervals are assumed to have no relation whatsoever and are estimated independently. Aware of this fact, Gore et al were extremely cautious making sure there was enough information in each interval ending up with a choice of 3 intervals (Gore et al., 1984, pg. 190). The ignorance in this case refers to the fact that, although the β's of different intervals measure the effect of the same covariates, there is no structure relating them.

In the later case, β is modelled as a specified function of time $\beta(t)$. Among the proposed ones are a linear function of time (Cox & Oakes, 1984, pg. 113 and Carter et al., 1983), a quadratic function of time (Stablein et al., 1981) and an exponential decay (Gore et al., 1984) where each component of β is set to

$$\beta_i(t) = c_i \exp(-\rho_i t) \quad , i = 1, ..., p \tag{3.5}$$

The parameters modelling these functions are generally unknown and are estimated from the observed data but the form that β depends on time is assumed completely known from the onset.

From the discussion in the previous chapter, it seems more sensible to allow for a dynamic model for the parameters. This provides a way of blending prior† assumptions with observed data. The approaches above seem to rely too heavily in either observed data in the first group and prior assumptions in the second one. A compromise is not only useful but desirable.

† where prior here stands for prior to observing the data.

The same comments apply to the baseline hazard function $\lambda_0(t)$. Survival models originally proposed it to be from a specific parametric family, generally the exponential or Weibull distribution. A major breakthrough came when Cox (1972) introduced the possibility of estimating λ_0 non-parametrically, that is, with no specific distributional assumption. After dividing the time scale in intervals, $\lambda_0(t)$ was estimated independently within each interval. Cox himself recognises this as unnecessary saying that some smoothness assumptions appear to be more reasonable. In fact, apart from a few exceptional cases, one expects the hazard to be a smooth function where values from adjacent intervals are related. Once again, a dynamic structure is appropriate by conveying smoothness assumptions without imposing a fully specified parametric form for the hazard.

Suppose the time scale is divided in N intervals and that corresponding to each interval I_i there is a vector of parameters $\boldsymbol{\beta}_i$ modelling the baseline hazard and the effect of the covariates. A dynamic structure can be set, as in chapter 2, via

$$\boldsymbol{\beta}_i = \mathbf{G}_i(\boldsymbol{\beta}_{i-1}) + \mathbf{w}_i \quad , i = 1, 2, ..., N \tag{3.6}$$

where $\mathbf{w}_i$ is an error increasing the modeller's uncertainty about the process as one moves from interval I_{i-1} to I_i. Each $\boldsymbol{\beta}_i$ can be partitioned into $\boldsymbol{\beta}_{i1}$ and $\boldsymbol{\beta}_{i2}$ modelling respectively the hazard function and the covariates' effects and $\mathbf{w}_i$ can be analogously partitioned into $\mathbf{w}_{i1}$ and $\mathbf{w}_{i2}$. The models discussed above can be put into the structure (3.6) as follows:

(a) complete ignorance about the covariates' effects change: $V^{-1}(\mathbf{w}_{i2}) \to 0, i = 1, 2, ..., N$.

(b) complete knowledge about the covariates' effects change: $\mathbf{w}_{i2} \equiv 0, i = 1, 2, ..., N$;

if, in addition, exponential decay is assumed for the p covariates, then $\boldsymbol{\beta}_{i2}^T = (\theta_{i1}, ..., \theta_{ip}, \rho_{i1}, \ldots, \rho_{ip})$, $\mathbf{G}_i(\boldsymbol{\beta}_{i-1}) = diag\,(\mathbf{G}_{i1}(\boldsymbol{\beta}_{i-1,1}), \mathbf{G}_{i2}(\boldsymbol{\beta}_{i-1,2}))$ where $\mathbf{G}_{i2}(\boldsymbol{\beta}_{i-1,2}) = diag\,(\theta_{i-1,1}\, exp(-\rho_{i-1,1}\,\delta t), ..., \theta_{i-1,p}\, exp(-\rho_{i-1,p}\,\delta t), 1, ..., 1)$ with δt as a measure of the distance between I_{i-1} and I_i generally taken as mid-interval distance.

(c) complete knowledge of the failure pattern of the population: $\mathbf{w}_{i1} \equiv 0, i = 1, 2, ..., N$.

(d) non-parametric estimation of the hazard function: $V^{-1}(\mathbf{w}_{i1}) \to 0, i = 1, 2, ..., N$.

It is now clear that the models that have been proposed do not fully explore a dynamic approach. They always go to the extreme of either having no system disturbances or having extremely large ones. The models developed here have a system equation like (3.6) that can eventually attain degenerate values for some of its components but in general stays in between them. This relates effects from different intervals in time by passing the information obtained within any of them. To do this, however, one has to be able to factorise the information from the observed data according to prespecified intervals. This is developed in the next section.

3.2 - Temporal Factorisation of the Likelihood

In survival models, every individual has a contribution to the likelihood of the form $p(y|\lambda)$ according to the notation established in chapter 1 where λ is the parameter modelling the survival time distribution for this individual. Suppose that after a partition $\tau = \{t_1, ..., t_{s-1}\}$ is performed in $\mathbb{R}^+$, the intervals $I_1 = [0, t_1]$, $I_i = (t_{i-1}, t_i]$, $i = 2, ..., s-1$ and $I_s = (t_{s-1}, \infty)$ are constructed. If $y \in I_i$, then

$$p(y|\lambda) = \frac{p(y|\lambda)}{S(t_{i-1}|\lambda)} \cdot \frac{S(t_{i-1}|\lambda)}{S(t_{i-2}|\lambda)} \cdots S(t_1|\lambda)$$

$$= \frac{p(y|\lambda)}{S(t_{i-1}|\lambda)} \prod_{j=1}^{i-1} \frac{S(t_j|\lambda)}{S(t_{j-1}|\lambda)} \tag{3.7}$$

where $S(t_0|\lambda) = 1$ since $t_0 = 0$. Using conditional probability, the above expression can be rewritten

$$p(y|\lambda) = p(y|Y > t_{i-1}, \lambda) \prod_{j=1}^{i-1} S(t_j|Y > t_{j-1}, \lambda) \tag{3.8}$$

The above expression tells that the total contribution from Y to the likelihood can be factored. This factorisation gives terms that are related to the information about Y obtained in one interval. For every interval I_j, $j = 1, ..., i-1$, all that is observed is that having Y entered it (the individual having survived up to the end of the previous interval), it survives through all the interval leading to the contribution $S(t_j|Y > t_{j-1}, \lambda)$. In I_i, one

observes that Y has survived up to the end of I_{i-1} but endures as far as y. The likelihood contribution $p(y \mid Y > t_{i-1}, \lambda)$ says that the observation stops at y either because the individual died in which case p is replaced by f or because the individual was censored and p is replaced by S. Of course, there is no further information about Y in intervals $I_{i+1},\ldots,I_s$. The factorisation also shows that, conditional on the survival up to the end of the previous interval, the interval contributions are independent. In fact, one can construct the variables $Y_1,\ldots,Y_s$ with

$$Y_i = \begin{cases} min(Y, t_i) \mid Y > t_{i-1} & \text{, if } Y > t_{i-1} \\ t_{i-1} & \text{, otherwise} \end{cases} \quad , i = 1,\ldots,s$$

Each Y_i gives the information about Y obtained in interval I_i. They are independent with a corresponding contribution to (3.8).

The information factorisation has wider uses and has long since been used. The celebrated product-limit (PL) estimator (Kaplan & Meier, 1958) uses this idea for non-parametric estimation in random samples. Rather than estimating an unknown distribution using empirical distribution functions, it estimates the survival function by equating each factor in (3.8) to the surviving proportion of individual at the end of the interval among those alive at its beginning. The case of a random sample is analysed in chapter 5 where the PL estimator is presented. It is worth pointing out that the actuarial estimator (Elandt-Johnson & Johnson, 1980, pg. 157) also uses the same ideas with a different partition of the time scale. The conditional likelihood used by Cox (1972) also uses this idea in extending the PL estimator for the case when covariates are present. It is a product of factors related to different intervals where each factor is based on individuals that have survived up to the end of the previous interval.

It seems that factorisation ideas are particularly useful for survival models due to the presence of censoring and are employed here to the full likelihood obtained from the observations. The result is valid for any countable partition of the time scale and for dependent observations as well.

Theorem 3.1 (TFL) Let $\mathbf{Y} = (Y_1, \ldots, Y_m)$ have joint density $f(\mathbf{y} \mid \lambda)$ in $(\mathbb{R}^+)^m$. Define $\mathcal{T} = \{t_1, t_2, \ldots, t_{n-1}\}$ where $0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = \infty$ and $I_1 = [0, t_1)$, $I_i = (t_{i-1}, t_i]$, $i = 2, 3, \ldots, n-1$ and $I_n = (t_{n-1}, \infty)$. Then, the likelihood for λ obtained after observing $\mathbf{y}$ is

$$\prod_{i=1}^{n} l_i(\mathbf{y}_i \mid D_{i-1}, \lambda) \tag{3.9}$$

where

$$l_i(\mathbf{y}_i \mid D_{i-1}, \lambda) = p(\mathbf{y}_i \mid D_{i-1}, \lambda) \, / \, p(y_{i1}, \ldots, y_{ir_i} \mid D_{i-1}, \lambda)$$

with $y_{ij} = min(y_j, t_i)$, y_j is the survival time of the j^{th} known to be alive at t_{i-1} $(j = 1, \ldots, r_i)$, r_i is the number of individual known to be alive at t_{i-1} and $D_i = \{D_0, \textit{data information obtained in } \bigcup_{j=1}^{i} I_j\}$.

Before going into the proof, it is important to establish that the main result is (3.9) ensuring that it is possible to factorise the likelihood according to arbitrary intervals conditioned to previous information. This factorisation is always based in a grid of times and is carried out taking this time ordering sequentially. The name temporal factorisation of the likelihood (TFL, in short) seems therefore appropriate to describe it and to emphasise its importance to time-dependent analysis. The labels $1, \ldots, r_i$ are given to the individuals contributing to l_i without loss of generality. Their labels are irrelevant as long as each individual can be readily identified at every stage of the factorisation.

Proof: Define the likelihood remainders as

$$\mathcal{R}_k = p(y_1, \ldots, y_{r_k} \mid D_{k-1}, \lambda) \quad , k = 1, \ldots, n$$

in the sense that the likelihood for λ after observing $\mathbf{y}$ is

$$\left[\prod_{j=1}^{i} l_i(\mathbf{y}_j \mid D_{j-1}, \lambda) \right] \mathcal{R}_{i+1} \quad , i = 0, 1, \ldots, n-1 \tag{3.10}$$

(3.10) is obtained from (3.9) applied with $\mathcal{T} = \{t_1, \ldots, t_k\}$. So, to prove the theorem, it suffices to prove that (3.10) is the likelihood for λ.

(i) if i=0, (3.10) reduces to $\mathcal{R}_1 = p(y_1, \ldots, y_m \mid D_0, \lambda)$ which is the likelihood for λ.

(ii) suppose (3.10) is the likelihood for λ for i=k-1. If

$$\mathcal{R}_k = l_k(y_k \mid D_{k-1}, \lambda)\, \mathcal{R}_{k+1} \tag{3.11}$$

then

$$\left[\prod_{j=1}^{k-1} l_j\right] \mathcal{R}_k = \left[\prod_{j=1}^{k-1} l_j\right] l_k\, \mathcal{R}_{k+1} = \left[\prod_{j=1}^{k} l_j\right] \mathcal{R}_{k+1}$$

so (3.10) is also valid for i=k and the inductive proof is completed. All that needs to be proven then is (3.11).

Let $\mathcal{F}_{k+1} = \{j ; Y_j \text{ dies after } t_k\}$, $\mathcal{C}_{k+1} = \{j ; Y_j \text{ censored after } t_k\}$, $\mathcal{F}_k = \{j ; Y_j \text{ dies in } I_k\}$ and $\mathcal{C}_k = \{j ; Y_j \text{ censored in } I_k\}$. Define the information sets

$$A = [\, Y_j \in \Delta Y_{k,j}, j \in \mathcal{F}_k \uplus \mathcal{F}_{k+1} ;\ Y_l > y_l, l \in \mathcal{C}_k \uplus \mathcal{C}_{k+1}]$$

$$A_1 = [\, Y_j \in \Delta Y_{k,j}, j \in \mathcal{F}_k ;\ Y_l > y_l, l \in \mathcal{C}_k ;\ Y_q > t_k, q \in \mathcal{F}_{k+1} \uplus \mathcal{C}_{k+1}]$$

$$A_2 = [\, Y_j \in \Delta Y_{k,j}, j \in \mathcal{F}_{k+1} ;\ Y_l > y_l, l \in \mathcal{C}_{k+1}]$$

where $\Delta Y_{k,j}$ is an interval of length $\Delta_{k,j}$ including y_j. Then

$$\mathcal{R}_k = \lim_{\rho \to 0} \frac{Pr(A \mid D_{k-1}, \lambda)}{\prod_{j \in \mathcal{F}_k \uplus \mathcal{F}_{k+1}} \Delta_{k,j}} \tag{3.12}$$

where $\rho = \max_j \{\Delta_{k,j}, j \in \mathcal{F}_k \uplus \mathcal{F}_{k+1}\}$.

$$Pr(A \mid D_{k-1}, \lambda) = Pr(A_1 \mid D_{k-1}, \lambda)\, Pr(A_2 \mid A_1, D_{k-1}, \lambda)$$

Substitution in (3.12) gives

$$\mathcal{R}_k = \lim_{\rho \to 0} \frac{Pr(A_1 \mid D_{k-1}, \lambda)}{\prod_{j \in \mathcal{F}_k} \Delta_{k,j}} \frac{Pr(A_2 \mid A_1, D_{k-1}, \lambda)}{\prod_{j \in \mathcal{F}_{k+1}} \Delta_{k,j}}$$

$$= \lim_{\rho \to 0} \frac{Pr(A_1 \mid D_{k-1}, \lambda)}{\prod_{j \in \mathcal{F}_k} \Delta_{k,j}} \lim_{\rho \to 0} \frac{Pr(A_2 \mid A_1, D_{k-1}, \lambda)}{\prod_{j \in \mathcal{F}_{k+1}} \Delta_{k,j}}$$

Taking limits as $\rho \to 0$ leads to

$$\mathcal{R}_k = p(t_k, \ldots, t_k, y_{r_{k+1}+1}, \ldots, y_{r_k} \mid D_{k-1}, \lambda)\, p(y_1, \ldots, y_{r_{k+1}} \mid D_k, \lambda)$$

$$= p(\min(y_1,t_k),\ldots,\min(y_{r_k},t_k) \mid D_{k-1},\lambda)\ \mathcal{R}_{k+1}$$

$$= l_k(y_k \mid D_{k-1},\lambda)\ \mathcal{R}_{k+1}$$

and the result is proven.

The TFL is needed in its full generality in section 3.5 to obtain the marginal likelihood for the observations. In that case, the parameters have already been integrated out and the parameter λ represents the model assumptions. Its likelihood is considered in relation to other possible models to evaluate the data support for them. Apart from that, the TFL is mainly used in its simplified form for independent observations conditional on the parameters where each l_i is

$$\prod_{j=1}^{r_i} p(y_{ij} \mid \lambda, D_{i-1}) = \prod_{w \in \mathcal{F}_i} f_w(y_{iw} \mid \lambda, D_{i-1}) \prod_{q \in S_i} S_q(y_{iq} \mid \lambda, D_{i-1}) \qquad (3.13)$$

where f_w is the density of Y_w and S_q is the survival of Y_q. If Y_j dies in I_i then $j \in \mathcal{F}_i$ and $y_{ij} = y_j$ and if Y_j survives through I_i or is censored in I_i then $j \in S_i$ and $y_{ij} = \min(y_j,t_i)$ and, of course, $\mathcal{F}_i \cup S_i = \{1,\ldots,r_i\}$. Even though the observations are independent, the dependence on the previous information D_{i-1} cannot be dropped from (3.13) because the occurrences related to each Y_j in I_i depend on what has already been observed about it in previous intervals. It is shown in the next section that this is easily accommodated in survival analysis. The grid $\mathcal{T}$ of points in time plays an important role in the subsequent analysis but apart from the monotonicity requirements there is no restriction for their choice. This choice is made by the modeller and it can be seen from the TFL that the points in $\mathcal{T}$ work as observation points where one pretends to observe what has happened in the preceding interval. In fact, the observation process is continuous in time and this discretisation is performed mainly for mathematical tractability of the problem and its practical use in applications. The analysis for a continuous observation process is discussed in chapter 7 giving rise to differential equations of the form (2.71). In practice, increasingly finer grids $\mathcal{T}$ can be defined up to the point of reaching the precision of the survival time measurement. Beyond that, any refinement is dubious since it is based on unavailable information.

3.3 - Piecewise Exponential Distributions

The idea of defining distributions in a piecewise sense is not new and has been explored in the areas of medical statistics, reliability and demography. In all those areas, a division of time is invariably performed for some reason and analysis of each interval separately applying ideas from the previous section is used. Elandt-Johnson & Johnson (1980) use a specific parametric model piecewisely defined in pre-determined intervals in the study of life tables. This point is discussed in chapter 5 where different models for the analysis for life tables are presented. The emphasis here is to model the hazard function without imposing any parametric family assumption and letting the data help the system in pointing out its direction. Suppose that, as in previous sections, a grid τ of points is given with corresponding intervals $I_1, ..., I_s$. Then, the simplest way to model the hazard function $\lambda(\cdot)$ is to assign a different constant value for each of the intervals. The choice of the intervals and their lengths are important bearing in mind that one is trying to capture the main features of a generally smooth functions using a piecewise constant or simple one. Cozzolino (1974) defined a piecewise constant failure rate tracing the origin of these ideas to Shooman (1968). The failure rate or hazard function is defined as

$$\lambda(y) = \begin{cases} \lambda_1 & , y \in I_1 = [0, t_1] \\ \cdot & \\ \cdot & \\ \cdot & \\ \lambda_s & , y \in I_s = (t_{s-1}, \infty) \end{cases} \tag{3.14}$$

Distributions having this form of hazard function are frequently called piecewise exponential distributions since within each interval it behaves like an exponential distribution. This class of distributions is quite rich and can be used to obtain any hazard function by letting the number of intervals go to ∞ and the lengths of the intervals go to 0. If Y has hazard (3.14), it is denoted $Y \sim PE(\boldsymbol{\lambda}, \tau)$ where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_s)$ and $\tau = \{t_1, ..., t_{s-1}\}$. It is not difficult to obtain that

$$p(y \mid \boldsymbol{\lambda}) = \lambda_i^{\kappa_i} \kappa_i \exp\left[-\lambda_i (y - t_{i-1})\right] \quad , y \in I_i \ , \ i = 1, ..., s \tag{3.15}$$

and by direct integration from (3.15)

$$E[Y|\lambda] = \frac{1}{\lambda_1} + \sum_{i=1}^{s-1} \kappa_{i+1} \left[\frac{1}{\lambda_{i+1}} - \frac{1}{\lambda_i} \right] \tag{3.16a}$$

$$E[Y^2|\lambda] = \frac{1}{\lambda_1^2} + \sum_{i=1}^{s-1} \kappa_{i+1} \left[\frac{1}{\lambda_{i+1}^2} - \frac{1}{\lambda_i^2} \right] \tag{3.16b}$$

where

$$\kappa_i = \exp\left[- \sum_{j=1}^{i-1} \lambda_j b_j \right] \quad , i = 1, \ldots, s$$

$$b_i = t_i - t_{i-1} \quad , i = 1, \ldots, s$$

and χ_i is the failure indicator for interval I_i, i=1, ..., s.

The models proposed by Breslow (1972, 1974) use PE distributions for the baseline hazard with τ = {successive death times}. Kalbfleisch & Prentice (1973) also use PE distributions for the baseline hazard without imposing restrictions on τ. The dynamic models for survival analysis assume this for the whole hazard function and not only for baseline. The PE distributions are used in the next section in conjunction with the TFL so that some conditional results are needed. Let B_i be the event $[Y > t_i]$, $i = 0, 1, \ldots, s-1$. Following (3.7),

$$p(y|\lambda, B_{i-1}) = \lambda_l^{\chi_l} \frac{\kappa_l}{\kappa_i} \exp[-\lambda_l (y - t_{l-1})] \quad , y \in I_l \, , \, l \geq i$$

and when $l = i$ this reduces to

$$p(y|\lambda, B_{i-1}) = \lambda_i^{\chi_i} \exp[-\lambda_i (y - t_{i-1})] \quad , y \in I_i \, , \, i = 1, \ldots, s \tag{3.17}$$

This is simply saying that given information previous to I_i, Y behaves in I_i like an exponentially distributed variable with origin shifted to t_{i-1}. The lack of memory of the exponential distribution extends to a piecewise lack of memory of the PE distribution within each interval. Technically, the important point is that all occurrences within I_i, given previous information, only provide likelihood for parameters directly related to that interval.

3.4 - Dynamic Survival Models and Parametric Updating

The material needed to present the dynamic models and their updating procedures have already been fully explained. The model is again presented in two levels: the observational and the system levels with the guide relationship providing the link between them. It consists of

(i) *observation equation* : $Y_j \sim PE(\lambda^{(j)}, \tau)$, $j = 1, \ldots, m$ independent given $\lambda^{(1)}, \ldots, \lambda^{(m)}$ with $\lambda^{(j)} = (\lambda_1^{(j)}, \ldots, \lambda_N^{(j)}, \lambda_{N+1}^{(j)})$ and $\tau = \{t_1, \ldots, t_N\}$.

(ii) *system equation* : $\beta_i = G_i(b_i)\beta_{i-1} + w_i$, $w_i \sim [0, W_i]$ (3.18)

based on a continuous system equation like in section 2.3 with $\beta_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})^T$, $i = 1, 2, \ldots, N, N+1$

(iii) *guide relationships* : $g(\lambda(t, z_j)) \simeq z_j \beta_i$, $t \in I_i$ or

$g(\lambda_i^{(j)}) \simeq z_j \beta_i$, $j = 1, \ldots, m$, $i = 1, 2, \ldots, N, N+1$.

The models are not as in chapter 2 because instead of one observation at any time, one has information from various observations at each time, obtained via the factorisation of the likelihood. The logarithmic link is generally assumed relating $\log \lambda_i^{(j)} \simeq z_j \beta_i$ although any other non-negative function could be used as well. The logarithmic link is chosen for interpretability reasons, the dynamic models being easily understood as extensions of the commonly used static models. The first component of the β's models the effect of the baseline hazard. It is accommodated into the analysis by extending the vector of p covariates z to $(1, z)$. The baseline hazard is no longer assumed unrelated to the regression parameters or to the baseline hazard in other intervals due to the system structure. The regression parameters and the baseline hazard are changing according to a system structure similar to the previously suggested in (3.6). (3.18) assumes a linear structure but the analysis can be extended to non-linear models as outlined in chapter 2. Indeed, an example of non-linearity in survival models is presented in chapter 5 for the analysis of life tables.

The full likelihood from the data can be factored using the TFL version for independent observations (equation (3.13)) with (3.17) and each likelihood factor l_i is

$$\prod_{j=1}^{r_i} \left[\lambda_i^{(j)}\right]^{\chi_{ij}} \exp\left[-\lambda_i^{(j)}(y_{ij}-t_{i-1})\right] \quad , i = 1, 2, \ldots, N, N+1 \tag{3.19}$$

since due to independence assumptions, each D_{i-1} in (3.13) can be reduced to $Y_j > t_{i-1}$ for the corresponding observation. (3.19) has the same form of the likelihood from r_i independent exponential observations with observed value $y_{ij} - t_{i-1}$, parameter $\lambda_i^{(j)}$ and death indicator χ_{ij}, $j = 1, \ldots, r_i$. It can be analysed as in example 1 of chapter 2 with conjugate Gamma prior distributions for the $\lambda_i^{(j)}$'s and a dynamic model for the system parameters. The value of t_N is chosen as a point after which there is no information so that $l_{N+1} = 0$. This keeps bounds to the problem avoiding having to take the system equation far beyond the point of practical interest. Since all $\lambda_i^{(1)}, \ldots, \lambda_i^{(r_i)}$ are related to the same system parameter β_i, the problem is one of a static model within each interval. The dynamics of the model appear when evolving through successive intervals. As pointed out in section 2.2, the dynamic GLM can also be used to analyse static regression. It updates the system using one observation at a time but the updated distribution of the parameter depends on this observation ordering.

Suppose that the r_i observations present in the likelihood for interval I_i are given labels from 1 to r_i and let $D_{i;j}$ be the information set including information previous to I_i and information from first j observations in I_i, $j = 1, \ldots, r_i$, $i = 1, 2, \ldots, N$. Observe that $D_{i-1;r_i} = D_i = D_{i;0}$, $i = 1, 2, \ldots, N$. Following example 1 of chapter 2, the model within I_i for r_i independent observations is:

(i') *observation equation* : $p(y_{ij} \mid \lambda_i^{(j)}, D_{i-1}) = \left[\lambda_i^{(j)}\right]^{\chi_{ij}} \exp\left[-\lambda_i^{(j)}(y_{ij}-t_{i-1})\right]$, $j = 1, \ldots, r_i$

(ii') *system equation* : $\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}_{i,j-1} = \boldsymbol{\beta}_i$, $j = 1, \ldots, r_i$

(iii') *prior distributions* : $[\boldsymbol{\beta}_{ij} \mid D_{i-1;j-1}] \sim [\mathbf{a}_{ij}; \mathbf{R}_{ij}]$ and $[\lambda_i^{(j)} \mid D_{i-1;j-1}] \sim G(\alpha_{ij}, \gamma_{ij})$ with $\alpha_{ij} = q_{ij}^{-1}$, $\gamma_{ij} = q_{ij}^{-1} \exp(-f_{ij})$, $f_{ij} = \mathbf{z}_j \mathbf{a}_{ij}$, $\mathbf{s}_{ij} = \mathbf{R}_{ij} \mathbf{z}_j^T$ and $q_{ij} = \mathbf{z}_j \mathbf{s}_{ij}$.

The prior distributions are based in the logarithmic link and the corresponding operational relations (2.19). The posterior distributions are $[\lambda_i^{(j)} \mid D_{i-1;j}] \sim G(\alpha_{ij} + \chi_{ij}, \gamma_{ij} + y_{ij} - t_{i-1})$ and, following (2.24)-(2.25), $[\boldsymbol{\beta}_{ij} \mid D_{i-1;j}] \sim [\mathbf{m}_{ij}, \mathbf{C}_{ij}]$ with

$$\mathbf{m}_{ij} = \mathbf{a}_{ij} + \frac{\mathbf{s}_{ij}}{q_{ij}} \log \left[\frac{1 + q_{ij}\chi_{ij}}{1 + q_{ij}(y_{ij} - t_{i-1})\exp(f_{ij})} \right] \tag{3.20a}$$

$$\mathbf{C}_{ij} = \mathbf{R}_{ij} - \frac{\chi_{ij}}{1 + q_{ij}} \mathbf{s}_{ij}\mathbf{s}_{ij}^T \tag{3.20b}$$

Since the system is static, $[\boldsymbol{\beta}_{i,j+1} \mid D_{i-1;j}] \sim [\mathbf{a}_{i,j+1}, \mathbf{R}_{i,j+1}]$ with $\mathbf{a}_{i,j+1} = \mathbf{m}_{ij}$ and $\mathbf{R}_{i,j+1} = \mathbf{C}_{ij}$. When $j = r_i$, then $[\boldsymbol{\beta}_{i,r_i} \mid D_{i-1;r_i}] \sim [\boldsymbol{\beta}_i \mid D_i] \sim [\mathbf{m}_i, \mathbf{C}_i]$ where $\mathbf{m}_i = \mathbf{m}_{i,r_i}$ and $\mathbf{C}_i = \mathbf{C}_{i,r_i}$.

Once all the observations in I_i have been processed, one can evolve in time by moving on to the next interval. This is done using the system equation (3.18) so that $[\boldsymbol{\beta}_{i+1} \mid D_i] \sim [\mathbf{a}_{i+1}, \mathbf{R}_{i+1}]$ with

$$\mathbf{a}_{i+1} = \mathbf{G}_{i+1}(b_{i+1})\,\mathbf{m}_i \tag{3.21a}$$

$$\mathbf{R}_{i+1} = \mathbf{G}_{i+1}(b_{i+1})\,\mathbf{C}_i\,\mathbf{G}_{i+1}^T(b_{i+1}) + \mathbf{W}_{i+1} \tag{3.21b}$$

The analysis can then proceed within I_{i+1} as above after initialising with $[\boldsymbol{\beta}_{i+1,1} \mid D_{i;0}] \sim [\mathbf{a}_{i+1,1}, \mathbf{R}_{i+1,1}]$ with $\mathbf{a}_{i+1,1} = \mathbf{a}_{i+1}$ and $\mathbf{R}_{i+1,1} = \mathbf{R}_{i+1}$. This sequential procedure goes on until it reaches I_N when the analysis finishes.

The order given to the observations is specially important since at every interval there are r_i independent observations and this number can be quite large if m is large. The procedure adopted here to minimize this problem is to allocate deaths before survivals, that is, at each stage i, if $\chi_{ij} = 1$ then $j \leq k$ for all k such that $\chi_{ik} = 0$. This ensures that the most informative contributions in the sense of reducing the variances are analysed first. To see that, let $\mathbf{v} = (v_1 \cdots v_{p+1}) \neq 0$ and $\xi = \mathbf{v}\boldsymbol{\beta}_i$. Then from (3.20b), $V[\xi \mid D_{i-1;j}] - V[\xi \mid D_{i-1;j-1}] = \mathbf{v}\mathbf{C}_{ij}\mathbf{v}^T - \mathbf{v}\mathbf{R}_{ij}\mathbf{v}^T = au^2 \geq 0$ where $a = \chi_{ij}/(1+q_{ij})$ and

$u = \mathbf{v}\mathbf{s}_{ij}$. So, the variance of any linear combination of $\boldsymbol{\beta}_i$'s components decreases when $\chi_{ij} = 1$ and remain constant when $\chi_{ij} = 0$. This allocation is shown in the next section to be supported by empirical evidence.

After the updating cycle is completed, one is left with the on-line distributions of $[\boldsymbol{\beta}_i \mid D_i]$, $i = 1, 2, ..., N$. Based on those, a trajectory can be drawn for the parameter and its evolution in time can be visualised. Those distributions are based on the information up to the given time. In survival data, the conclusions are often drawn only after all the data is observed. The inference procedure can then be made more accurate after obtaining the parametric distributions based on all the information collected. This is done with the smoothed distributions for $[\boldsymbol{\beta}_i \mid D_N]$, $i = 1, 2, ..., N$. It is not difficult to see by comparison with the proof of theorem 2.1 and the comments about different time intervals of section 2.3 that $[\boldsymbol{\beta}_i \mid D_N] \sim [\mathbf{m}_i^N, \mathbf{C}_i^N]$, $i = 1, 2, ..., N$ with

$$\mathbf{m}_i^N = \mathbf{m}_i + \mathbf{C}_i \mathbf{G}_{i+1}^T(b_{i+1}) \mathbf{R}_{i+1}^{-1} (\mathbf{m}_{i+1}^N - \mathbf{a}_{i+1}) \tag{3.22a}$$

$$\mathbf{C}_i^N = \mathbf{C}_i - \mathbf{C}_i \mathbf{G}_{i+1}^T(b_{i+1}) \mathbf{R}_{i+1}^{-1} (\mathbf{R}_{i+1} - \mathbf{C}_{i+1}^N) \mathbf{R}_{i+1}^{-1} \mathbf{G}_{i+1}(b_{i+1}) \mathbf{C}_i \tag{3.22b}$$

where $\mathbf{m}_N^N = \mathbf{m}_N$ and $\mathbf{C}_N^N = \mathbf{C}_N$ provide the initialisation. Referring back to the proof of the theorem, the main change is on the distribution of future occurrences, $f(y_{t+1}, ..., Y_N \mid \boldsymbol{\beta}_{t+1}, \boldsymbol{\beta}_t, D_t)$ in the proof, to $p(y_1, ..., y_{r_{i+1}} \mid \boldsymbol{\beta}_{i+1}, \boldsymbol{\beta}_i, D_i)$ in the context of survival models. It can also be proven following exactly the same arguments of (2.44)-(2.48) that

$$p(y_1, ..., y_{r_{i+1}} \mid \boldsymbol{\beta}_{i+1}, \boldsymbol{\beta}_i, D_i) = p(y_1, ..., y_{r_{i+1}} \mid \boldsymbol{\beta}_{i+1}, D_i) \quad , i = 1, ..., N-1$$

so that, given information up to the present and the next state of the system, future occurrences do not depend on the present state of the system. Apart from that and the replacement of $\mathbf{G}_i$ by $\mathbf{G}_i(b_i)$ due to unequal time intervals, the derivation is the same for dynamic GLM and survival data. The smoothed distributions of the parameters should be more accurate reflecting the presence of more information. In general, they do not provide very different a trajectory for $\boldsymbol{\beta}$ from the one with the on-line distributions but a smoother one. Their main importance, however, is in the construction of survival

functions for new individuals in order to make predictive statements.

The models and their analysis were presented here using linear relations for the parameters. There is no problem to extend it by replacing the linear form $\mathbf{z}_j\boldsymbol{\beta}_i$ in (iii) by a general $\mathbf{z}_j(\boldsymbol{\beta}_i)$ or the system transfer function by a non-linear transfer $\mathbf{G}_i(b_i,\boldsymbol{\beta}_i)$ as in section 2.2. The analysis follows the one for the NLM and the smoothed distributions are as in corollary 2.1. A non-linear transfer function can be useful to model exponential decay of the parameters. The dynamic structure of the model permits learning about the values of ρ_i in (3.5) and letting it change with time. The case of a non-linear regression term $\mathbf{z}_j(\boldsymbol{\beta}_i)$ is illustrated in section 5.2 where the adoption of a piecewise logistic hazard precludes any form of linear regression.

The construction of the system evolution in (ii) leads to prior moments given by (3.21). Alternatively, discount ideas could be used to set

$$\mathbf{R}_{i+1} = \mathbf{B}_{i+1}^{-1/2}\,\mathbf{G}_{i+1}(b_{i+1})\,\mathbf{C}_i\,\mathbf{G}_{i+1}^{T}(b_{i+1})\,\mathbf{B}_{i+1}^{-1/2} \tag{3.23}$$

This approach should be taken with great care specially if the interval lengths vary over a wide range of values but can be equally applied if the modeller has more feel for discounts then for additive system disturbances. The smoothed distributions are still the same but for the values of $\mathbf{R}_i$, $i = 1, 2, \ldots, N$.

3.5 - Model Criticism

An important part of the inference procedure is to validate the particular model adopted. This may be approached in Bayesian analysis by comparison of the posterior probabilities of different models, If M_1 and M_2 are two possible models, then, via Bayes' theorem,

$$Pr(M_1|\mathbf{y}) = c\ Pr(M_1)\,l(\mathbf{y}|M_1)$$
$$Pr(M_2|\mathbf{y}) = c\ Pr(M_2)\,l(\mathbf{y}|M_2)$$

for some $c > 0$ so that

$$\frac{Pr(M_1|\mathbf{y})}{Pr(M_2|\mathbf{y})} = \frac{Pr(M_1)}{Pr(M_2)}\,\frac{l(\mathbf{y}|M_1)}{l(\mathbf{y}|M_2)}$$

and if there is prior indifference between the two models

$$\frac{Pr(M_1 \mid \mathbf{y})}{Pr(M_2 \mid \mathbf{y})} = \frac{l(\mathbf{y} \mid M_1)}{l(\mathbf{y} \mid M_2)} \tag{3.24}$$

since $Pr(M_1) = Pr(M_2)$. The RHS of (3.24) is often called the Bayes factor (Good, 1985). $l(\mathbf{y} \mid M_1)$ and $l(\mathbf{y} \mid M_2)$ are the predictive or marginal likelihoods of M_1 and M_2 respectively, after observing $\mathbf{y}$. They are called marginal because they are obtained after integrating out the parameters from the likelihood with prior weights. (3.24) tells us that the odds in favour of M_1 are the ratio of the predictive likelihoods if one has no preference between the two models before observing the data. (3.24) is still approximately true if there is only a slight preference for one of the models. In practice, different models have a ratio of predictive likelihoods that override the prior odds unless these are very extreme. The task here therefore is to obtain the predictive likelihood for any given model. Every model has a number of assumptions regarding selection of covariates, the grid of observation points, the system covariances and so forth and this is denoted hereafter by M. All the distributions until now should in fact have had M explicitly in the conditioning part but it was dropped for simplicity without problems since the model was always the same. In this section, M is to be made explicit in the conditioning part.

The value of $l(\mathbf{y} \mid M)$ is

$$\int l(\mathbf{y} \mid \boldsymbol{\lambda}, M)\, f(\boldsymbol{\lambda} \mid M)\, d\boldsymbol{\lambda}$$

but cannot be obtained this way because the approach adopted does not involve the specification of joint prior distributions. This was deliberately done to avoid tedious numerical integrations that would inevitably follow. Instead, the approach with sequential updating keeps complications to a minimum without jeopardizing the dynamic nature of the process. It can be used also to obtain $l(\mathbf{y} \mid M)$ with the use of the TFL. The observations are no longer independent because the system structure now relates them but the TFL ensures that

$$l(\mathbf{y} \mid M) = \prod_{i=1}^{N} l_i(\mathbf{y}_i \mid D_{i-1}, M) \tag{3.25}$$

From theorem 1.1, each likelihood factor l_i can be written as

$$l_i(\mathbf{y} \mid D_{i-1}, M) = \prod_{j=1}^{r_i} p(y_{ij} \mid y_{i1}, \ldots, y_{ij-1}, D_{i-1}, M)$$

$$= \prod_{j=1}^{r_i} p(y_{ij} \mid D_{i-1;j-1}, M) \tag{3.26}$$

since $D_{i-1;j-1}$ contains D_{i-1} and information from the first $j-1$ individuals in I_i, $i = 1, 2, \ldots, N$. Each factor in (3.26) is

$$p(y_{ij} \mid D_{i-1;j-1}, M) = \int_0^\infty p(y_{ij} \mid \lambda_i^{(j)}, D_{i-1;j-1}, M)\, f(\lambda_i^{(j)} \mid D_{i-1;j-1}, M)\, d\lambda_i^{(j)}$$

$$= \int_0^\infty [\lambda_i^{(j)}]^{\chi_{ij}} \exp[-\lambda_i^{(j)}(y_{ij} - t_{i-1})] \frac{\gamma_{ij}^{\alpha_{ij}}}{\Gamma(\alpha_{ij})} [\lambda_i^{(j)}]^{\alpha_{ij}-1} \exp(-\lambda_i^{(j)}\gamma_{ij})\, d\lambda_i^{(j)}$$

$$= \left[\frac{\alpha_{ij}}{\gamma_{ij} + y_{ij} - t_{i-1}}\right]^{\chi_{ij}} \left[1 + \frac{y_{ij} - t_{i-1}}{\gamma_{ij}}\right]^{-\alpha_{ij}} \tag{3.27}$$

obtained after observing that $p(y_{ij} \mid \lambda_i^{(j)}, D_{i-1;j-1}, M) = p(y_{ij} \mid \lambda_i^{(j)}, D_{i-1}, M)$ and that the distribution of $[\lambda_i^{(j)} \mid D_{i-1;j-1}, M]$ is given in (iii') above. Substitution of (3.27) in (3.26) and then in (3.25) gives

$$l(\mathbf{y} \mid M) = \prod_{i=1}^{N} \left\{ \prod_{j=1}^{r_i} \left[\frac{\alpha_{ij}}{\gamma_{ij} + y_{ij} - t_{i-1}}\right]^{\chi_{ij}} \left[1 + \frac{y_{ij} - t_{i-1}}{\gamma_{ij}}\right]^{-\alpha_{ij}} \right\} \tag{3.28}$$

The predictive likelihood is very easy to obtain along the updating procedure. It takes only a couple of lines in any computer programme and it is advisable to work with its logarithm because it reaches very small values very fast. Its evaluation can be inserted after the construction of the prior for $[\lambda_i^{(j)} \mid D_{i-1;j-1}, M]$ when α_{ij} and γ_{ij} are available along with χ_{ij} and $y_{ij} - t_{i-1}$ from the observation to be processed. The problem of order dependence is still present here and the observations should be given the same ordering they were given in the sequential updating. The allocation of deaths before survivals has improved the likelihood with respect to other allocations in all applications confirming suggestions made at the previous section. This is adopted hereafter as a standard pro-

cedure.

The predictive likelihood of a model is a major tool to be used to criticise a model, relative to others. In the applications, it has given some empirical evidence in favour of particular assumptions. In terms of choice of the grid τ, it seems to favour grids whose times are determined by the observed death times. Also, very fine grids based on the time measurement accuracy give higher likelihoods than other arbitrary grids with longer interval lengths. This reinforces the intuitive feeling that the analysis should be done as close to the continuous as practically possible because the observation process is continuous. A grid based on the measurement unit can sometimes be prohibitive due to the sheer amount of points to be covered and the grid based on the death times can be as good. It is this latter grid that Kaplan & Meier (1958), Cox (1972) and Breslow (1974) adopt for the baseline hazard apparently due to the presence of substantial amount of information (deaths) at these points.

Graphical outputs are the other main form of model criticism. Each parameter has a trajectory in time. By plotting its on-line and smoothed means one has a good idea of the changes taking place and comparison between them gives some insight into the particular model assumptions. Sometimes, assumptions on the system disturbances can dampen completely the smoothed effects of some covariates. A system without disturbances gives constant smoothed trajectories irrespective of the changes indicated by the on-line means. In other cases, the smoothed effects can be grossly overweighted. A compromise suitable to the particular problem can help in chosing the adequate model by visual inspection.

This is about all that can be done to select a model in the light of the given data. Another important test for any model is its ability to predict future occurrences. This is developed in the next section.

3.6 - Prediction of Future Occurrences

After going through the estimation procedure along with smoothing and checking that the model is suitable, one has to consider the implications of the adopted model. It

gives valuable information about the population under study but the difficult test for any model is to provide an account as accurate as possible of what has not yet been observed. This task is called prediction because it is based on the predictive distribution conditioned on all the information collected so far. Its need in survival analysis can be divided in two cases: follow-up predictions and new individuals.

The first case arises when, during an observation process one is interested in predicting the survival time of one of the individuals still known to be alive. This is useful in a controlled clinical trial when the origin of the survival times of different individuals coincides in calendar time. After a certain amount of time, one has sufficient information to make predictions about the remaining survival of the individuals still alive and based on those take decisions like changing treatment of some patients or withdrawing others. The prediction of new individuals is more useful however. Often, after collecting sample data from a population, one is faced with a new individual. Given the covariates, the survival pattern can be established and decisions made. This division is made only for presentation and it may well be the case that a mixture of them is required. In any case, the prediction for both cases is presented here and it is hoped that after that the mixture case remain an obvious exercise. In both cases, factorisation is the key to obtaining sequentially the required predictive distributions.

3.6.1- Follow-up prediction

Suppose the observation for a group of individuals is stopped at a specific point, say t_k, $k = 1, 2, ...$ Let Y_t be the survival time of the individual whose follow-up performance is required. This can happen for many different reasons typically related to the problem under study. An obvious requirement is that this individual is known to be alive at t_k so that it makes sense to think about its survival after t_k. The interval (t_k, ∞) can then be partitioned in M subintervals, $I_{k+j} = (t_{k+j-1}, t_{k+j}]$, $j = 1, ..., M-1$ and $I_{k+M} = (t_{k+M-1}, \infty)$. These intervals are obtained after consideration of points where the modeller feels the system is noticeably changing. In the absence of such information, the

intervals can be set according to multiples of the measurement unit. The number of intervals is governed by this choice and can be any positive integer. The case $M=1$ represents the belief that the system is basically static after t_k.

Since the information that $Y_s > t_k$ is in D_k, the interest lies in obtaining $p(y_s \mid D_k)$ for $y_s > t_k$. Let $D_{k+j}^* = \{D_k, Y_s > t_{k+j}\}$, $j = 0, 1, ..., M-1$ so that $D_k^* = D_k$. From the standpoint of t_k, D_{k+j}^* contains a mixture of actual information from D_k with the virtual information from the events $\{Y_s > t_{k+j}\}$ yet to be realised. The same factorisation of (3.7) leads to

$$p(y_s \mid D_k) = p(y_s \mid D_k, Y_s > t_{k+l-1}) \prod_{j=1}^{l-1} S(t_{k+j} \mid D_k, Y_s > t_{k+j-1})$$

$$= p(y_s \mid D_{k+l-1}^*) \prod_{j=1}^{l-1} S(t_{k+j} \mid D_{k+j-1}^*) \quad , y_s \in I_{k+l} \tag{3.29}$$

Supposing that $[\lambda_{k+j}^{(s)} \mid D_{k+j-1}^*] \sim G(\alpha_{k+j}; \gamma_{k+j})$, each factor in (3.29) is obtained as

$$p(y_s \mid D_k) = \int_0^\infty p(y_s \mid \lambda_{k+j}^{(s)}, D_{k+j-1}^*) f(\lambda_{k+j}^{(s)} \mid D_{k+j-1}^*) \, d\lambda_{k+j}^{(s)}$$

$$= \int_0^\infty [\lambda_{k+j}^{(s)}]^{\chi_{k+j,s}} \exp[-\lambda_{k+j}^{(s)}(y_s - t_{k+j-1})] f(\lambda_{k+j}^{(s)} \mid D_{k+j-1}^*) \, d\lambda_{k+j}^{(s)}$$

$$= \left[\frac{\alpha_{k+j}}{\gamma_{k+j} + y_s - t_{k+j-1}}\right]^{\chi_{k+j,s}} \left[1 + \frac{y_s - t_{k+j-1}}{\gamma_{k+j}}\right]^{-\alpha_{k+j}} , \; y_s \in I_{k+j} \; , \; j = 1, ..., M$$

where $\chi_{k+j,s}$ is the death indicator for Y_s in interval $k+j$. Substitution in (3.29) gives

$$p(y_s \mid D_k) = \left[\frac{\alpha_{k+l}}{\gamma_{k+l} + y_s - t_{k+l-1}}\right]^{\chi_{k+l,s}} \left[1 + \frac{y_s - t_{k+l-1}}{\gamma_{k+l}}\right]^{-\alpha_{k+l}} \prod_{j=1}^{l-1} \left[1 + \frac{b_{k+j}}{\gamma_{k+j}}\right]^{-\alpha_{k+j}} \tag{3.30}$$

for $y_s \in I_{k+l}$, $l = 1, ..., M$. The distribution for $[\lambda_{k+j}^{(s)} \mid D_{k+j-1}^*]$, $j = 1, ..., M$ can be obtained sequentially as follows:

(a) set $j=1$;

(b) from $[\boldsymbol{\beta}_{k+j-1} \mid D_{k+j-1}^*] \sim [\mathbf{m}_{k+j-1}, \mathbf{C}_{k+j-1}]$, one can construct $[\boldsymbol{\beta}_{k+j} \mid D_{k+j-1}^*] \sim [\mathbf{a}_{k+j}, \mathbf{R}_{k+j}]$ with the system equation as before and with the

use of covariates $\mathbf{z}_*$ obtain $f_{k+j} = \mathbf{z}_* \mathbf{a}_{k+j}$, $\mathbf{s}_{k+j} = \mathbf{z}_* \mathbf{R}_{k+j}$ and $q_{k+j} = \mathbf{s}_{k+j} \mathbf{z}_*^T$. The guide relationship then gives $[\lambda_{k+j}^{(*)} \mid D_{k+j-1}^*] \sim G(\alpha_{k+j}, \gamma_{k+j})$ with $\alpha_{k+j} = q_{k+j}^{-1}$ and $\gamma_{k+j} = q_{k+j}^{-1} \exp(-f_{k+j})$;

(c) using the likelihood from $[Y_* > t_{k+j} \mid D_{k+j-1}^*]$, update to the posterior distributions $[\lambda_{k+j}^{(*)} \mid D_{k+j}^*] \sim G(\alpha_{k+j}, \gamma_{k+j} + b_{k+j})$ and $[\beta_{k+j} \mid D_{k+j}^*] \sim [\mathbf{m}_{k+j}, \mathbf{C}_{k+j}$ with $\mathbf{m}_{k+j} = \mathbf{a}_{k+j} - \{\log[1 - q_{k+j} b_{k+j} \exp(f_{k+j})]\} \mathbf{s}_{k+j} / q_{k+j}$ and $\mathbf{C}_{k+j} = \mathbf{R}_{k+j}$;

(d) if $j < M$ add 1 to j and return to (b), otherwise the cycle is finished.

The distribution of $[Y_* \mid D_k]$ is then completely specified and predictive statements can be made. It is useful to plot the predictive survival $S(y_* \mid D_k)$ to observe graphically the expected behaviour of Y_*, and to note summaries such as the median, the quartiles and others. The mean, obtained using (3.30), is

$$E[Y_* \mid D_k] = \sum_{j=1}^{M-1} \left\{ \left[\prod_{l=1}^{j-1} c_{k+l} \right] \left[t_{k+j-1} + \frac{\gamma_{k+j}}{\alpha_{k+j} - 1} - c_{k+j} \left(t_{k+j-1} + \frac{\gamma_{k+j} + b_{k+j} \alpha_{k+j}}{\alpha_{k+j} - 1} \right) \right] \right\}$$

$$+ \left[\prod_{l=1}^{M-1} c_{k+l} \right] \left(t_{k+M-1} + \frac{\gamma_{k+M-1}}{\alpha_{k+M-1} - 1} \right)$$

where

$$c_{k+l} = \left(1 + \frac{b_{k+l}}{\gamma_{k+l}} \right)^{-\alpha_{k+l}}, \quad l = 1, \ldots, M-1$$

A simple way of deriving this result is shown in the next subsection for the more general case of new observations.

3.6.2- Prediction of new observations

In many ways, the ability to predict outcomes related to new observations is the real test of a model. It can always be set to fit a given set of observations to any accuracy but this gives no idea about how useful the model can be outside this data. Due to the constant presence of censoring, it is difficult to compare predictions of different models for

observations that end up being censored but for those who die, one can resort to devices like predictive survival or summaries like mean survival time or median survival time. One drawback of the present approach is that predictions can only be made individually. The sequential updating meant that only one of the λ's was used at any one time thus avoiding the specification of a joint predictive distribution for different λ's. Although this was useful for the computations, it now prevents the specification of a joint predictive distribution for a group of individuals. The predictions here are therefore the marginal predictions obtained independently for individuals one at a time.

Suppose that after observing the data, the smoothed distributions for $[\boldsymbol{\beta}_i \mid D_N]$, $i = 1, 2, \ldots, N$ are available and a new individual with survival time Y_* and covariates $\mathbf{z}_*$ appears. This can be either a real individual whose prediction is of importance for decisions about its treatments and so forth or an hypothetical exercise where one is mainly concerned with a possible situation determined by the covariates and wants to see its prospective performance. In either case, the quantity of interest is clearly $p(y_* \mid D_N)$. This again can be factorised for $y_* \in I_i$, $i = 1, 2, \ldots, N, N+1$ to

$$p(y_* \mid D_N) = p(y_* \mid D_N, Y_* > t_{i-1}) \prod_{j=1}^{i-1} S(t_j \mid D_N, Y_* > t_{j-1}) \tag{3.31}$$

Supposing that $[\lambda_j^{(*)} \mid D_N, Y_* > t_{j-1}] \sim G(\alpha_{j,*}, \gamma_{j,*})$ as in the previous section, each factor in (3.31) is

$$\begin{aligned} p(y_* \mid D_N, Y_* > t_{j-1}) &= \int_0^\infty p(y_* \mid \lambda_j^{(*)}, D_N, Y_* > t_{j-1})\, f(\lambda_j^{(*)} \mid D_N, Y_* > t_{j-1})\, d\lambda_j^{(*)} \\ &= \int_0^\infty [\lambda_j^{(*)}]^{\chi_{j,*}} \exp[-\lambda_j^{(*)}(y_* - t_{j-1})]\, f(\lambda_j^{(*)} \mid D_N, Y_* > t_{j-1})\, d\lambda_j^{(*)} \\ &= \left[\frac{\alpha_{j,*}}{\gamma_{j,*} + y_* - t_{j-1}}\right]^{\chi_{j,*}} \left[1 + \frac{y_* - t_{j-1}}{\gamma_{j,*}}\right]^{-\alpha_{j,*}}, \quad y_* \in I_j,\ j = 1, \ldots, N, N+1 \end{aligned}$$

where $\chi_{j,*}$ is the death indicator for Y_* in I_j. Finally, substitution in (3.31) gives for $y_* \in I_i$, $i = 1, 2, \ldots, N, N+1$

$$p(y_s|D_N) = \left[\frac{\alpha_{i,s}}{\gamma_{i,s}+y_s-t_{i-1}}\right]^{\lambda_{i,s}} \left[1+\frac{y_s-t_{i-1}}{\gamma_{i,s}}\right]^{-(\alpha_{i,s}+\lambda_{i,s})} \prod_{j=1}^{i-1}\left[1+\frac{b_j}{\gamma_{j,s}}\right]^{-(\alpha_{j,s}+\lambda_{j,s})} \quad (3.32)$$

The difference here is that the distributions for $[\lambda_i^{(s)} | D_N, Y_s > t_{i-1}]$, $i = 1,2,...,N,N+1$ are not as easy to obtain as the ones in 3.6.1 due to the presence of $[Y_s > t_{i-1}]$ in the conditioning part. The problem is to obtain the distributions for $[\beta_i | D_N, Y_s > t_{i-1}]$, $i = 1,2,...,N,N+1$ because from them and the guide relation the required distributions can be obtained. This is done for $i = 1,2,...,N$ as follows: use of Bayes' theorem at time t_i gives

$$f(\beta_i,\beta_{i+1} | Y_s > t_i, D_N) \propto p(\mathbf{y}_{i+1}, y_{i,s} | \beta_i,\beta_{i+1}, D_N)\, f(\beta_i,\beta_{i+1} | D_i) \quad (3.33)$$

where D_N is divided with respect to t_i as $\{D_i, \mathbf{y}_{i+1}\}$ and $y_{i,s} = min(y_s, t_i) = t_i$. Given the system parameters, the observations are independent implying that

$$p(\mathbf{y}_{i+1}, y_{i,s} | \beta_i,\beta_{i+1}, D_i) = p(\mathbf{y}_{i+1} | \beta_i,\beta_{i+1}, D_i)\, Pr(Y_s > t_i | \beta_i,\beta_{i+1}, D_i)$$

$$= p(\mathbf{y}_{i+1} | \beta_{i+1}, D_i)\, S_s(t_i | \beta_i,\beta_{i+1}, D_i)$$

The RHS of (3.33) can be written as

$$[p(\mathbf{y}_{i+1} | \beta_{i+1}, D_i) f(\beta_{i+1} | D_i)]\,[S_s(t_i | \beta_i,\beta_{i+1}, D_i) f(\beta_i | \beta_{i+1}, D_i)]$$

which is proportional to

$$f(\beta_{i+1} | D_i, \mathbf{y}_{i+1})\, f(\beta_i | \beta_{i+1}, D_i, Y_s > t_i) \quad (3.34)$$

by use of Bayes' theorem within each pair of brackets. Since (3.34) is a product of densities, it can be equated to the LHS of (3.33) and rejoining $D_N = \{D_i, \mathbf{y}_{i+1}\}$ gives

$$f(\beta_i,\beta_{i+1} | Y_s > t_i, D_N) = f(\beta_{i+1} | D_N)\, f(\beta_i | \beta_{i+1}, D_i, Y_s > t_i)$$

So

$$E[\beta_i | D_N, Y_s > t_i] = E[E[\beta_i | \beta_{i+1}, D_i, Y_s > t_i] | D_N] \quad (3.35a)$$

$$V[\beta_i | D_N, Y_s > t_i] = E[V[\beta_i | \beta_{i+1}, D_i, Y_s > t_i] | D_N] + V[E[\beta_i | \beta_{i+1}, D_i, Y_s > t_i] | D_N] \quad (3.35b)$$

Suppose now that the data is reanalysed with an extra observation with covariates $\mathbf{z}_s$ and death time larger than t_N. This observation plays the role of Y_s and is the last to be analysed at every interval. It only gives survival contributions to the intervals and therefore do not change previous variances. This leads to the joint distribution

$$\left[\begin{pmatrix} \beta_i \\ \beta_{i+1} \end{pmatrix} | D_i, Y_s > t_i \right] \sim \left[\begin{pmatrix} \mathbf{m}_i^{(s)} \\ \mathbf{a}_{i+1}^{(s)} \end{pmatrix}, \begin{pmatrix} \mathbf{C}_i & \mathbf{C}_i \mathbf{G}_{i+1}^T(b_{i+1}) \\ \mathbf{G}_{i+1}(b_{i+1})\mathbf{C}_i & \mathbf{R}_{i+1} \end{pmatrix} \right] \quad (3.36)$$

where the superscripts (s) are present in the means $\mathbf{m}_i^{(s)}$ and $\mathbf{a}_{i+1}^{(s)} = \mathbf{G}_{i+1}(b_{i+1})\mathbf{m}_i^{(s)}$ to emphasise their difference to the means from the standard updating of section 3.4 but the variances remain unchanged throughout. Linear Bayesian methods can be applied to (3.36) leading to

$$E[\beta_i | \beta_{i+1}, D_i, Y_s > t_i] = \mathbf{m}_i^{(s)} + \mathbf{C}_i \mathbf{G}_{i+1}^T(b_{i+1}) \mathbf{R}_{i+1}^{-1} (\beta_{i+1} - \mathbf{a}_{i+1}^{(s)})$$

$$V[\beta_i | \beta_{i+1}, D_i, Y_s > t_i] = \mathbf{C}_i - \mathbf{C}_i \mathbf{G}_{i+1}^T(b_{i+1}) \mathbf{R}_{i+1}^{-1} \mathbf{G}_{i+1}(b_{i+1}) \mathbf{C}_i$$

which can be replaced in (3.35) to

$$E[\beta_i | D_N, Y_s > t_i] = \mathbf{m}_i^{(s)} + \mathbf{C}_i \mathbf{G}_{i+1}^T(b_{i+1}) \mathbf{R}_{i+1}^{-1} (\mathbf{m}_{i+1}^N - \mathbf{a}_{i+1}^{(s)})$$

$$V[\beta_i | D_N, Y_s > t_i] = \mathbf{C}_i - \mathbf{C}_i \mathbf{G}_{i+1}^T(b_{i+1}) \mathbf{R}_{i+1}^{-1} (\mathbf{R}_{i+1} - \mathbf{C}_{i+1}^N) \mathbf{R}_{i+1}^{-1} \mathbf{G}_{i+1}(b_{i+1}) \mathbf{C}_i$$

Defining $\mathbf{e}_i = \mathbf{m}_i^{(s)} - \mathbf{m}_i$ as the correction on the mean due to the presence of the events $[Y_s > t_i]$, the expression for the mean can be rewritten as

$$E[\beta_i | D_N, Y_s > t_i] = \mathbf{m}_i + \mathbf{C}_i \mathbf{G}_{i+1}^T(b_{i+1}) \mathbf{R}_{i+1}^{-1} [\mathbf{m}_{i+1}^N - \mathbf{a}_{i+1}] + \mathbf{e}_i^{(s)} - \mathbf{C}_i \mathbf{G}_{i+1}^T(b_{i+1}) \mathbf{R}_{i+1}^{-1} \mathbf{G}_{i+1}(b_{i+1}) \mathbf{e}_i^{(s)}$$

$$= \mathbf{m}_i^N + [\mathbf{I} - \mathbf{C}_i \mathbf{G}_{i+1}^T(b_{i+1}) \mathbf{R}_{i+1}^{-1} \mathbf{G}_{i+1}(b_{i+1})] \mathbf{e}_i^{(s)}$$

and $V[\beta_i | D_N, Y_s > t_i] = \mathbf{C}_i^N$. Equations (3.20) is used again in updating

$$E[\beta_i | D_N, Y_s > t_i] = E[\beta_i | D_N, Y_s > t_{i-1}] - \frac{\mathbf{s}_i}{q_i} \log [1 + q_i b_i exp(f_i)] \quad (3.37a)$$

$$V[\beta_i | D_N, Y_s > t_i] = V[\beta_i | D_N, Y_s > t_{i-1}] = \mathbf{C}_i^N \quad (3.37b)$$

where $f_i = \mathbf{z}_s E[\beta_i | D_N, Y_s > t_{i-1}]$, $\mathbf{s}_i = \mathbf{C}_i^N \mathbf{z}_s^T$ and $q_i = \mathbf{z}_s \mathbf{s}_i$. So, the variance of the distribution of $[\beta_i | D_N, Y_s > t_{i-1}]$ is already available but the mean can only be obtained from (3.37a) by numerical methods. The fact that this has to be done for every interval along with the need to reanalyse the data all over makes this an unrealistic task from a practical implementation point of view. In general, D_N contains a great deal of information and is reasonable to approximate $E[\beta_i | D_N, Y_s > t_{i-1}]$ by $\mathbf{m}_i^N = E[\beta_i | D_N]$. All this implies is is that, in the presence of D_N, the events $[Y_s > t_{i-1}]$ are negligible in expectation. If this approximation is adopted, writing $f_i = \mathbf{z}_s \mathbf{m}_i^N$ and $q_i = \mathbf{z}_s \mathbf{C}_i^N \mathbf{z}_s^T$ gives

$$\begin{cases} \alpha_{i,s} = q_i^{-1} \\ \gamma_{i,s} = q_i^{-1} \exp(-f_i) \end{cases}, \quad i = 1, 2, \ldots, N$$

The parametric distributions for the last interval are difficult to obtain because the system evolution is not specified. If a transition matrix $\mathbf{G}_{N+1}$ and a variance matrix $\mathbf{W}_{N+1}$ can be specified then $[\boldsymbol{\beta}_{N+1} \mid D_N] \sim [\mathbf{a}_{N+1}, \mathbf{R}_{N+1}]$ with $\mathbf{a}_{N+1} = \mathbf{G}_{N+1}\mathbf{m}_N$ and $\mathbf{R}_{N+1} = \mathbf{G}_{N+1}\mathbf{C}_N\mathbf{G}_{N+1}^T - W_{N+1}$ as usual. In general, this is irrelevant because there will be little probability left after t_N.

After (3.32) is completely specified, the survival function $S(y_s \mid D_N)$ can be drawn and the median obtained if necessary. The best way to do this is to localize visually the interval that contains the median and then solve analytically $S(m \mid D_N) = .5$ for m, or simply draw a vertical line at .5 and read off the corresponding coordinate. The mean can also be obtained as

$$E[Y_s \mid D_N] = \int_0^{\infty} y\, f(y \mid D_N)\, dy = \sum_{i=1}^{N+1} \int_{t_{i-1}}^{t_i} y\, f(y \mid D_N)\, dy$$

From (3.32),

$$E[Y_s \mid D_N] = \sum_{i=1}^{N+1} \left[\prod_{j=1}^{i-1} c_j \right] \int_{t_{i-1}}^{t_i} \frac{y\, \alpha_{i,s} \gamma_{i,s}^{\alpha_{i,s}}}{[\gamma_{i,s} + y - t_{i-1}]^{\alpha_{i,s}+1}}\, dy$$

$$= \sum_{i=1}^{N} \left\{ \left[\prod_{j=1}^{i-1} c_j \right] \left[t_{i-1} + \frac{\gamma_{i,s}}{\alpha_{i,s} - 1} - c_i \left(t_{i-1} + \frac{\gamma_{i,s} + \alpha_{i,s} b_i}{\alpha_{i,s} - 1} \right) \right] \right\}$$

$$+ \left[\prod_{j=1}^{N} c_j \right] \left(t_N + \frac{\gamma_{N+1,s}}{\alpha_{N+1,s} - 1} \right) \tag{3.38}$$

provided $\alpha_{i,s} > 1$, $i = 1, 2, \ldots, N$ and with

$$c_i = \left(1 + \frac{b_i}{\gamma_{i,s}} \right)^{-\alpha_{i,s}}, \quad i = 1, 2, \ldots, N$$

If $\alpha_{i,s} \leq 1$, the integrals are improper and the mean is indeterminate. If a vague prior is adopted, this could well be the case if the system has large disturbances. Otherwise the

mean can be obtained if the system equation is extended to I_{N+1} although the median is far easier to obtain and can be used as a summary for location. The predictive density (3.32) is piecewise continuous but the survival function is always continuous in $\mathbf{R}^+$.

References

Anderson, J. A. and Senthilselvan, A. (1982). A two-step regression model for hazard functions. *Appl. Statist.*, **31**, 44-51.

Breslow, N. (1972). Comment on Cox's paper. *J. R. Statist. Soc., Ser. B,* **34**, 216-217.

-----(1974). Covariance analysis of censored survival data. *Biometrics,* **30**, 89-99.

Carter, W. H., Wampler, G. L. and Stablein, D. M. (1983). *Regression Analysis of Survival Data in Cancer Chemotherapy.* New York: Marcel Dekker.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc., Ser. B,* **34**, 187-220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

Cozzolino, M. (1974). Conjugate distributions for incomplete observations. *J. Amer. Statist. Ass.,* **69**, 264-266.

Elandt-Johnson, R. C. and Johnson, N. L. (1980). *Survival Models and Data Analysis.* New York: Wiley.

Good, I. J. (1985). Weight of evidence: a brief survey. In *Bayesian Statistics 2* (J. M. Bernardo et al., eds.). Amsterdam: North-Holland, 249-270.

Gore. S. M., Pocock, S. J. and Kerr, G. R. (1984). Regression models and non-proportional hazards in the analysis of breast cancer survival. *Appl. Statist.,* **33**, 176-195.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika,* **60**, 267-278.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Ass.,* **53**, 457-481.

Shooman, M. L. (1968). *Probabilistic Reliability.* New York: McGraw-Hill.

Stablein, D. M., Carter, W. H. and Novak, J. W. (1981). Analysis of survival data with non-proportional hazard functions. *Cont. Clin. Trials,* **2**, 149-159.

CHAPTER FOUR

NUMERICAL APPLICATIONS

In this chapter, the models proposed in the previous chapter are used to analyse various sets of data. These all come from real applications and are used here to highlight different aspects of the models. The first data set comes from a cancer treatment study and is analysed with the simplest type of model as a two-sample problem. Despite its simplicity, it illustrates how a simple dynamic model can cope with changes in the difference between the two treatments and how a dynamic system equation can be useful in describing such changes. Next a case of time-dependent covariates -as opposed to time-dependent effects- is studied. Another application from medicine, this time heart transplantation. is used to illustrate this point. Finally, an application to the study of personal unemployment spells in Britain is presented. This application in economics not only provides a more comprehensive analysis but also stresses the usefulness of the models in areas other than medicine. In presenting it, the use of switching models where the relevant covariates at different times are different is briefly introduced. This is another feature exclusive to dynamic models.

In all the applications, unless otherwise stated, analysis is assumed to start from $[\beta_0 \mid D_0] \sim [0, 10^3 \mathbf{I}_{p+1}]$ where p is the number of covariates of the model. This represents the prior of a modeller with vague information about the value of the parameter at the beginning of times.

While analysing the data, a number of interesting theoretical problems like modelling a specific differential system equation and extending the models to cope with covariates changing with time arise. They are discussed and their solutions obtained.

4.1 - The Two-sample Problem

The data set analysed here was originally presented in Stablein, Carter & Novak

(1981) and is displayed in Table 1. It consists of the survival times of 90 patients equally divided in two treatment groups: chemotherapy and combined treatment of chemotherapy and radiation. It is not clear from that reference whether the allocation of patients to treatment was random or whether the origin in time was the same feature for all individuals, such as the first time the disease was diagnosed, but both assumptions are made here. These are important to ensure that there was no treatment bias from the onset and no bias in the measurement of the survival times of the patients.

TABLE 1

Survival times (days) of gastric cancer patients

(Stablein, Carter & Novak, 1981)

Chemotherapy	17, 42, 44, 48, 60, 72, 74, 95, 103, 108, 122, 144, 167, 170, 183, 185, 193,
and *Radiation*	195, 197, 208, 234, 235, 254, 307, 315, 401, 445, 464, 484, 528, 542, 567, 577, 580, 795, 855, 1174*, 1214, 1232*, 1366, 1455*, 1585*, 1622*, 1626*, 1736*
Chemotherapy	1, 63, 105, 125, 182, 216, 250, 262, 301, 301, 342, 354, 356, 358, 380, 383, 383, 388, 394, 408, 460, 489, 499, 523, 524, 535, 562, 569, 675, 676, 748, 778, 786, 797, 955, 968, 977, 1245, 1271, 1420, 1460*, 1516*, 1551, 1690*, 1694

* - censored.

A quick look at Table 1 indicates that at the beginning there are more deaths in the combined treatment group but towards the end this does not seem to be the case. The evidence is far from pointing to one treatment as being consistently better than the other one in terms of smaller hazard. This is confirmed in the above reference by obtaining a non-significant value for the constant treatment difference using Cox's model. The trouble is that (the parameter modelling) the treatment difference is actually changing with time. Aware of that, Stablein, Carter & Novak (1981) tried to fit a model with three covariates:

the treatment indicator z_1 (1, for the combined treatment and 0, for chemotherapy alone), a linear interaction term $z_1 t^*$ and a quadratic interaction term $z_1 t^{*2}$ where $t^* = t/30$. Using Cox's model again, the third covariate is shown to be non significant and later, Carter, Wampler and Stablein (1983) reanalysed the data with only the first two covariates. In the notation of this thesis, this is saying that β_1, the treatment difference parameter, can be written as $\beta_1(t) = \theta_1 + \theta_2 t^*$. Their analysis with Cox's model gives the estimates $\hat{\theta}_1 = 1.2711$ and $\hat{\theta}_2 = -.0794$.

This data set therefore seems particularly suited to analysis using a dynamic model. The first group of models used have system equation

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}_{i-1} + \mathbf{w}_i, \ \mathbf{w}_i \sim [\mathbf{0}, \mathbf{W}_i], \ \mathbf{W}_i = b_i \boldsymbol{\sigma}_i \boldsymbol{\sigma}_i^T \tag{4.1}$$

where $\boldsymbol{\beta}_i^T = (\beta_{i,0}, \beta_{i,1})$, $\boldsymbol{\sigma}_i^T = (\sigma_{i,0}, \sigma_{i,1})$ and $\mathbf{w}_i^T = (w_{i1}, w_{i2})$. $\boldsymbol{\beta}$'s first component models the logarithm of the baseline hazard and the second one the difference between the two treatments. So $\mathbf{z} = (1, z_1)$ with z_1 as before. Also the allocation of deaths before survivals is adopted as standard. The values of the marginal likelihood for some combinations of $\boldsymbol{\sigma}_i$ and T are shown in Table 2.

TABLE 2

Logarithm of the predictive likelihood

$\boldsymbol{\sigma}_i^T$	Interval division	
	{20,40,60,...,200,250,300,...,600,700,800,...,1800}	{observed death times}
(0, 0)	-607.15	-607.79
(0, .04)	-603.12	-603.60
(.05, 0)	-606.76	-608.45

The models in the first line assume a static system for the parameters, that is, exponential baseline distribution and proportional hazards. The second line assumes a dynamic difference between the treatments which is clearly supported by the data. The third line can be viewed as a test for non-exponentiality when compared with line 1 and it does not

seem to provide any improvement in the fit. Comparison between columns should have given the relative merits of different choices of the grid τ but the two chosen ones seem to be practically equivalent. A comparison for a wider choice of grids τ of observation points when σ_1^T (0, 0.04), is therefore displayed in Table 3.

TABLE 3

Logarithm of the marginal likelihood for different grids

# τ	notation	points on the grid	*log l*
30	τ_{30}	20, 40, 60,..., 200, 250, 300,..., 600, 700, 800,..., 1800	-603.12
60	τ_{60}	10, 20, 30,..., 200, 225, 250,..., 600, 650, 700,..., 1800	-627.75
75	τ_{75}	15, 30, 45,..., 750, 780, 810,..., 1200, 1250, 1300,..., 1800	-640.44
100	τ_{100}	10, 20, 30,..., 700, 725, 750,..., 1200, 1250, 1300,..., 1800	-627.71
250	τ_{250}	4, 8, 12,..., 700, 710, 720,..., 1200, 1220, 1240,..., 1700	-624.53
500	τ_{500}	2, 4, 6,..., 700, 705, 710,..., 1200, 1210, 1220,..., 1700	-615.83
1700	τ_{1700}	1, 2, 3,..., 1700	-615.83
75	τ_{ODT}	observed death times	-603.60

Apart from the grid τ_{ODT}, the grids have increasingly worse fit as the number of observation points increases up to a point where they begin to be more and more supported by the data. The observation points were selected to have roughly the same density observed in the data. It is surprising however that the worse fit is obtained at the same number of points of the grid of observed death times which gives about the best fit of the data. The lower limit of # τ = 30 was chosen because it does not make much sense to push it further

down if one is interested in learning about the dynamics of the process. The upper limit was determined by the unit of measurement.

The on-line means of β in different times form trajectories depicted in Figure 1 for the grids τ_{75}, τ_{100}, τ_{500} and τ_{1700}. It is clear from Figure 1.a that the improvement on the fit of the model is related to a more positive mean for β_1 in early stages of the process. Figure 2 shows the trajectory of the on-line and smoothed means for the grid of the observed death times and confirms this point. With this model, the predictive survival function for individuals under each of the two treatments is shown in Figure 3. This is also a good visual test of the significance of the difference between the two treatments for the proximity of the two curves gives an idea of the usefulness or otherwise of the inclusion of the parameter in the analysis. Extreme care should be exercised in summarising survival functions like the ones in Figure 3. In this case, the medians are roughly 320 for the combined treatment and 440 for chemotherapy but are giving a very partial account of the changing pattern observed in the different survivals.

The assumption that $\sigma_{i,j} = 0$ for all i and for a fixed j implies that the model assumed β_j to be constant throughout time. As a consequence the smoothed means are all equal to $m_{N,j} = E[\beta_i \mid D_N]$ and the smoothed variances are all equal to $C_{N,jj} = V[\beta_i \mid D_N]$. This can also be checked from equation (3.22). The fact that $\sigma_{i,0} = 0$ can be criticised on the grounds that the smoothed mean's trajectory for β_0 is failing to acknowledge the changes observed during the on-line estimation for the early stages of the process. The mean trajectories for β in Figure 4 are obtained after changing the value of $\sigma_{i,0}$ to .1. Although there is a closer agreement between the smoothed and on-line mean trajectories for β_0, the smoothed mean trajectory for β_1 is completely dampened out. Also, as was previously suggested in Table 1, this model has a log likelihood value of -613.25. The consequences of this model in terms of predictions are depicted in Figure 5 where, as expected from Figure 4, the survival for the two treatments are undistinguishable. The same predictive effect is observed when using Cox's model with z_1 as the only covariate. Cox's model estimates the baseline hazard independently within each interval which is

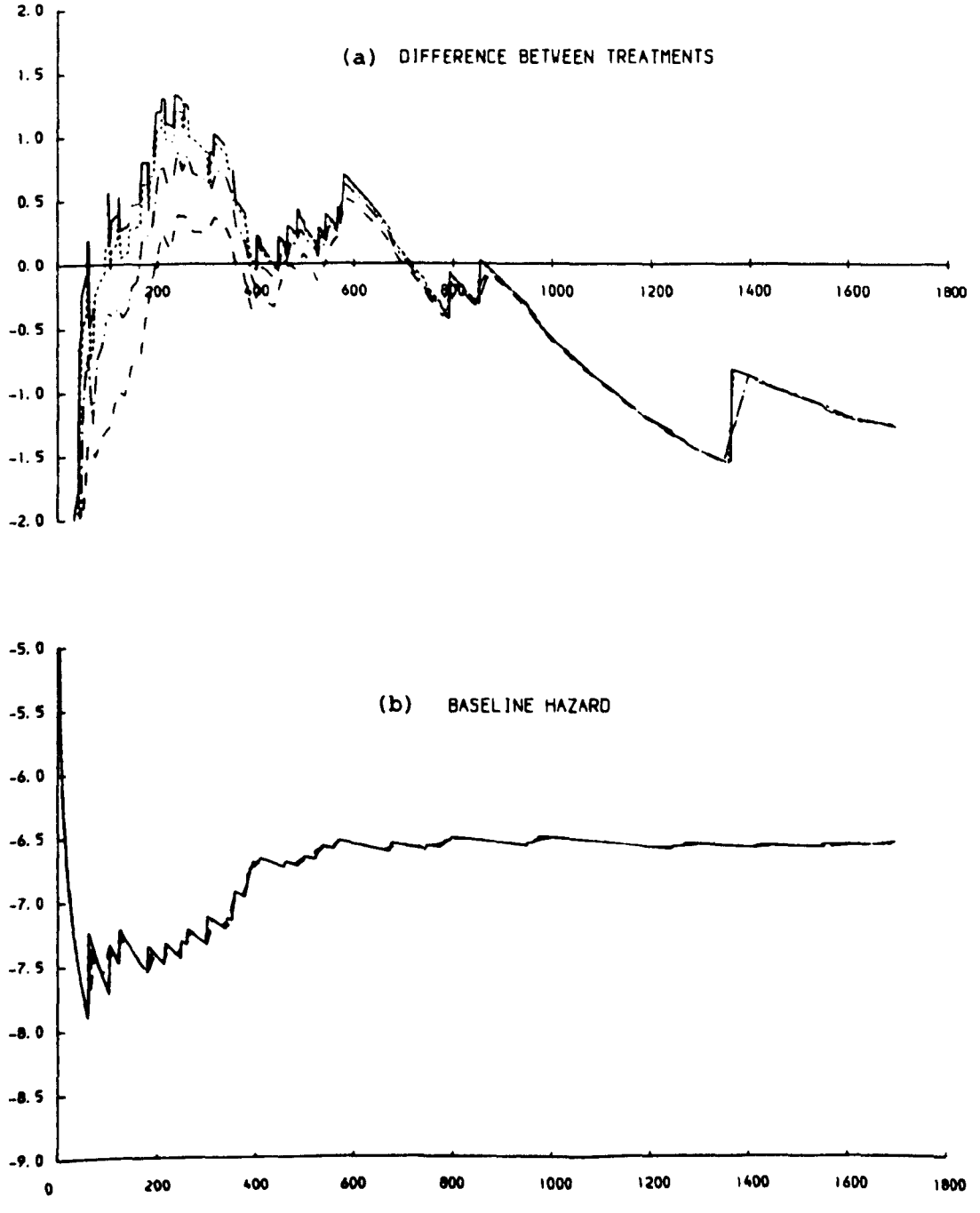


Figure 1. On-line mean trajectories for different grids (———, # τ = 1700; ·····, # τ = 500; -·-·-, # τ = 100; ----, # τ = 75)

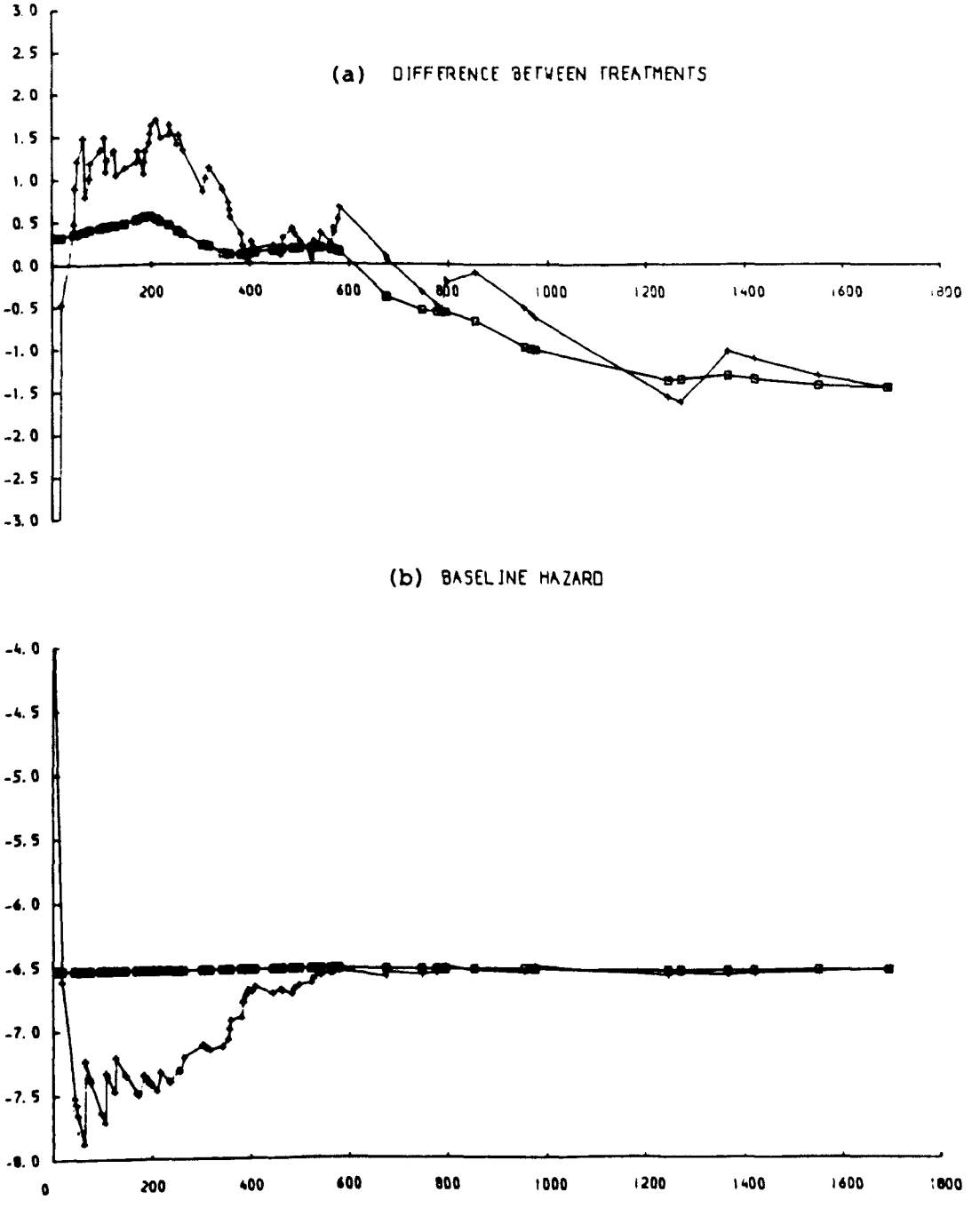


Figure 2. Mean trajectories for Υ_{ODT} (□, smoothed; +, on-line)

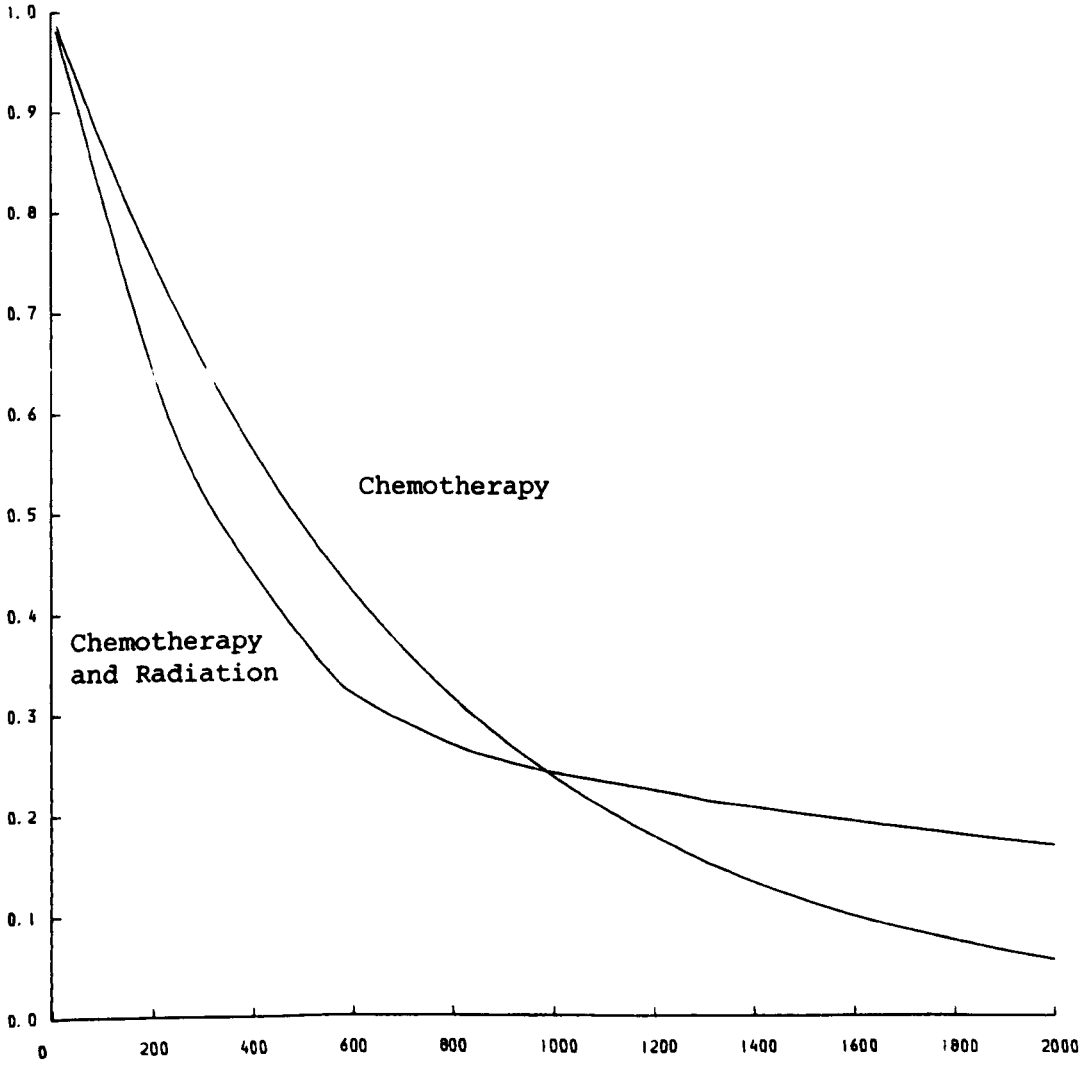


Figure 3. Predictive survival functions for τ_{ODT}

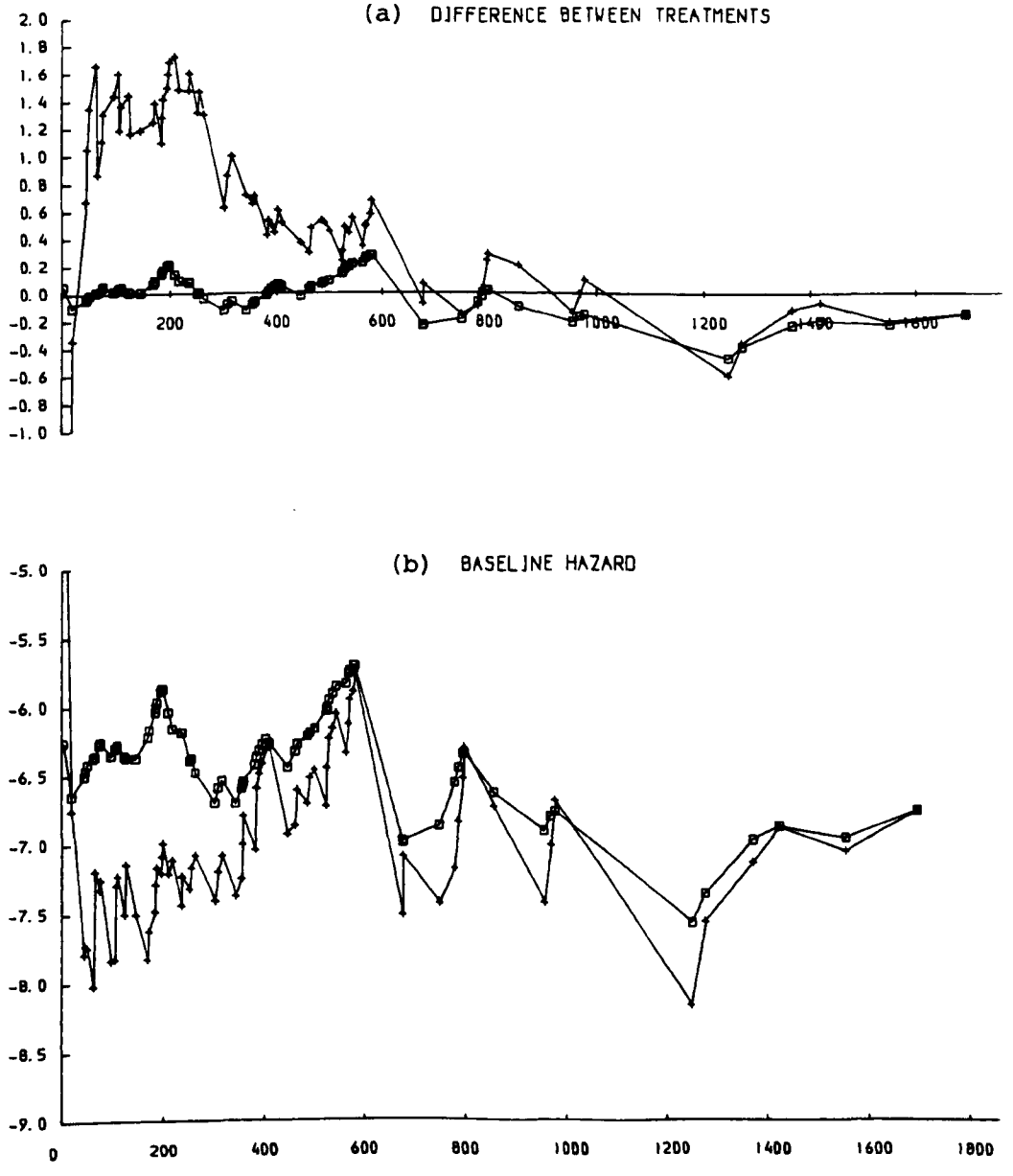


Figure 4. Mean trajectories for γ_{ODT} and $\sigma_{i,0}$ = .1 (□, smoothed; +, on-line)

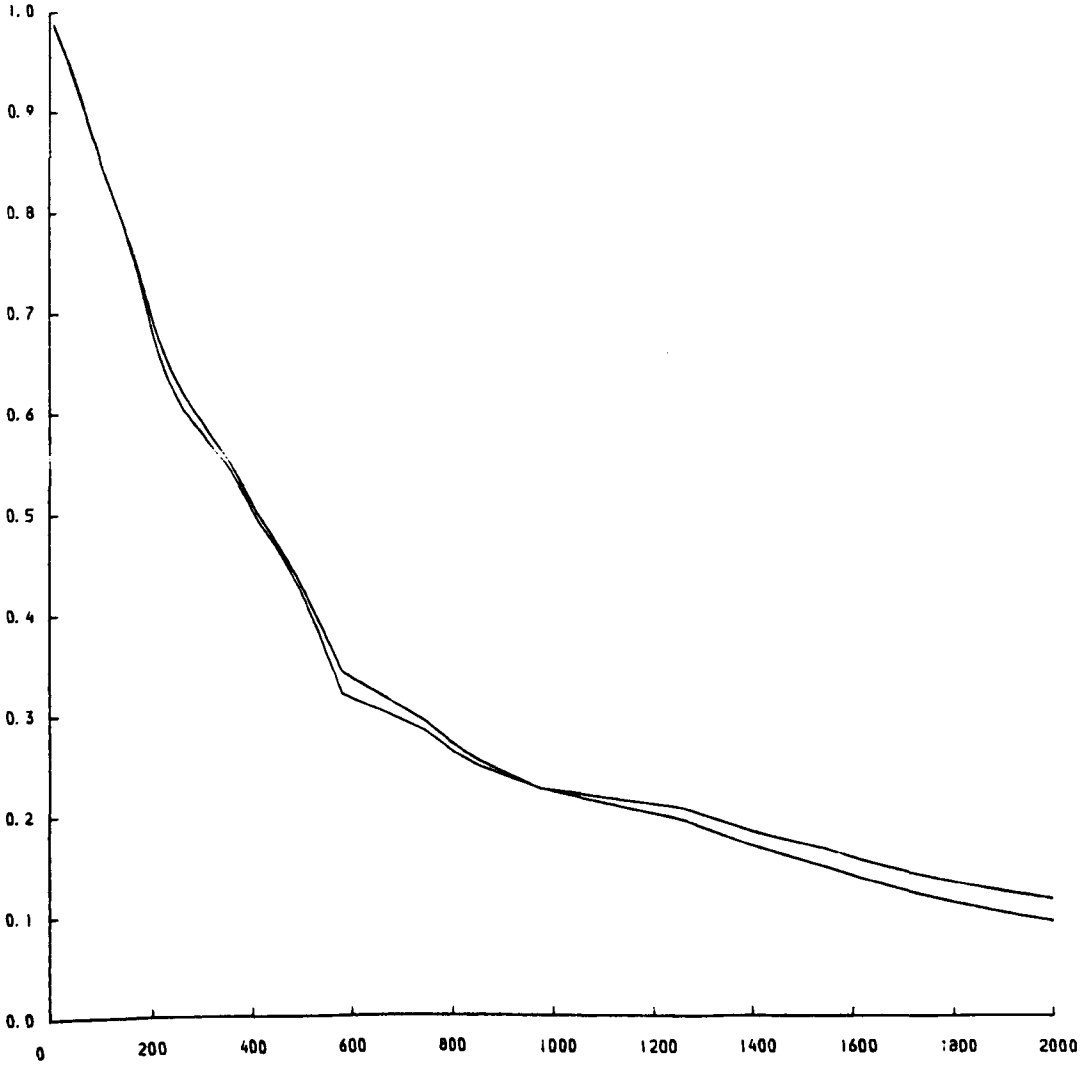


Figure 5. Predictive survival functions for γ_{ODT} and $\sigma_{i,0} = .1$

equivalent to $\sigma_{i,0} \to \infty$ (more about that in the next chapter). So, the evidence is that the further $\sigma_{i,0}$ goes away from 0 the worse is the data support for the model.

One possibility suggested by the amount of changes experienced by β_1 is to introduce a further parameter β_2 representing the slope of β_1. This is done with the system equation

$$\boldsymbol{\beta}_i = \mathbf{G}_i(b_i)\boldsymbol{\beta}_{i-1} + \mathbf{w}_i \,, \quad \mathbf{w}_i \sim (\mathbf{0}, \mathbf{W}_i) \tag{4.2}$$

with

$$\mathbf{G}_i(b_i) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & b_i^* \\ 0 & 0 & 1 \end{pmatrix}, \; b_i^* = \frac{b_i}{30} \text{ and } \mathbf{w}_i = \begin{pmatrix} w_{i1} \\ w_{i2} \\ w_{i3} \end{pmatrix}$$

and as before $\mathbf{W}_i$ is taken as $b_i \boldsymbol{\sigma}_i \boldsymbol{\sigma}_i^T$ with $\boldsymbol{\sigma}_i^T = (\sigma_{i,0}, \sigma_{i,1}, \sigma_{i,2})$. A similar evolution equation was also used in Smith et al. (1983). The derivation of (4.2) involves the specification of a differential system equation for $\boldsymbol{\beta}$ like the ones in section 2.3. The modelling procedures and calculations leading to (4.3) are an useful exercise and are presented in Appendix 4.1. These models are in line with the ones used by Carter et al. (1983) but for two major differences. The first is that the change in the effect of the difference between the treatments is specified at the system level as opposed to the observational level. This makes it easier to understand the model by decomposition of its effects. Also, the models allow for the slope of the difference between the treatments to change with time if there is indication for it. Observe that $\mathbf{G}_i$ is no longer an identity matrix since its element (2,3) is b_i^* and $\mathbf{z}$ is now $(1, z_1, 0)$. The values of the marginal likelihood for some choices of T and $\boldsymbol{\sigma}_i$ are displayed in Table 4. The striking fact as compared to the previous models is the strong preference manifested for the grid of observed death times. There is still support for $\sigma_{i,0} = 0$ but the data support for a dynamic β_1 is no longer present. All those analysis were again based on the allocation of deaths before survivals. A number of alternative allocations were tried giving very similar estimates and slightly smaller likelihood values when compared to the standard allocations. For instance, with $\boldsymbol{\sigma}_i^T = (0.6 \; .1 \; 0)$ and T_{30} the data was analysed in the order they were give in Table 1 and in the reversed order from Table 1 and the respective values of the log likelihood were -642.28 and -644.38.

TABLE 4

Logarithm of the predictive likelihood for the linear growth model

σ_i	Grid of observation points	
	τ_{30}	τ_{ODT}
(0 0 0)	-665.30	-611.56
(.06 0 0)	-653.59	-613.08
(.06 .1 0)	-641.90	-617.88
(0 .04 0)	-	-611.81

Figure 6 shows the mean trajectories for β with $\sigma_{i,0} = 0$, $\sigma_{i,1} = .04$ and $\sigma_{i,2} = .02$ for $t < 401$ and $\sigma_{i,2} = .01$ for $t > 401$ and with the grid τ_{ODT}. Although it attains a lower value of -614.29 for the log likelihood than the model with static slope (see Table 4), it more closely represents the changes undergone by the slope observed in Figure 2. The error variance is increased in the early stages to cope with this highly changing period. The trajectory of the baseline hazard and the predictive survival for the two treatment groups are very similar to the ones obtained in the model with two parameters and $\sigma_i^T = (0\ 0.04)$. A couple of points can be discussed taking this linear growth model as the basis for comparison.

The effect of an informative prior can be appreciated from Figure 7. It gives the mean trajectory for β with the same model but for the prior

$$[\beta_0 \mid D_0] \sim \left[\begin{pmatrix} -7 \\ 1 \\ .1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & .316 & 0 \\ 0 & 0 & .1 \end{pmatrix} \right]$$

The fit of the model increases to -595.47 but from $t = 200$ onwards, the mean trajectories for β are undistinguishable. Also, the test for a non-static baseline hazard is made with the same model but for $\sigma_{i,0} = .1$ as before. Again, this model has less support from the data ($\log l = -617.85$). It can be seen from Figure 8 that the smoothed mean for β_1 is no longer dampened out but the smoothed and on-line mean trajectories for the slope β_2 show a sharp contrast in the early stages (t from 0 to 300). Also in Figure 8 is the mean

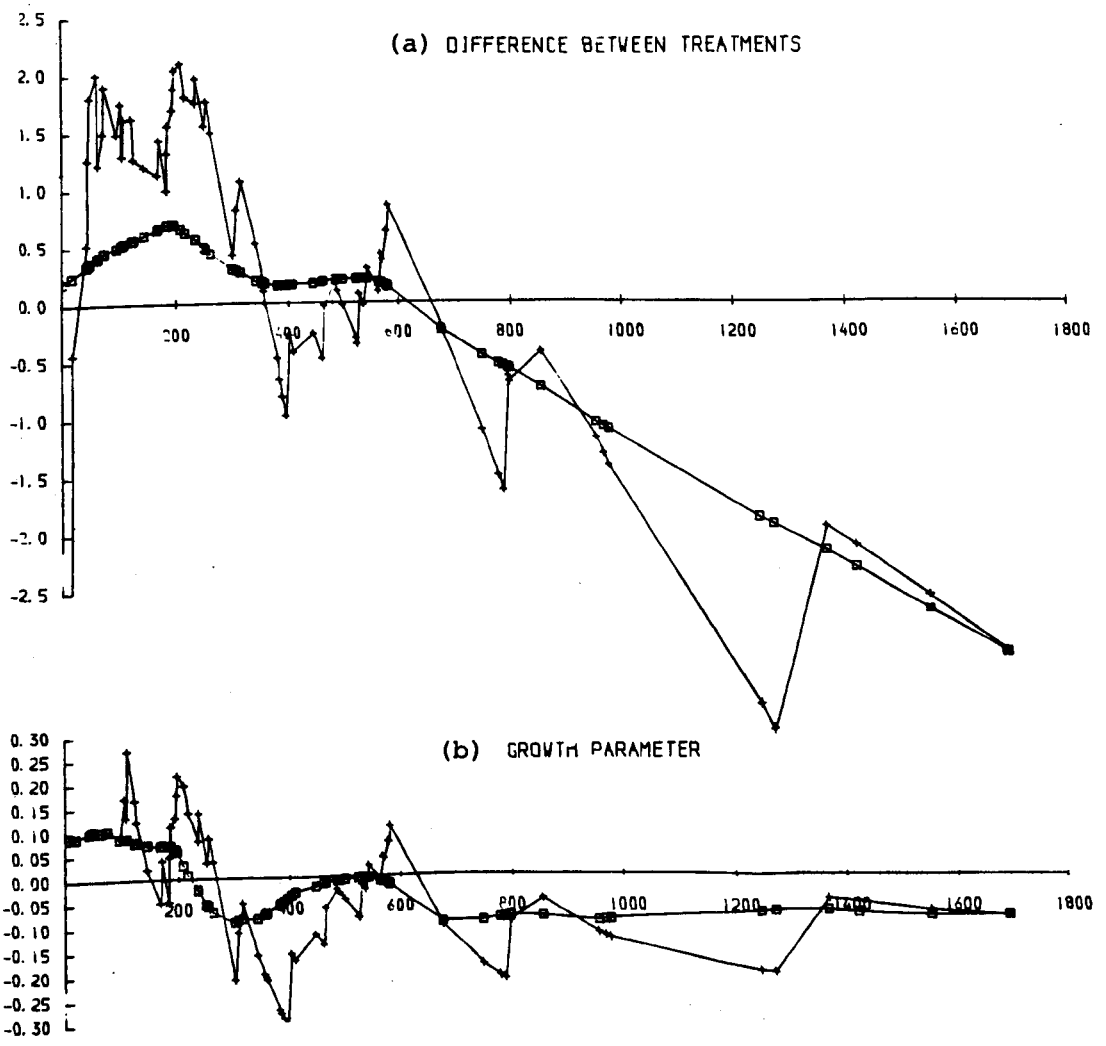


Figure 6. Mean trajectories for linear growth model (□, smoothed; +, on-line)

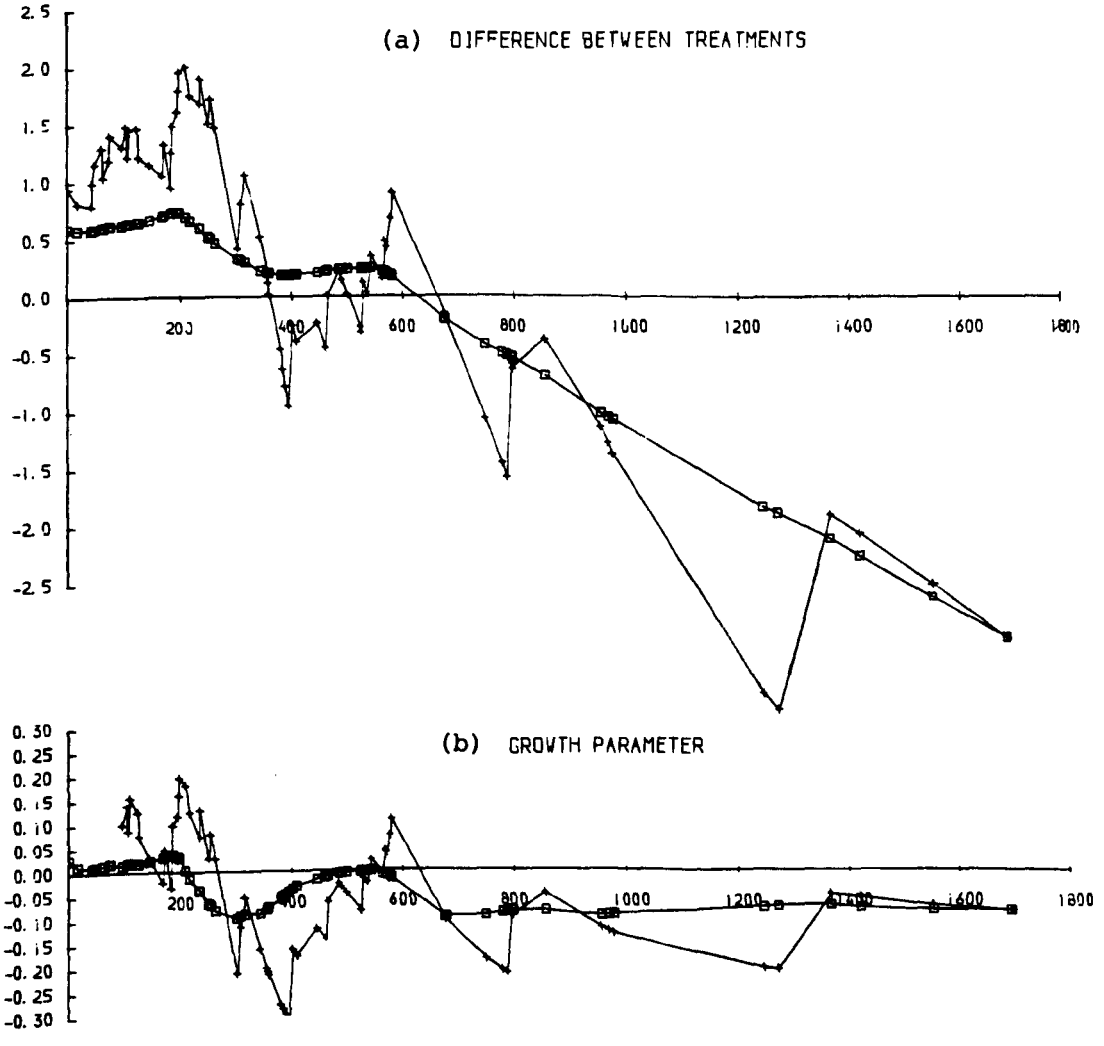


Figure 7. Mean trajectories with an informative prior (□, smoothed; +, on-line)

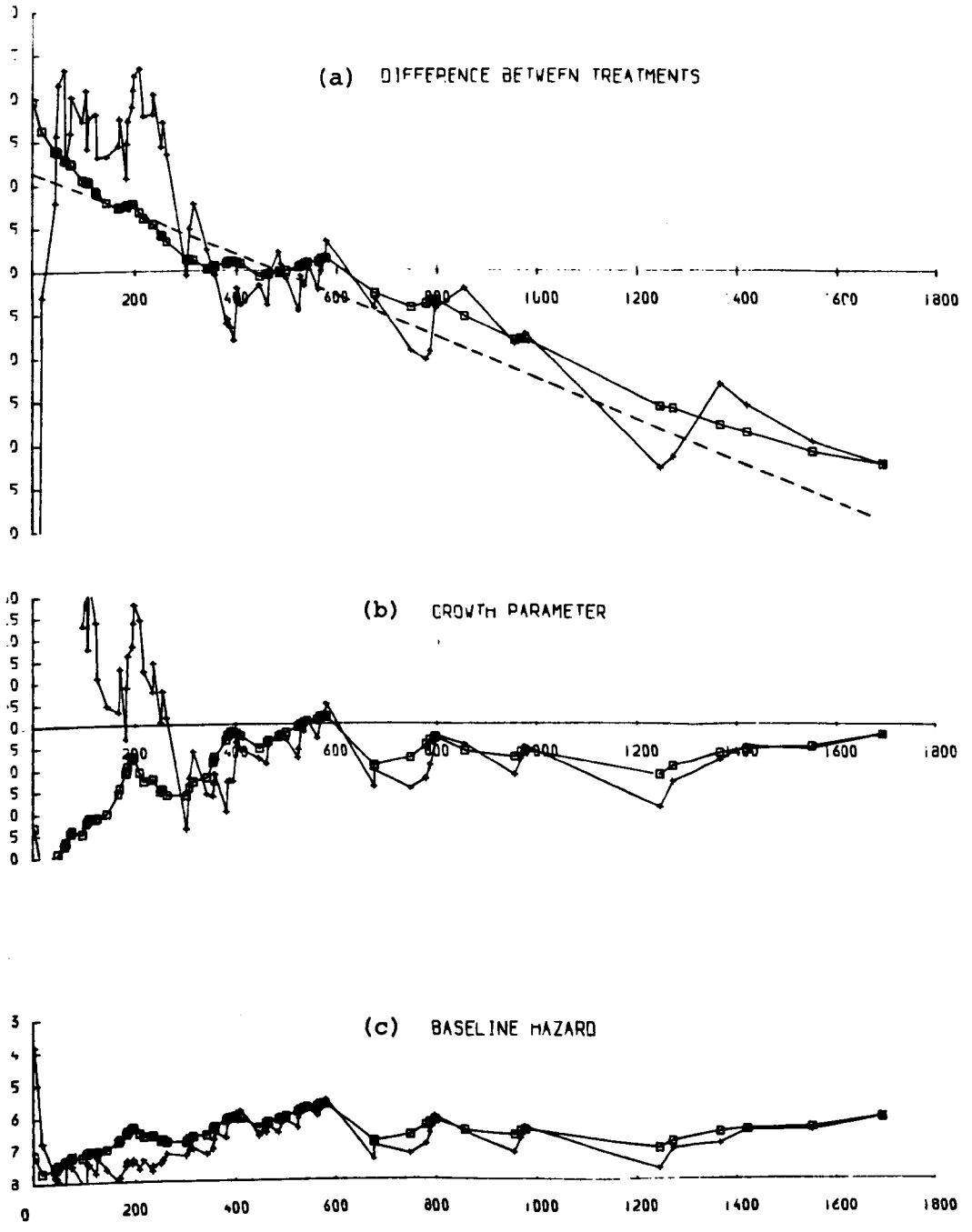


Figure 8. Mean trajectories for $\sigma_{i,0}$ = .1 (■, smoothed; +, on-line)

trajectory for β_1 as estimated by Carter et al. (1983). It seems to agree with the smoothed trajectory for β_1 because the assumption of a constant growth but this is seriously questioned by data evidence specially for $t < 600$ as can be appreciated in Figure 6.

Finally, it is interesting to observe that the model with an extra parameter can be useful in identification and description of the process but carries the risk of giving a worse fit of the data. This is clearly seen after comparing the data support for the two sets of models entertained and reminds that parsimony is a criteria that should always be considered. Comparison of models with parameters of different dimensionalities is however particularly difficult with vague prior distributions (see Smith & Spiegelhalter, 1980). One should not therefore rule out the linear growth model based on the analyses of this section. These are performed here in an exploratory exercise. In practice, consultation with the medical staff should lead to the specification of an informative prior for β.

4.2 - Time-dependent Covariates

The problem of time-dependent covariates does not bring any novelty to the models but needs some special attention. In this section, the case of covariates that are piecewisely constant is considered. This is what typically occurs in practice when some factors change their positions after specific trigger events take place. An application to heart transplantation data is used to illustrate this point later in this section. Although the changes considered here allow for analysis with covariates changing their values at many points in time, they tend to change very sparsely, only at major occurrences. On the other hand, the case of a time-continuous covariate can be analysed as that of a covariate that keeps changing at the time measurement units.

The solution for this is not in fact very difficult. Suppose that an individual with survival time Y_j is observed either to die or to be censored at y_{ij} in interval $I_i = (t_{i-1}, t_i]$, $i = 1, 2, \ldots, N$. There is no restriction to the value of y_{ij} within I_i and therefore it can be t_i possibly representing a survival through I_i. In short, consider a general individual contribution to l_i. Suppose now that at y_{ij_1}, Y_j has its covariates changing

from $\mathbf{z}_{j1}$ to $\mathbf{z}_{j2}$ where $t_{i-1} < y_{i,j_1} < y_{ij}$. The contribution from Y_j to l_i is therefore

$$\exp\{-\lambda_i^{(j_1)}(y_{i,j_1} - t_{i-1})\} \cdot \left[\lambda_i^{(j_2)}\right]^{\gamma_{ij}} \exp\{-\lambda_i^{(j_2)}(y_{ij} - y_{i,j_1})\} \tag{4.3}$$

with the guide relations

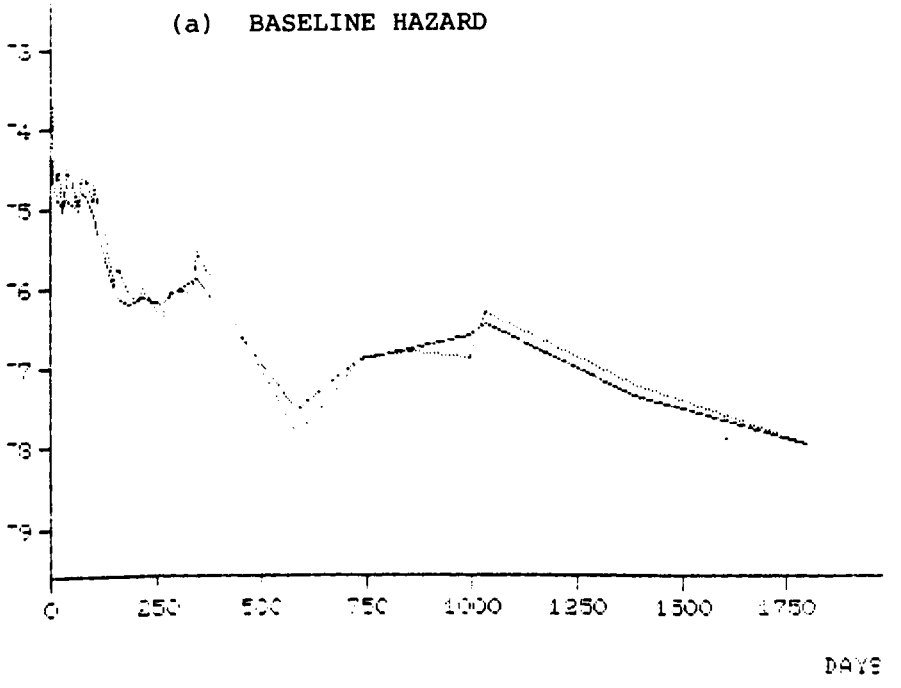
$$\log \lambda_i^{(j_1)} \simeq \mathbf{z}_{j1}\boldsymbol{\beta}_i$$
$$\log \lambda_i^{(j_2)} \simeq \mathbf{z}_{j2}\boldsymbol{\beta}_i$$

There are, in fact, two contribution from Y_j; the first reporting on the interval (t_{i-1}, y_{i,j_1}) and the latter from (y_{i,j_1}, y_{ij}). It is straightforward to generalise for the case that the covariates for Y_j change from $\mathbf{z}_{j1}$ to $\mathbf{z}_{j2}$ up to $\mathbf{z}_{jk}$ at $y_{i,j_1}, y_{i,j_2}, \ldots, y_{i,j_k}$. Sometimes, it is advantageous to have the times the covariates change included in the grid $\mathcal{T}$ of observation points. In cohort studies such as presented in the next section, the covariates change at the same time points for all individuals. Inclusion of these points in $\mathcal{T}$ helps with the computation but also emphasises their importance and can be an aid in determining possible change points of the process. Another alternative is to think of the points in $\mathcal{T}$ as the actual observation points over the intervals. The information about each individual is therefore the last recorded one including covariates. As one is only pretending to have $\mathcal{T}$ as the set of observation points and is really continuously observing the process, this procedure is not using the whole information available. Nevertheless it is useful in applications where the covariates for each individual change at different times as in the application that follows.

The next data set comes from the Stanford Heart Transplantation Programme and was initially presented by Crowley & Hu (1977). It consists of data on 103 individuals joining the programme since commencement up to the 1st of April of 1974, covering a period of about six years. The data consists of the waiting time to transplant since acceptance for the programme, survival time after transplant for those who had it and a number of covariates. The main feature is that the total survival time is the sum of the waiting time and the survival time after transplant and some covariates are bound to

change at time of transplant. This data set not only serves as a good application for the material of this section but, as shown below, provides a comparison between classical analysis and static Bayesian analysis. A number of different models have been analysed by Crowley & Hu using the discrete baseline hazard of Breslow's models. For illustration purposes, I shall concentrate on the analysis with the 99 patients with no missing covariates, using models with the following covariates: z_1= transplant indicator, z_2= previous open-heart surgery indicator and z_3= age at transplant. z_1 and z_3 are 0 until there is a transplant, changing then and remaining constant thereafter. τ is set with the observed death times, the allocation of deaths first is adopted and it is also assumed that $\mathbf{W}_i = diag\,(b_i\sigma_0^2, 0, 0, 0), \forall i$. This assumption enables comparison with the (static) classical analysis since it forces the regression parameters to be constant. This can be seen by substitution into the system equation (3.18) and makes the smoothed distributions for the regression parameters all the same, for all i. The only stochastic input to the system is the constant diffusion disturbance for the baseline hazard. This model with a number of different values of σ_0^2 leads to the results summarised in Table 5 along with the results from classical analysis.

The classical analysis estimates the baseline hazard parameter independently within each interval. This corresponds to $\sigma_0^2 \to \infty$ as discussed in section 3.1. It can be seen from Table 5 that the larger σ_0^2, the closer are the posterior means of regression parameters to the MLE estimates, although with smaller precision. The reasons for this are unclear but it should be pointed out that the data is strongly supporting smaller values of σ_0^2 with more precise estimates. The mean trajectories of the parameters for $\sigma_0^2 = .1$ are shown in Figure 9. It shows that a non-exponential ($\sigma_0^2 \neq 0$) baseline hazard assumption is appropriate with indications of a decreasing hazard. So survival of the patients in the early stages of the programme is crucial and once through it, they stand a good chance of a relatively lengthy survival. The figures also indicate that there is a great deal of uncertainty with respect to the regression parameters in the early stages of the process (the first year or so) with stability of their estimates being achieved well after that. This shows that the dynamic



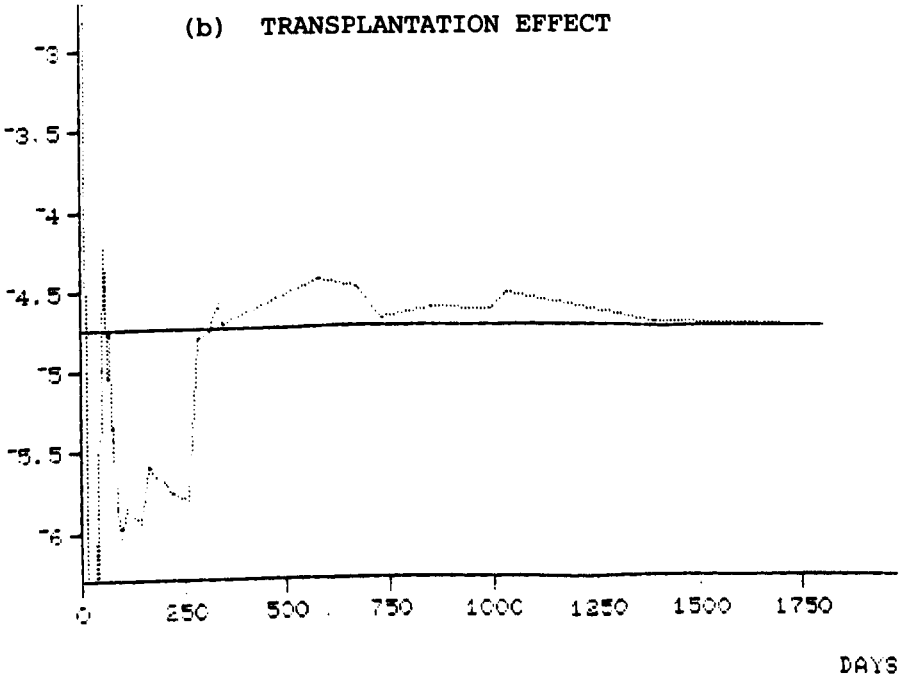
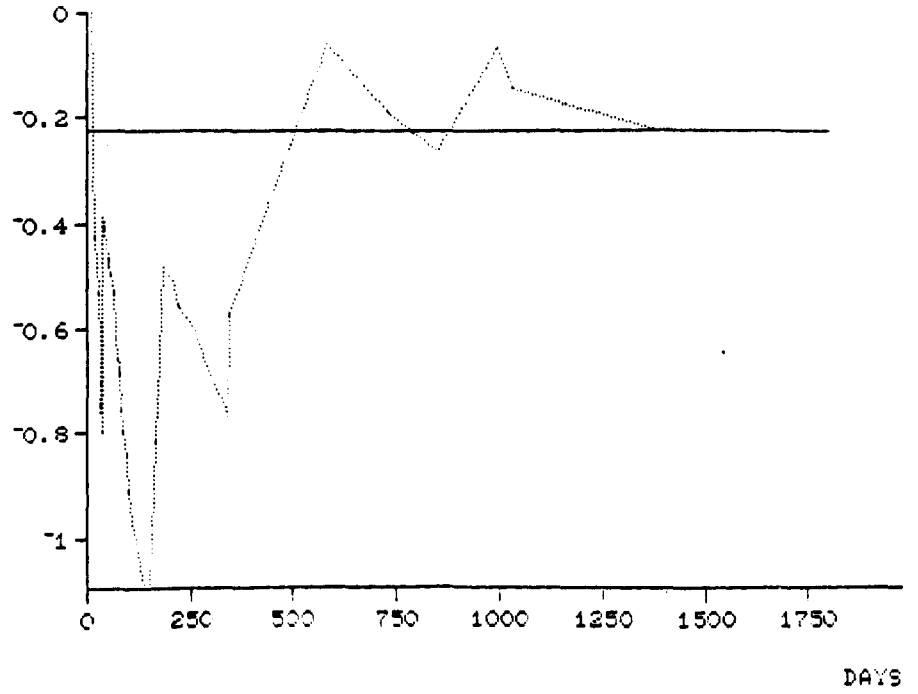


Figure 9(a,b). Mean trajectories for heart transplant data (———, smoothed; ---, on-line)

(c) PREVIOUS SURGERY EFFECT



(d) EFFECT OF THE AGE AT TRANSPLANT

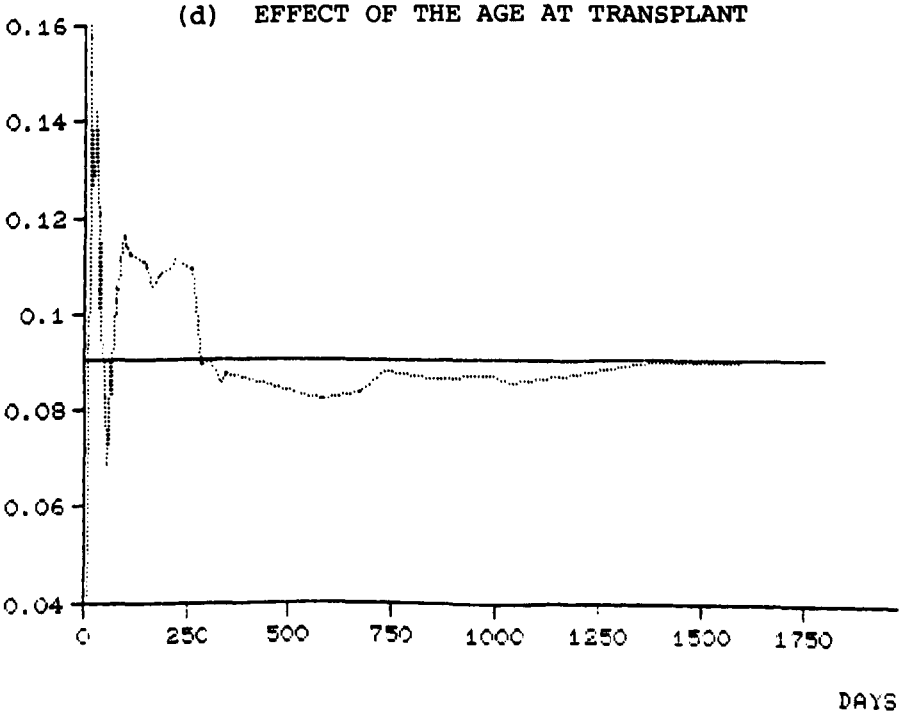


Figure 9(c,d). Mean trajectories for heart transplant data (——, smoothed; ---, on-line)

approach can be useful even within the static environment pointing out movements in the parametric estimation procedure that would have otherwise gone unnoticed.

TABLE 5

Parameter estimation for heart transplant data

σ_0^2	*log l*	β_1	β_2	β_3
.01	-499.02	-3.941	-.324	.0769
		(1.059)	(.419)	(.0214)
.1	-514.32	-3.704	-.239	.0727
		(1.185)	(.519)	(.0242)
1	-553.48	-2.693	-.415	.0574
		(1.455)	(.731)	(.0230)
4	-595.69	-2.481	-.651	.0565
		(1.764)	(.919)	(.0359)
10	-632.73	-2.545	-.801	.0599
		(2.014)	(1.042)	(.0398)
100	-746.8	-2.639	-.969	.0675
		(2.608)	(1.196)	(.0451)
Crowley & Hu's	-	-2.652	-.903	.0550
Analysis		(1.119)	(.402)	(.0219)

(standard deviations in brackets)

4.3 - An econometric application

Although the emphasis in this thesis is towards medical applications, there is no denying of the usefulness of the models in other areas of science. Problems in demography and reliability are dealt with in the next chapters. Another important area is economics where the so called duration times- as opposed to survival times- until certain economic events occur are of interest. The attractiveness of these models in economics

stems mainly from the typically dynamic nature of economic events and their importance to our lives is beyond doubt.

In this section, the duration times of unemployment spells of British men are studied. The data came from the DHSS Cohort Study of the Unemployed containing data on 2332 men registering for unemployment benefit at the autumn of 1978. Along with the duration of their unemployed spell, there is a mass of information concerning each individual. These include a number of factors already available plus the material collected during the 3 interviews that they were subjected to, amounting to around 800 variables for each individual. The information is collected for the time span of 1 year so that the only censored individuals are those remaining unemployed after that.

The study of human behaviour whilst unemployed is important on its own right but, among the many questions that can be asked, particular interest is placed on the effect of the unemployed benefit on the duration of the spell (Nickell, 1979). There seems to be general agreement of the deterrent effect of the benefit and the question is posed more in quantitative rather than qualitative terms. This is very important because a strong effect of the benefit can lead to a government policy of reduction of benefits as a controlling factor of the unemployed level.

In this study, however, I shall concentrate in analysing the effect of different covariates and describe the way in which their changes take place. Due to the very nature of economics, one is quite willing to accept that effects are bound to change with time and there is not much controversy on use of dynamic models there.

Narendranathan et al (1985) studied this data set and concluded, among other things, that out of the immense number of covariates only 20 or so are statistically important. From those, the most significant ones were selected for the present analysis. They are:

z_1 = age of individual (in years) when registering unemployed

z_2 = marriage indicator (1 if married, 0 otherwise)

z_3 = education indicator (1 for at least CSE, 0 otherwise)

z_4 = staying power indicator (1 if less than 12 months in last full-time job, 0 otherwise)

z_5 = previous unemployment indicator (1 if registered in last 12 months, 0 otherwise)

z_6 = local unemployment rate (in %)

z_7 = log (estimated income at work)

z_8 = log (income whilst unemployed)

z_7, before logging, is taken as the mean income for the sector of the labour market in which the individual is seeking work. z_8, before logging, includes unemployment benefit but excludes tax rebates and redundancy payments. Note that z_6, z_7 and z_8 are adjusted quarterly and are, therefore, time dependent. z_8 is not recorded during the first 4 weeks and z_7 was not evaluated for those individuals obtaining work within the first 4 weeks. These features led us to analyse the data from $t = 28$ days, conditioning on individuals remaining unemployed for a month or more. This, and the exclusion of withdrawals from the register due to events other than obtaining work (ex: retirement, etc.), reduces the original data set to 1538 individuals.

There are many ways to tackle the problem created by the absence of two covariates at the beginning of the time scale. The first one, adopted by Narendranathan et al (1985) and Gamerman and West (1987), is to analyse the data conditioned on the duration time being greater than 28 days thus reducing the number of individuals to 1538. This can be described as the safest way because it avoids the problem of missing covariates and potential misspecification of the model. Before going any further, it should be pointed out that this approach is completely sound for $t > 28$ but leaves the first four weeks of unemployment without being analysed. Other possibilities are explored later on.

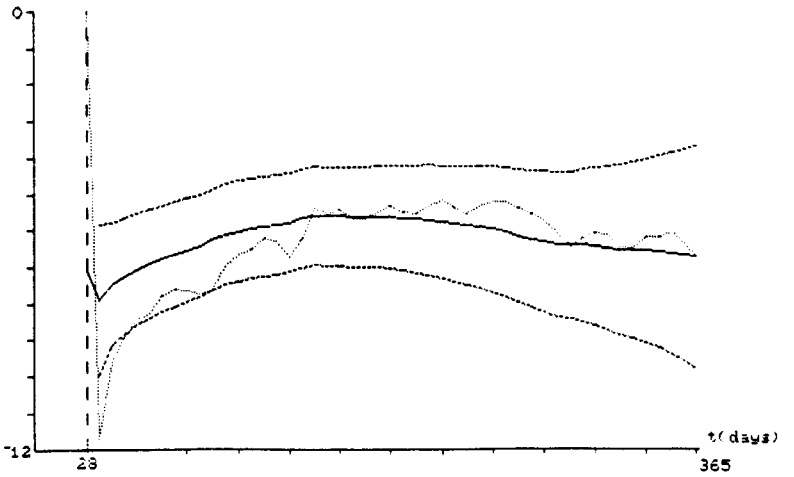
The grid of times is determined by every full week elapsed. Economic indicators are generally provided on a weekly basis so that it is sensible in this case to assume that the parameters are constant within each week changing afterwards. Also, it is noted that there is enough information within each week and that the number of time points on such

grids is reasonably large to capture dynamic movements in the parametric trajectories. As a bonus, it results that all the intervals have equal length allowing many forms of evolution to be adopted. In particular, it is a useful case for discount evolutions since the cautious note in Chapter 2 about its use with grids of unequal interval lengths is no longer applicable. The discount approach used here is the simplest of all with a single discount $d \in (0,1]$ applicable to all the parameters and constant throughout time. Namely, in the notation of Chapter 2, $\mathbf{B}_i = d\, I_{p+1}$ implying through (2.9) that $\mathbf{R}_i = \mathbf{C}_{i-1}/d$. Thinking in terms of precision or inverse variance as a measure of information, it can be rewritten as $\mathbf{R}_i^{-1} = d\, \mathbf{C}_{i-1}^{-1}$ saying that $100.d\%$ of the information passes to the next interval and $100(1-d)\%$ is lost as time passes. Also note that a single discount as above ensures that $\mathbf{R}_i$ is positive semidefinite if $\mathbf{C}_{i-1}$ is which is by no means guaranteed with a general discount matrix as in (2.9).

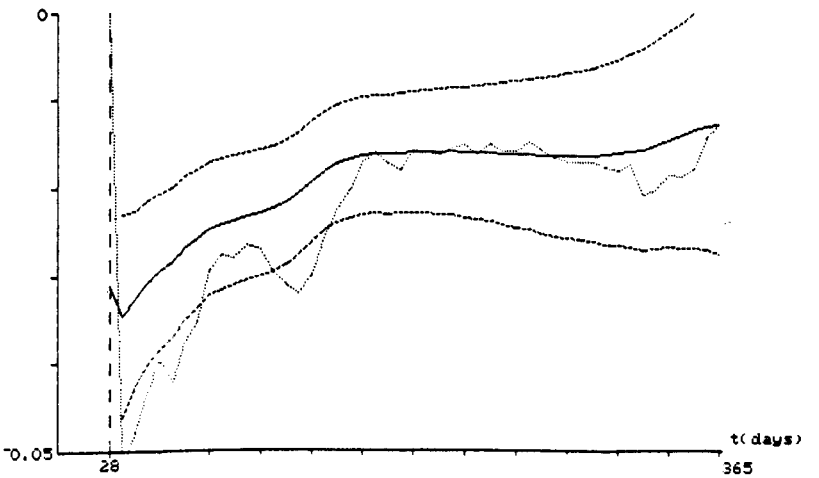
For the present application, the values of $d = .9$ and $d = 1$ were selected. The first one represents a truly dynamic model discounting 10% of the information as a full week elapses and the second one represents a static model since it preserves variances through time. A comparison of these two models reveals that their respective likelyhoods are -8054.1 and -8115.8. The Bayes factor of exp(61.7) is overwhelmingly in favour of a dynamic model as expected. In fairness, the adoption of $d = 1$ implies not only a static model for the regression parameters but for the baseline hazard as well with an implicit assumption of exponentiality. This could have impaired the dynamic vs static comparison were it not for the fact that the baseline hazard even for the dynamic model behaves very nearly like an exponential one as can be seen in Figure 10.a.

Also in Figure 10 are the trajectories of the regression parameters with the 2 standard deviation bounds on the smoothed mean trajectories providing an idea of the uncertainty associated with the parameters. The main feature of the Figure is the decay in the explanatory power of the covariates as time passes. This is particularly evident in Figure 10.i. The effect of the unemployment income is clearly negative as anticipated and very precise at the start but as time passes it steadily weakens and its precision decreases and

(a) Baseline hazard



(b) Age effect



(c) Effect of being married

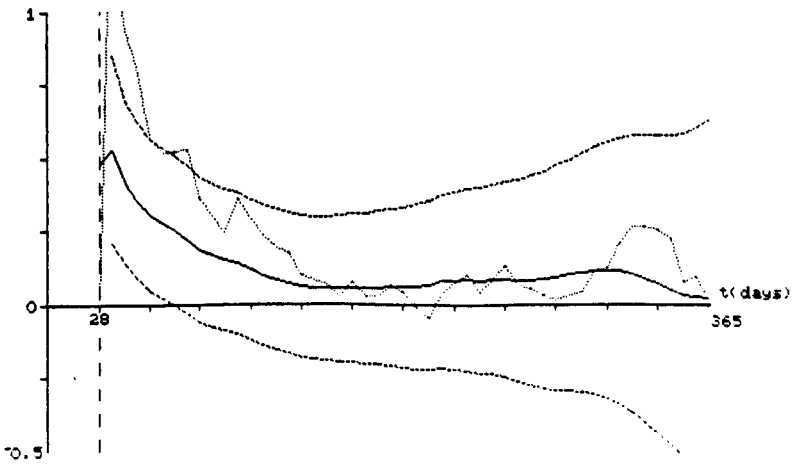
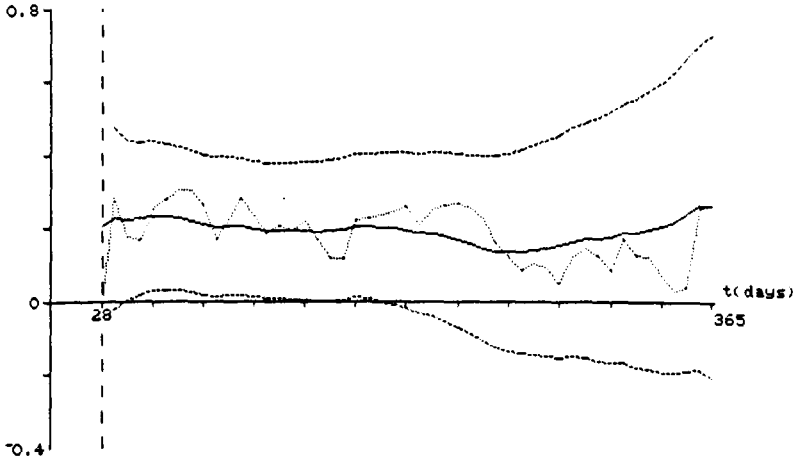
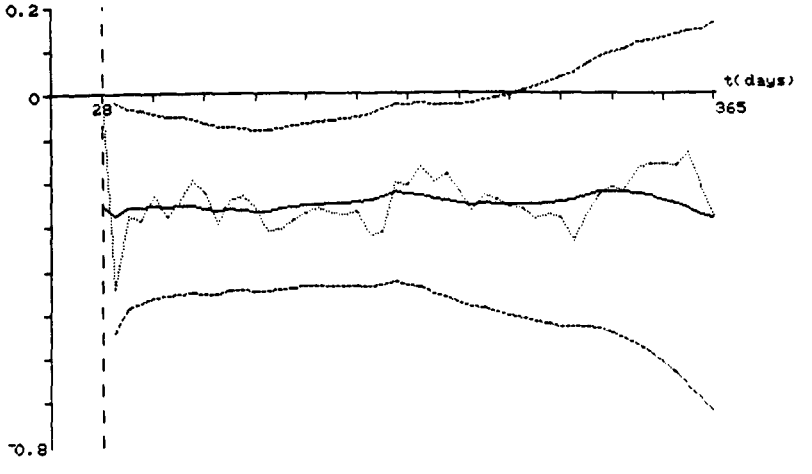


Figure 10(a,b,c). Mean trajectories for unemployment data (——, smoothed; ---, 2 S.D. limits; ····, on-line)

(d) Education indicator



(e) Less than 1 yr. in last job ind.



(f) Effect of being unemployed last year

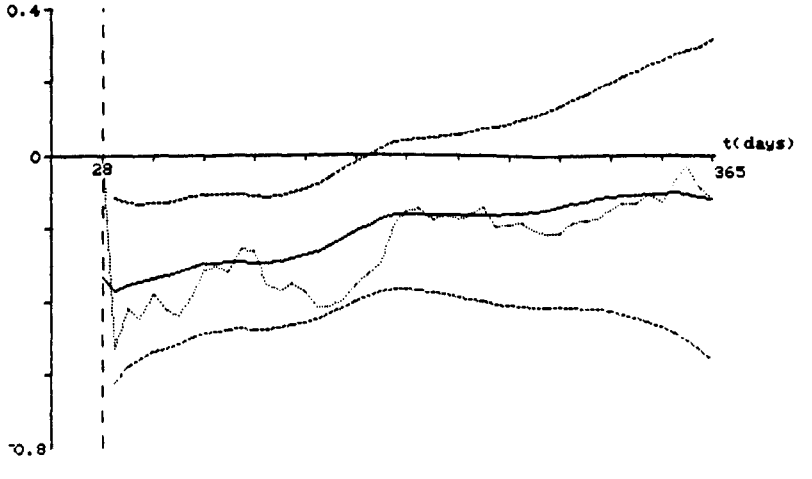
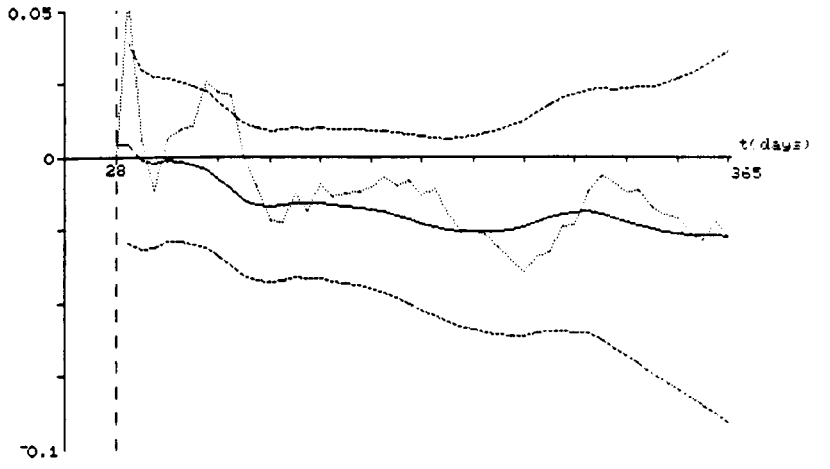
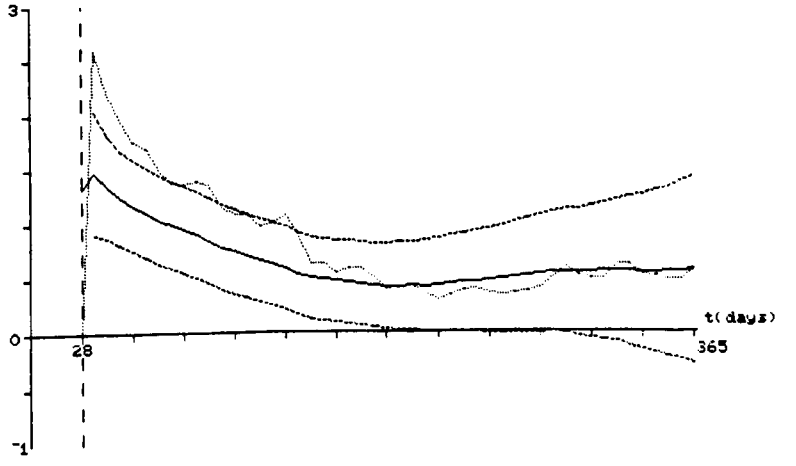


Figure 10(d,e,f). Mean trajectories for unemployment data
(——, smoothed; ---, 2 S.D. limits; ····, on-line)

(g) Unemployment rate effect



(h) Effect of income at work



(i) Unemployment income effect

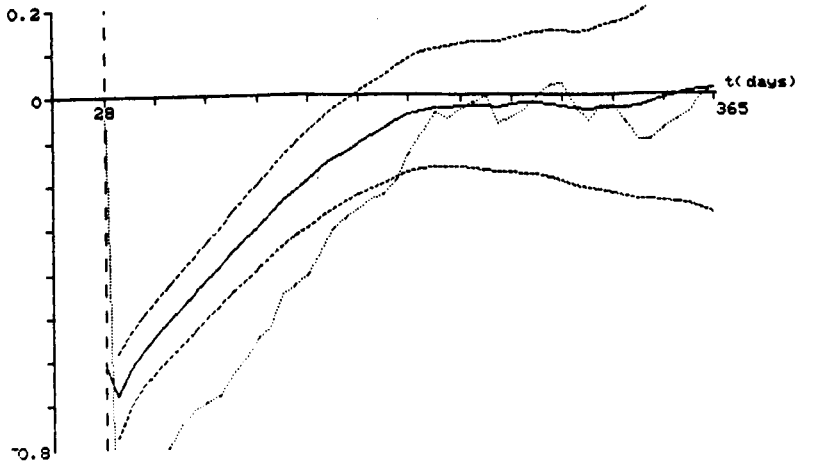


Figure 10(g,h,i). Mean trajectories for unemployment data
(——, smoothed; ---, 2 S.D. limits; ····, on-line)

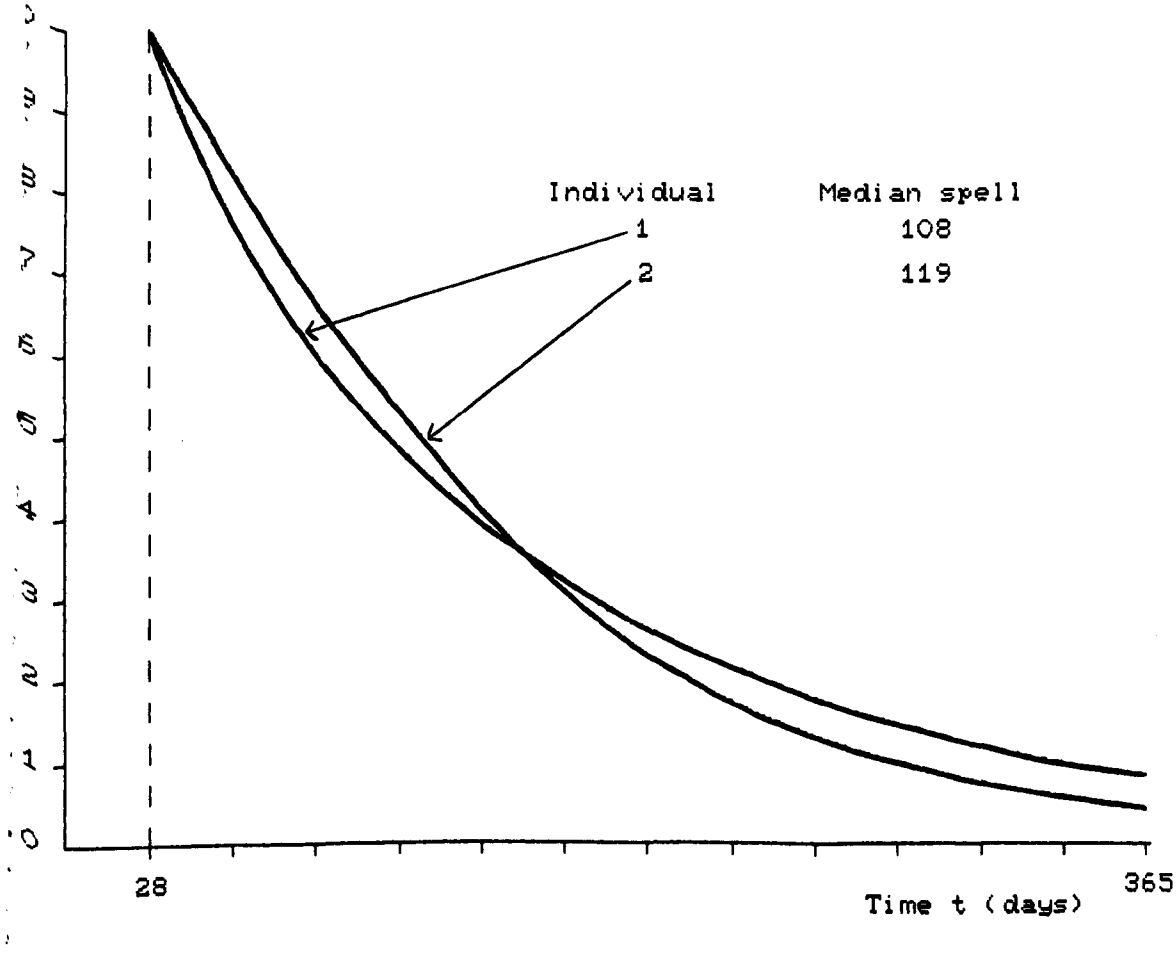


Figure 11. Survival functions for two individuals

after around 6 mouths it does not provide any explanation to the model. This is in line with the view that, as unemployment spell increases, individuals react similarly irrespective of their characteristics, indicating a very strong, common social effect. The only covariates not following this pattern are the education factor, the indicator of being less than 12 months in the last full time job (staying power indicator) and the local unemployment rate. The first two remain fairly constant throughout showing the importance of education and previous labour history while the last one, in fact, steadily departs from 0 but with an increasingly large variance preventing it from taking over as a limiting explanatory variable. The baseline hazard, although remaining fairly constant, has some indication of increasing up to around the half year mark and decreasing thereafter. Of course, this needs further exploration from both theoretical and empirical grounds that could lead to a better understanding of the unemployment process as a whole. The final estimates obtained with a static model ($d = 1$) are presented in Table 6.

TABLE 6

Parameter estimation with the static model

Parameter	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
mean	-.0238	.253	.243	-.270	-.311	-.0103	.821	-.366
standard deviation	.0023	.084	.066	.065	.063	.0091	.126	.031

Predictive survival for different individuals in a dynamic model can be expected to cross. Figure 11 shows it to happen for two individuals with respective covariates (20, 1, 0, 1, 1, 16, log 100, log 50) and (30, 1, 1, 0, 0, 16, log 100, log 200). The very high unemployment income and greater age of the second individual make it more unlikely for him to go back to work in the early stages of their unemployment spell. As time passes, these factors lose their importance and his more favourable education and previous labour history characteristics make it more likely that he will get a job in the end.

The problem created by the absence of two important covariates in the first 4 weeks

can be tackled in many alternative ways. Indeed, one could produce an estimate of those covariates based on the data or on previous information and replace it for the period in question. There is a large area of research involving the problem of missing variables that one could look at for help. I shall concentrate, however, on alternatives provided by the flexibility of the dynamic approach. Recursive estimation means that the models are capable of coping not only with time-dependent covariates but also with changes, addition and deletion of covariates. In the general case, suppose that there are $1+p+r$ relevant parameters for interval $\mathbf{I}_{i-1}$ with on-line distribution

$$\left[\begin{pmatrix} \beta_I \\ \beta_{II} \end{pmatrix}_{i-1} \mid D_{i-1} \right] \sim \left[\begin{pmatrix} \mathbf{m}_I \\ \mathbf{m}_{II} \end{pmatrix}_{i-1} , \begin{pmatrix} \mathbf{C}_I & \mathbf{C}_{I,II}^T \\ \mathbf{C}_{I,II} & \mathbf{C}_{II} \end{pmatrix}_{i-1} \right] \tag{4.4}$$

where β_I (including the baseline hazard) and β_{II} are $(1+p)$-dimensional and r-dimensional parameters respectively. Suppose now that β_{II} is no longer relevant for the events in interval I_i but a q-dimensional parameter β_{III} is. One proceeds first by evolving (4.4) as usual with $\mathbf{G}_i(b_i)$ and $\mathbf{W}_i$ both $(1+p+r) \times (1+p+r)$ matrices. Then, the marginal distribution $[\beta_{I,i} | D_{i-1}] \sim [\mathbf{a}_{I,i} , \mathbf{R}_{I,i}]$ is taken from it and combined with the prior $[\beta_{III,i} | D_{i-1}] \sim [\mathbf{a}_{III,i} , \mathbf{R}_{III,i}]$ and with covariance term $Cov[(\beta_I , \beta_{III}) | D_{i-1}] = \mathbf{R}_{I,III;i}$ to form the prior distribution for the parameters relevant to interval I_i. From there, one proceeds within I_i with the usual updating to obtain the on-line distribution for $[(\beta_I^T , \beta_{III}^T) | D_i]$. Of course, β_{II} can be reinstated in interval I_{i+1} if that is the case with moments unaffected by the data information in I_i but only by the system evolution, e.g $V[\beta_{II,i} | D_i] = V[\beta_{II,i} | D_{i-1}]$.

Applied to the present example, there are 4 intervals (the first ones) with two absent parameters, β_7 and β_8. So, $\beta_I = (\beta_0 \ \beta_1 \ \cdots \ \beta_6)^T$, $\beta_{III} = (\beta_7 \ \beta_8)^T$ and there is not a β_{II} to be deleted. One possible analysis uses β_I for the first 4 weeks ending up with a prior distribution for interval $I_5 = (28,35]$

$$[\beta_{I,5} \mid D_4] \sim [\mathbf{a}_{I,5} ; \mathbf{R}_{I,5}]$$

The above distribution is in fact the distribution of $[\beta_{I,5} \mid D_4, \beta_{III,i}=0, i=1,2,3,4]$ since in the present case z_7 and z_8 are relevant but missing. It is however a feasible working approximation. At this stage, β_{III} is input to complete the partly vague prior

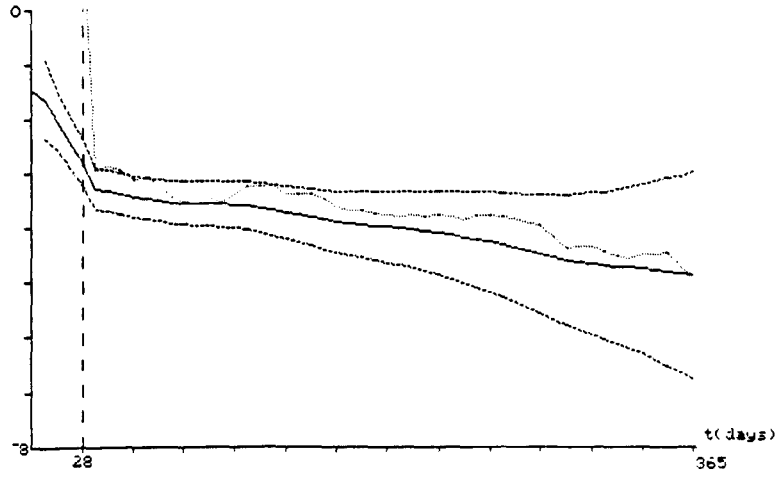
$$\left[\begin{pmatrix} \beta_I \\ \beta_{III} \end{pmatrix}_5 \mid D_4 \right] \sim \left[\begin{pmatrix} \mathbf{a}_I \\ 0 \end{pmatrix}_5 , \begin{pmatrix} \mathbf{R}_I & 0 \\ 0 & 10^3 \mathbf{I}_2 \end{pmatrix}_5 \right] \tag{4.5}$$

From there on, the parameters remain unchanged and analysis proceeds as usual. This has been done here again with the single discount approach for $d=0.9$ and $d=1$. The data set used includes some 120 individuals finding a job within the first 4 weeks along with the previous 1538 individual histories taken from $t=0$ rather than conditioned to $t>28$. Comparison between the dynamic and the static models again provides a Bayes factor of exp(81.3) in favour of a dynamic model. The mean trajectories for the parameters are basically the same as the previous one for $t=28$. The important parameter for the unemployed income, in particular, is remarkably stable for the change in the analysis. The main differences observed occur in the trajectories for β_0, β_6 and β_7 also shown in Figure 12. The baseline hazard is monotonically decreasing now showing more signs of departure from exponentiality. The unemployment rate effect is now more constant and has far less uncertainty associated with it although it is not clear what is the influence of the very extreme values in the early stages. The effect of income at work has been substantially reduced and, apart from the first couple of months, is judged to be irrelevant.

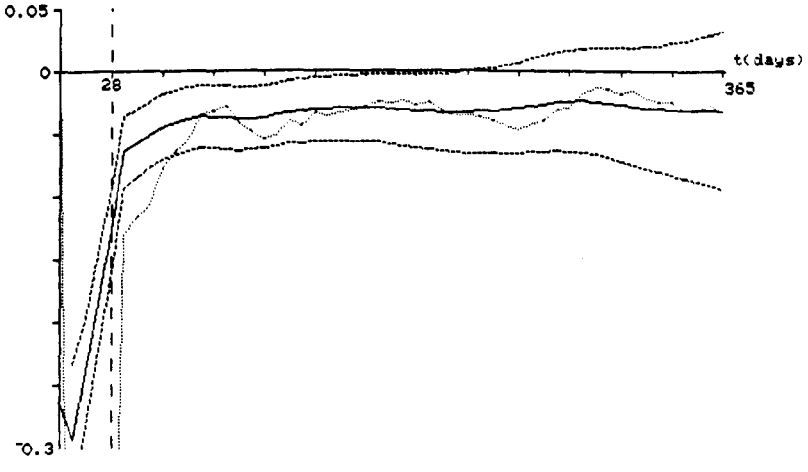
Another possibility is a compromise between the previous ones. The information collected in the first 4 weeks can be used only as far as the prevision of an informative prior for $[\beta_{I,1} \mid D_0]$ where $\beta_{I,1}$ is the parameter for interval $I_1=(28,35]$ and D_0 includes all the events observed in the first 4 weeks. The prior can be completed as in (4.5) and recursive estimation starts from there. The difference between this approach and the first one is on the specification of a more precise marginal prior distribution for $[\beta_{I,i} \mid D_0]$.

The trajectories are, of course, the same in the second approach but for the absence of the first 4 weeks. The comparison of this model with $d=.9$ with the first one with the

(a) Baseline hazard



(b) Unemployment rate effect



(c) Effect of income at work

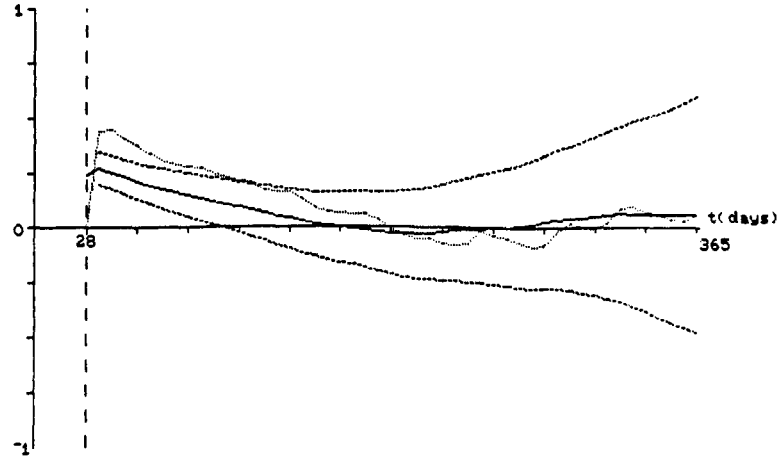


Figure 12. Mean trajectories for unemployment data from t = 0 (____, smoothed; ----, 2 S.D. limits; ...., on-line)

fairly vague prior shows a Bayes factor of exp(3.56)=35.1 in favour of the precise prior one. Due to the amount of data and the vagueness of the prior in the first model, this factor is somewhat low. This could be due to model misspecification as suggested by the trajectories of β_7 in figure 12.c. A comparison static versus dynamic models when both have precise prior distribution gives a factor of exp(81.8) in favour of a dynamic model once again.

Appendix 4.1

Dynamic linear growth model

Let β_0, β_1 and β_2 be the baseline hazard parameter, a general regression parameter and a parameter representing the slope of the trajectory experienced by β_1 as it moves through time. This model is commonly describe as a linear growth model but it should be understood that this growth measured by β_2 could well be negative.

For the particular problem of section 4.1, it is useful to look in the time scale t^* measured in months or rather $t^* = t/30$ where t is measured in days but the derivations of the model as such are the same once the time scale is agreed upon and fixed. Comparison with the classical results presented at the beginning of that section are made easier through this time rescaling and for ease of notion, t^* is hereafter denoted by t until further notice.

Given no further information for the system, the differential equations for β_0 and β_2 are

$$d\beta_0(t) = dw_1(t) \qquad \text{(A4.1a)}$$

$$d\beta_2(t) = dw_3(t) \qquad \text{(A4.1b)}$$

where w_1 and w_3 are OII processes defined in section 2.3 not necessarily independent. So, β_0 and β_2 are modelled as random walk-like processes with uncertainty about them increasing according to the respective diffusion of w_1 and w_3 . The differential equation for β_1, however, should also take into account of the modelling assumption that β_2

represents its slope thus leading to

$$d\beta_1(t) = \beta_2(t)\,dt + dw_2(t) \tag{A4.1c}$$

where w_2 is yet another OII process. Equations (A4.1) can be written in the more compact form

$$d\boldsymbol{\beta}(t) = \mathbf{F}\,\boldsymbol{\beta}(t)dt + d\mathbf{w}(t) \tag{A4.2}$$

where

$$\boldsymbol{\beta}(t) = \begin{pmatrix} \beta_0(t) \\ \beta_1(t) \\ \beta_2(t) \end{pmatrix}, \quad \mathbf{F} \quad \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } \mathbf{w}(t) = \begin{pmatrix} w_1(t) \\ w_2(t) \\ w_3(t) \end{pmatrix}$$

The term $d\mathbf{w}(t)$ can be replaced by the more general $\boldsymbol{\Gamma}(t)d\mathbf{w}(t)$ as before. Equation (A4.2) is in standard linear form already discussed in chapter 2 with solution

$$\boldsymbol{\beta}(t) = \mathbf{G}(t,0)\boldsymbol{\beta}(0) + \int_0^t \mathbf{G}(t,s)\boldsymbol{\Gamma}(s)\,d\mathbf{w}(s)$$

with

$$\mathbf{G}(t,s) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & t-s \\ 0 & 0 & 1 \end{pmatrix}$$

given the initial condition $\boldsymbol{\beta}(0)$. Recursive procedures lead to

$$\boldsymbol{\beta}(t) = \mathbf{G}(t,v)\boldsymbol{\beta}(v) + \int_v^t \mathbf{G}(t,s)\boldsymbol{\Gamma}(s)\,d\mathbf{w}(s) \tag{A4.3}$$

where $t > v$. Once a grid $\mathcal{T}$ of points is set and the identifications $\boldsymbol{\beta}(t_i) \equiv \boldsymbol{\beta}_i$ and $\mathbf{G}(t,t_{i-1}) \equiv \mathbf{G}_i(t-t_{i-1})$ are made, (A4.3) reduces to

$$\boldsymbol{\beta}_i = \mathbf{G}_i(b_i)\boldsymbol{\beta}_{i-1} + \mathbf{w}_i\,, \quad \mathbf{w}_i \sim [0, \mathbf{W}_i] \tag{A4.4}$$

where $\mathbf{w}_i$ is the integral term of (A4.3) and its covariate matrix $\mathbf{W}_i$ is given by

$$\int_{t_i}^{t_{i-1}} \mathbf{G}(t_i,s)\boldsymbol{\Gamma}(s)\mathbf{Q}(s)\boldsymbol{\Gamma}^T(s)\mathbf{G}^T(t_i,s)\,ds$$

where $\mathbf{Q}(s)$ is the (time-varying) diffusion of the OII process $\mathbf{w}(t)$. Assuming $\boldsymbol{\Gamma}(s) = \mathbf{I}$ and $\mathbf{Q}(s) = \mathbf{Q} = (q_{ij})$ for $s \in I_i = (t_{i-1}, t_i]$ gives

$$\mathbf{W}_i = b_i\,\mathbf{Q} + b_i^2 \begin{pmatrix} 0 & 2q_{13} & 0 \\ 2q_{13} & q_{23} + (b_i/3)q_{33} & 2q_{33} \\ 0 & 2q_{33} & 0 \end{pmatrix} \tag{A4.5}$$

If $b_i < 1$, the second term in the R.H.S. of (A4.5) can be neglected. As previously suggested, a useful assumption is that the process is driven by a single OII process with different magnitudes, that is,

$$\mathbf{Q}(s) = Q \ , \ \Gamma^T(s) = \frac{1}{\sqrt{Q}}(\sigma_{i,0} \ \sigma_{i,1} \ \sigma_{i,2}) \ , \ s \in I_i = (t_{i-1}, t_i]$$

giving

$$\mathbf{W}_i = b_i \boldsymbol{\sigma}_i \boldsymbol{\sigma}_i^T + b_i^2 \begin{pmatrix} 0 & 2\sigma_{i,0}\sigma_{i,2} & 0 \\ 2\sigma_{i,0}\sigma_{i,2} & \sigma_{i,1}\sigma_{i,2} - (b_i/3)\sigma_{i,2} & 2\sigma_{i,2}^2 \\ 0 & 2\sigma_{i,2}^2 & 0 \end{pmatrix}$$

and the same comments apply if $b_i < 1$. So, the adoption of the relation $\mathbf{W}_i = b_i \boldsymbol{\sigma}_i \boldsymbol{\sigma}_i^T$ constitutes a good practical approximation as the leading term in the expression above.

References

Carter, W. H., Wampler, G. L. and Stablein, D. M. (1983). *Regression Analysis of Survival Data in Cancer Chemotherapy*. New York: Marcel Dekker.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant data. *J. Amer. Statist. Ass.*, **72**, 27-36.

Gamerman, D. and West, M. (1987). A time series application of dynamic survival models in unemployment studies. To appear in *The Statistician*.

Narendranathan, W., Nickell, S. and Stern, J. (1985). Unemployment benefits revisited. *Economic Journal*, **95**, 307-329.

Nickell, S. (1979). The effect of unemployment and related benefits on the duration of unemployment. *Economic Journal*, **89**, 34-49.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *J. R. Statist. Soc., Ser B*, **42**, 213-220.

Smith, A. F. M., West, M., Gordon, K., Knapp, M. S. and Trimble, I. (1983). Monitoring kidney transplant patients. *The Statistician*, **32**, 46-54.

Stablein, D. M., Carter, W. H. and Novak, J. W. (1981). Analysis of survival data with non-proportional hazard functions. *Cont. Clin. Trials*, **2**, 149-159.

CHAPTER FIVE

SPECIAL MODELS

The models proposed in Chapter 3 are general and encompass a wide range of usages. In the previous chapter, general regression and analysis of variance models were used to analyse medical and econometric data. In this chapter, some special cases of the general formulation are considered. This is done here to explore further the characteristics of the particular models and also to show applicability in other areas.

The first case is the analysis of random samples. This has received a lot of attention in the literature following the paper by Kaplan & Meier (1958), which introduced the product-limit (PL) estimator (see Breslow and Crowley (1974) for a list of references). The general framework can be specialised to the case of no covariates and properties of the model are derived. The models are shown to yield to a more general analysis with far more modelling tools available to the practitioner than the other approaches. Comparison with the PL is established at theoretical and practical levels through some simulated samples.

The second case concerns the analysis of life tables. The emphasis is in the analysis with specific parametric forms for the force of mortality. Examples provided include the logistic force of mortality that leads to a non-linear analysis.

Finally, the analysis of competing risk is presented. This is yet another example of the generality of the model and provides a useful insight into further extensions of the models towards observation of multivariate point processes.

5.1 - Analysis of Random Samples

The problem in question is the inference about an unknown distribution from which a sample $\mathbf{Y} = (Y_1, \ldots, Y_m)$ is drawn. Interest lies mainly in the population survival function $S(u) = P(Y > u)$. The only information to be obtained about the sample is the observed values of $\mathbf{Y}$ and the vector $\chi = (\chi_1, \ldots, \chi_m)$ of death indicators. No additional

information, such as covariates, is available, implying an homogeneous population. The quantities **Y** therefore have conditional density given by

$$\prod_{i=1}^{m} f(y_i \mid \lambda)$$

for some parameter λ. This problem is analysed directly from this set-up by Gamerman (1987). Here, this task is simplified by the use of the structure provided by the dynamic survival models.

5.1.1- Estimation procedures and models

The absence of covariates leads to $z_j = 1, j = 1,...,m$ and $\beta = \beta_0$. The fact that $z_j = 1$ for all individuals implies that $\lambda_i^{(j)} = \lambda_i$, $j = 1,...,r_i$, $i = 1, 2, ..., N$ and the guide relations are simply

$$\log \lambda_i \simeq \beta_{0,i}, \; i = 1, 2, ..., N \tag{5.1}$$

Once again, the last interval I_{N+1} can be neglected after assigning to t_N a value after which there is no data information. Finally, the factorised likelihoods $\{l_i, i = 1, 2, ..., N\}$ reduce to

$$\prod_{j=1}^{r_i} \lambda_i^{x_{ij}} \exp\{-\lambda_i (y_{ij} - t_{i-1})\} = \lambda_i^{d_i} \exp\{-\lambda_i v_i\}, \; i = 1, 2, ..., N \tag{5.2}$$

with

$$d_i = \sum_{j=1}^{r_i} x_j = \text{number of deaths observed in } I_i$$

$$v_i = \sum_{j=1}^{r_i} (y_{ij} - t_{i-1}) = \text{total survival time observed in } I_i$$

If the system (represented by the baseline hazard parameter β_0) is assumed to be disturbed by an OII process, then the difference system equation is

$$\beta_{0,i} = \beta_{0,i-1} + w_i, \; w_i \sim [0, W_i] \tag{5.3}$$

If, in addition, the OII process has a constant diffusion Q, then $W_i = b_i Q$. A full cycle in the sequential estimation procedure is as follows:

(i) let $[\beta_{0,i-1} \mid D_{i-1}] \sim [m_{i-1}, C_{i-1}]$;

(ii) the system equation (5.3) gives $[\beta_{0,i} \mid D_{i-1}] \sim [a_i, R_i]$ with $a_i = m_{i-1}$ and $R_i = C_{i-1} + W_i$

(iii) the guide relations (5.1) imply as before, $[\lambda_i \mid D_{i-1}] \sim G(\alpha_i, \gamma_i)$ with $\alpha_i = R_i^{-1}$ and $\gamma_i = R_i^{-1} exp(-a_i)$;

(iv) the likelihoods (5.2) update the distributions in (iii) to $[\lambda_i \mid D_{i-1}] \sim G(\alpha_i + d_i, \gamma_i + v_i)$;

(v) linear Bayesian methods lead to $[\beta_{0,i} \mid D_i] \sim [m_i, C_i]$ with

$$m_i = a_i + \log\left(\frac{1 + R_i d_i}{1 + R_i v_i \exp(a_i)}\right) \tag{5.4a}$$

$$C_i = R_i \frac{\alpha_i}{\alpha_i + d_i} \tag{5.4b}$$

In fact, equations (5.4) are also obtained if the analysis is done as before: one observation at a time. In this case, the order in which the observations are analysed is irrelevant. This is a special case of a result ensuring that, in the static case, (conditionally) independent pieces of survival information with the same covariates can be treated as one. The result is proven as a lemma in the next section where it is needed in a general form.

An interesting consequence of the approach is provided by a comparison of the distributions of the consecutive λ's , λ_i and λ_{i+1} say. This gives the effect on the hazard level of the evolution of β_0. Starting from (v) above, the system evolution leads to $[\beta_{0,i+1} \mid D_i] \sim [m_i, C_i + W_{i+1}]$. The guide relation gives $[\lambda_{i+1} \mid D_i] \sim G(\alpha_{i+1}, \gamma_{i+1})$ with $\alpha_{i+1} = (C_i + W_{i+1})^{-1}$ and $\gamma_{i+1} = (C_i + W_{i+1})^{-1}\exp(-m_i)$. Using (5.4), and the guide relations in (iii), it can be shown that

$$\alpha_{i+1} = \phi_{i+1}(\alpha_i + d_i) \tag{5.5a}$$

$$\gamma_{i+1} = \phi_{i+1}(\gamma_i + v_i) \tag{5.5b}$$

with $\phi_{i+1} = [1 + (\alpha_i + d_i) W_{i+1}^{-1}]^{-1} [1 - C_i^{-1} W_{i+1}]^{-1}$ within the interval (0,1). Comparison with the posterior distribution for λ_i in (iv) shows that

$$E[\lambda_{i+1} \mid D_i] = E[\lambda_i \mid D_i]$$

$$V[\lambda_{i+1} \mid D_i] = \phi_{i+1}^{-1} V[\lambda_i \mid D_i]$$

The system equation therefore implies a smooth evolution for the hazard by maintaining the previous level with increased uncertainty. This seems to be the correct way to model the hazard function although any relevant information about its form can be put into the model via a non-standard ω_i possibly with non-zero, system-dependent mean. This point is explored in a more systematic way in the next section.

The smooth distributions of $\beta_{0,i} \mid D_N$, $i = 1,\ldots,N$ are obtained as in chapter 3. Equation (3.22) can be specialised to the present situation giving

$$m_i^N = m_i - \phi_{i+1} (m_{i+1}^N - m_i) \tag{5.6a}$$

$$C_i^N = C_i - \phi_{i+1}^2 (C_i + W_i - C_{i+1}^N) \tag{5.6b}$$

In the same way, the predictive likelihood can be obtained from (3.28). This was derived with the one-at-a-time updating required for the general model. Use of the particular structure of random samples makes this derivation simpler. From (3.25), (5.2) and (ii) above, the predictive likelihood for any particular model is

$$\begin{aligned} l(\mathbf{y} \mid M) &= \prod_{i=1}^{N} l_i(\mathbf{y}_i \mid D_i, M) \\ &= \prod_{i=1}^{N} \int_0^\infty l_i(\mathbf{y}_i \mid \lambda_i, D_i, M)\, f(\lambda_i \mid D_i, M)\, d\lambda_i \\ &= \prod_{i=1}^{N} \int_0^\infty \lambda_i^{d_i} \exp(-\lambda_i v_i)\, \frac{\gamma_i^{\alpha_i}}{\Gamma(\alpha_i)} \lambda_i^{\alpha_i - 1} \exp(-\lambda_i \gamma_i)\, d\lambda_i \\ &= \prod_{i=1}^{N} \left\{ (\gamma_i + v_i)^{-d_i} \left[1 + \frac{v_i}{\gamma_i} \right]^{-\alpha_i} \prod_{j=1}^{d_i} (\alpha_i + j - 1) \right\} \end{aligned} \tag{5.7}$$

The above result can be obtained from (3.28) after observing that $\alpha_{i,j} = \alpha_{i,j-1} + 1$ if j^{th} individuals dies in I_i and $\alpha_{i,j} = \alpha_{i,j-1}$, otherwise. Some terms vanish and others cancel

out leading finally to (5.7).

5.1.2- Predictive survival function

Since there are no regression parameters to be estimated, the primary interest in the analysis of random samples is to make inference about the survival function. Once again, the theory of chapter 3 can be used to provide the distribution of $Y_* | D_N$ where Y_* is an hypothetical, or otherwise, individual from the same homogeneous population. This predictive approach to inference seems natural at this stage but it should be mentioned that it is not the most common one. A number of authors (Susarla & Van Ryzin, 1976; Martz & Waller, 1982; Mashhoudy, 1985) adopt the estimation approach. After defining a parametric model $P(\cdot|\lambda)$ for the observations and specifying a prior distribution $\pi(\lambda)$, the survival function at a point t, $S(t)$, is estimated as $S(t|\hat{\lambda})$ where $S(\cdot|\lambda)$ is the survival function induced by P and $\hat{\lambda}$ is an estimator obtained from the posterior distribution $\pi(\lambda|D_N)$. In fact, one is faced with a family of posterior distributions for $\{S(t|\lambda), t>0\}$ as functions of λ. The problem of dealing with a continuous family of (interdependent) posterior distributions are obvious and one has to resort to a summary $\hat{\lambda}$ to be able to make inference. The predictive approach not only avoids this problem but provides the required answer for the practitioner's question: after collecting information from a sample, what can be said about an individual from this population (not belonging to the sample)? Having said that, the estimation approach could be used here as well but the calculations are not presented. The predictive density is easily obtained from section 3.6.2 as

$$p(y_* | D_N) = \left[\frac{\alpha_i}{\gamma_i + y_* - t_{i-1}}\right]^{\chi_{i,*}} \left[1+\frac{y_* - t_{i-1}}{\gamma_i}\right]^{-\alpha_i} \prod_{j=1}^{i-1}\left[1+\frac{b_j}{\gamma_j}\right]^{-\gamma_j} \qquad (5.8)$$

for $y_* \in I_i$, $i=1,2,...,N$ and with $\chi_{i,*}$ = death indicator for $Y_* \in I_i$, $\alpha_i = (C_i^N)^{-1}$ and $\gamma_i = (C_i^N)^{-1}\exp(-m_i^N)$.

5.1.3- Comparison with other estimators

It is interesting at this stage to compare the survival function implied by (5.8) with PL estimator derived by Kaplan & Meier (1958). The latter is given by

$$\hat{S}(t) \quad \prod_{i:\, y_i < t} \left[1 - \frac{d_i}{r_i} \right] \tag{5.9}$$

where y_i is the i^{th} ordered death time. This is a piecewise constant function with jumps at the observed death times. It is therefore advisable to restrict comparison with inference based on the grid τ_{ODT} of death times. Before doing that, comparison of (5.9) with (3.8) reveals that all the PL estimator is doing is estimating the probability of surviving between two successive death times by the sample frequency of individuals not dying in that interval. Thus, there is no passage of system information between intervals. This has been extensively advocated in this thesis to be unreasonable but for comparison purposes a non-informative model can be chosen by letting $W_i \to \infty, \forall i$, or, equivalently, $\phi_i \to 0, \forall i$. In this limiting case, $\lambda_i | D_N \sim \lambda_i | D_i \sim G(d_i, v_i)$. To make clearer the comparison that follows assume further that $v_i = b_i r_i$, that is, all the individuals contributing to I_i have their survival times equal to t_i, even the censored ones. Substitution of this information $(\alpha_i = d_i, \gamma_i = b_i r_i)$ into (5.8) gives

$$S(y_s | D_N) = \left[1 - \frac{y_s - t_{i-1}}{b_i r_i} \right]^{-d_i} \prod_{j=1}^{i-1} \left[1 + \frac{1}{r_j} \right]^{-d_j}, \quad y_s \in I_i, i = 1, 2, \ldots, N \tag{5.10}$$

Each of the multiplying factors of (5.10) is a function of $r_k^{-1}, k = 1, \ldots, i$. For large and moderate samples, r_k is fairly large, particularly for i not close to N, and a linear approximation can be sought for r_k^{-1} near 0. This leads to

$$S(y_s | D_N) \doteq \left[1 - \frac{d_i}{r_i} \frac{y_s - t_{i-1}}{b_i} \right] \prod_{j=1}^{i-1} \left[1 - \frac{d_j}{r_j} \right], \quad y_s \in I_i, \; i = 1, 2, \ldots, N \tag{5.11}$$

Comparison of (5.11) with (5.9) shows that $\hat{S}(t) \doteq S(y_i | D_N)$ as $t \uparrow y_i$. In between death times, the non-informative survival function is approximately linearising the PL survival. The approximations above can be very poor for small samples or when i approaches N, but the examples below confirm the non-informative approach as a smoothing procedure with respect to the PL estimator.

It is also worth noting the Bayesian estimator proposed by Susarla & Van Ryzin

(1976). This approach starts by identifying the sampling distribution of the observations as the parameter to be estimated. A Dirichlet process prior (Ferguson, 1973) is assigned to the distribution and the estimator is the mean of the Dirichlet process that results as the posterior distribution (for the sampling distribution) after observing the data. Their estimate is smoother than the step function-like PL estimator but it still has the drawback of leading to discontinuities at death times. The same comments previously made with respect to the PL estimator therefore apply to the Susarla & Van Ryzin estimator albeit to a lesser extent. It seems to be situated between the piecewise constant PL estimator and the continuous non-informative survival. Surprisingly, there is no mention in their paper of a predictive approach although it should not be that difficult to obtain it within that conjugate Bayesian framework.

5.1.4- Numerical examples

Two sets of data are used. The first consists of 3 simulated samples and attention is concentrated on the predictive survival function. The second comes from medicine and illustrates model choice with respect to a couple of user-designed parameters. In all examples, analysis starts with vague prior $\lambda_0 \sim G(\epsilon,\epsilon)$ with $\epsilon=10^{-3}$. This implies $V[\lambda_0|D_0] = 10^3$ representing ignorance as to the value of λ_0. Also, for β_0 it is equivalent to an initial prior with zero mean and variance 10^3 as in the numerical applications of the previous chapter.

One sample of size 30 without censoring was generated from the Weibull(3,100), Weibull(5,100) and logistic(2,100) distributions. These have respective hazard functions

$$.03\,(\log 2)\left(\frac{t}{100}\right)^{2}\ ,\quad .005\,(\log 2)\left(\frac{t}{100}\right)^{5}\ \text{and}\ \ .02\,\frac{t/100}{1+(t/100)^{2}}\ ,\ t>0$$

Their common median and scale parameter is 100 and they represent quite different failure patterns. As shown in figure 13, the first is monotonically increasing, and the second is monotonically decreasing, whilst the logistic has increasing hazard up to the median and is steadily decreasing thereafter. There are practical circumstances leading to

each of those three types of hazards but the main point is that they represent three different forms of hazard of practical interest. The samples are presented in Table 7.

TABLE 7

Simulated samples

Weibull(3, 100)	9.096, 61.682, 62.351, 67.196, 72.896, 75.980, 76.470, 87.726, 91.422, 94.118, 94.964, 96.101, 96.229, 107.930, 109.247, 113.651, 113.900, 114.388, 117.629, 120.161, 120.165, 122.107, 128.743, 131.816, 133.954, 135.492, 138.762, 145.127, 155.529, 161.479
Weibull(.5, 100)	0.449, 2.365, 3.341, 3.478, 3.823, 10.939, 12.565, 29.081, 41.726, 46.260, 47.245, 63.671, 103.620, 146.009, 151.213, 172.863, 204.324, 232.389, 296.789, 461.138, 463.015, 535.003, 609.774, 1066.881, 1218.685, 1777.358, 1933.132, 2177.856, 2207.325, 2972.893
logistic(2, 100)	12.330, 27.851, 50.289, 54.753, 68.584, 71.551, 71.850, 81.580, 95.676, 96.184, 101.033, 120.885, 123.105, 134.265, 136.881, 140.598, 144.199, 144.199, 150.121, 160.427, 161.067, 165.579, 186.192, 187.258, 192.537, 230.816, 284.441, 355.186, 534.879, 695.853

For each of the samples, analysis based on a grid $\mathcal{T}_{ODT}$ is performed for two types of evolution: a dynamic with constant diffusion Q and a non-informative. The respective values of Q adopted for the samples were two types of evolution. The Figures present also the PL estimator and survival function of the generating distribution. It can be seen that in all cases, the non-informative survival is essentially a smoothed version of the PL estimator, confirming earlier comments. A reasonable degree of smoothness, however, of the sort exhibited by the generating survival function, is only achieved with a proper dynamic model. The difficult case of a decreasing hazard leads to a overweight of the resulting survival functions in all the approaches.

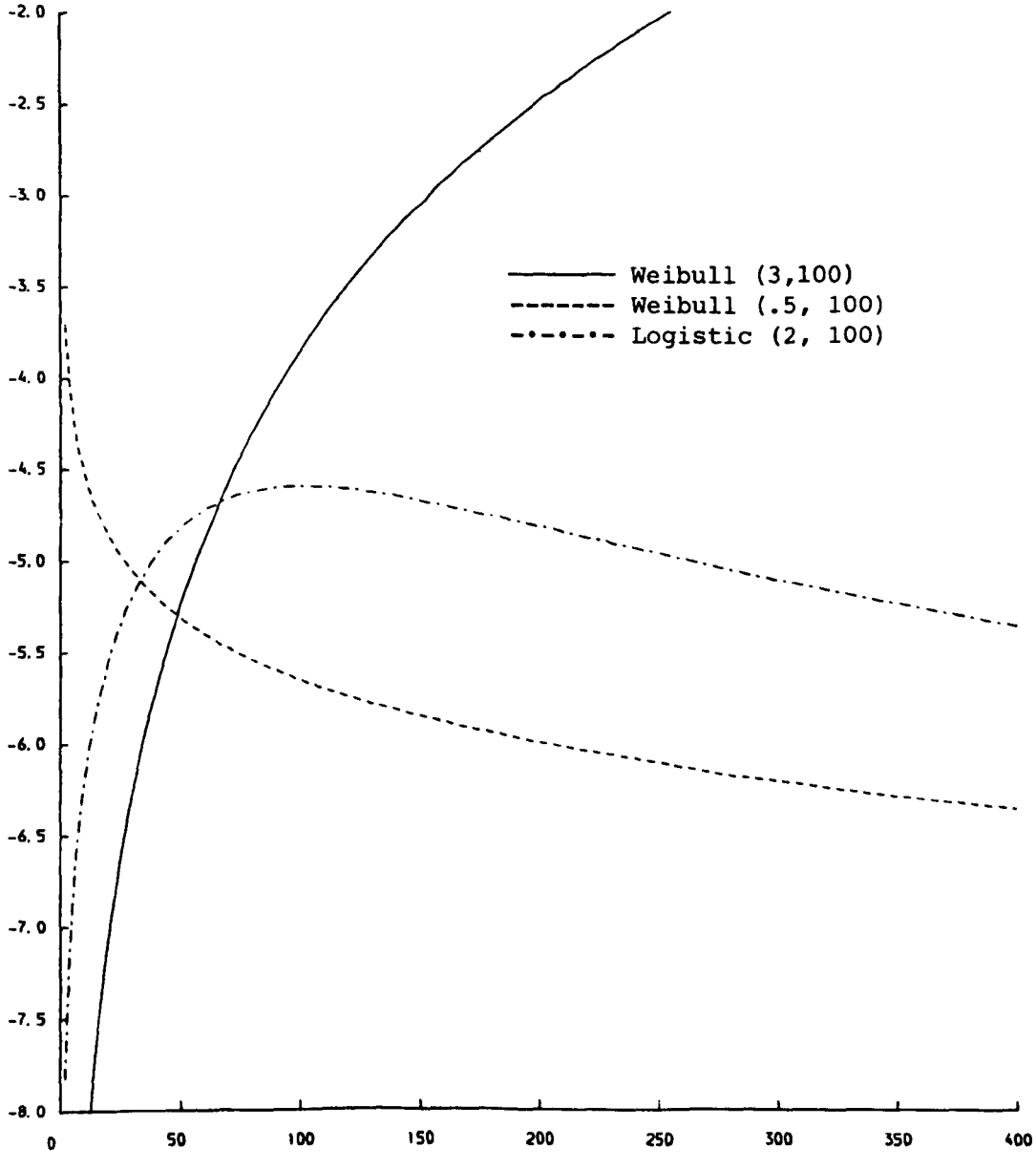


Figure 13. Hazard functions (in the log scale)

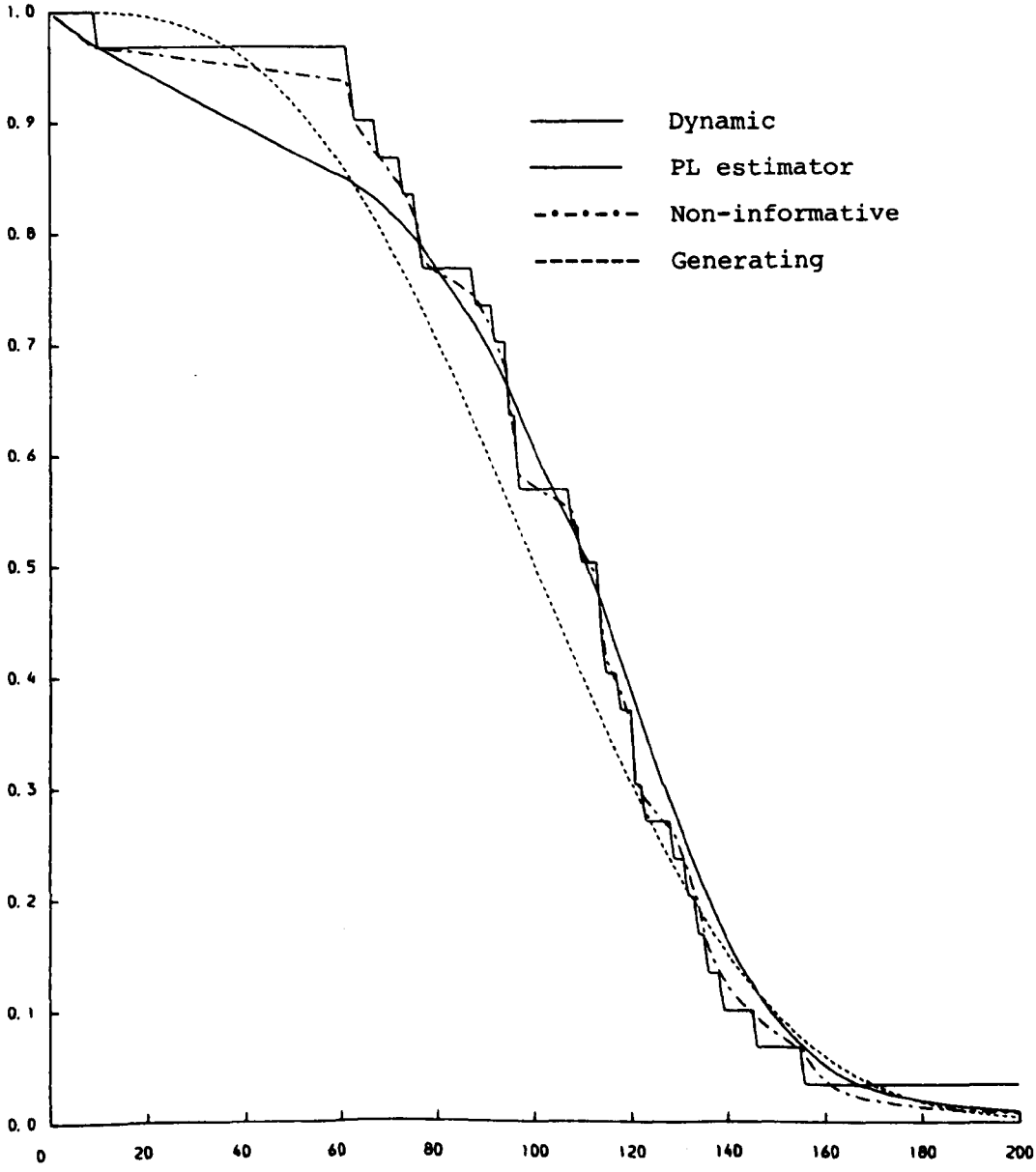


Figure 14(i). Survival functions for the Weibull (3,100).

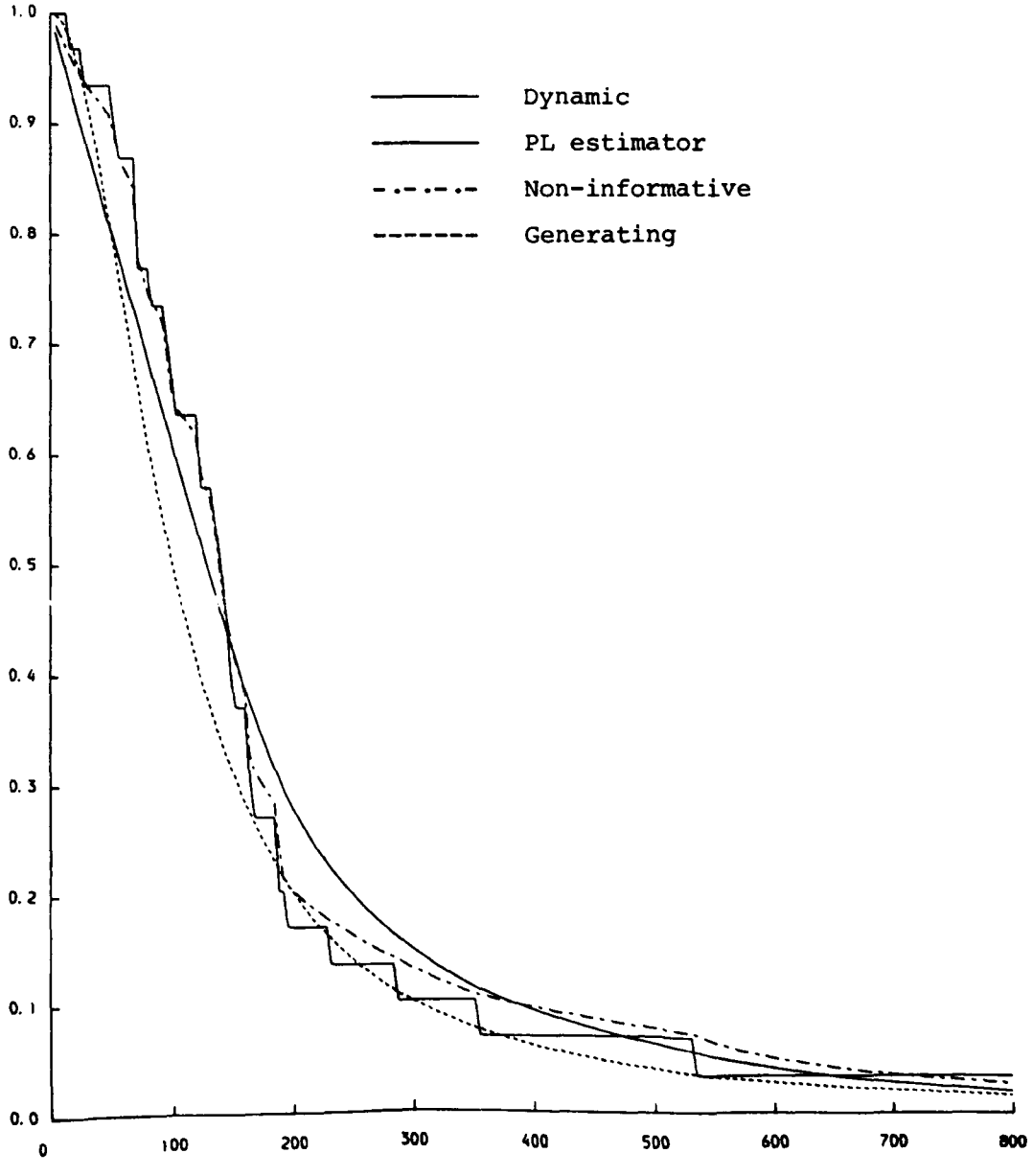


Figure 14(ii). Survival functions for the Logistic (2,100).

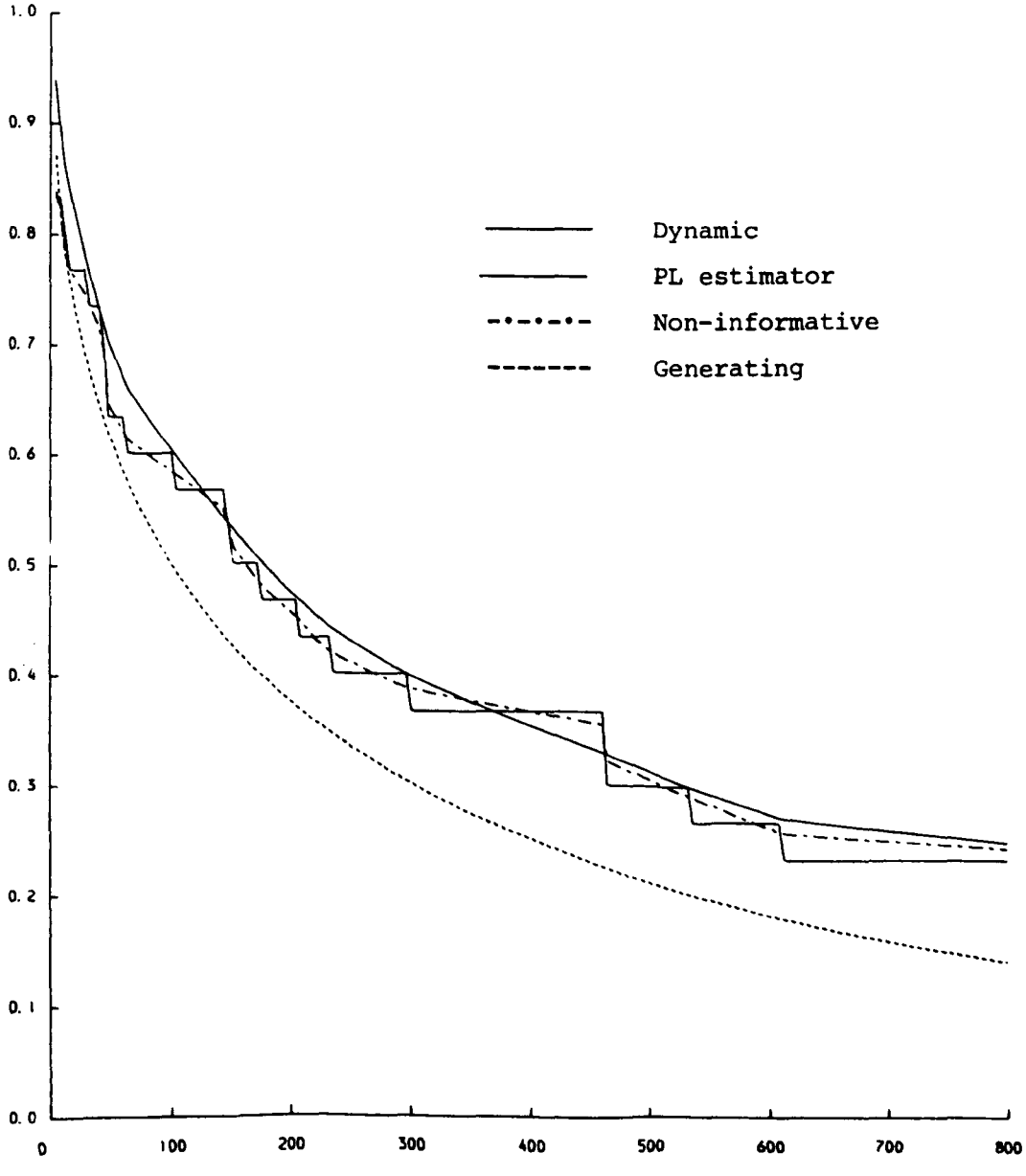


Figure 14(iii). Survival functions for the Weibull (.5,100).

Supposing the practitioner's choice is the simplest model (5.3) with a constant diffusion, he is still left the choice of the grid τ and the value of the diffusion. The predictive likelihood can be used for these model choices. The leukaemia data from Gehan (1965) is used to illustrate this point. They consist of remission times of leukaemia patients in two groups: treatment and control. Both groups have size 21 and although no observation is censored in the control group, there is heavy censoring in the treatment group. This data can be analysed in a two-sample model as in section 4.1 but the two groups are analysed here independently. A number of combinations of values of τ and Q are used for each sample and the corresponding marginal likelihoods are given in Table 8. Once again, the grids τ_{ODT} and of the measurement unit are better than an arbitrary one and the adoption of τ_{ODT} from the onset appears to be a good practical policy. The main feature of the table is the preference manifested for a static model ($Q=0$) in both groups. This represents the model with no parametric evolution, namely the exponential distribution. In this special case, there is no need for a dynamic analysis since $\lambda_i = \lambda, \forall i$ and the posterior distribution is $[\lambda | D_N] \sim G(d, v)$ with

$$d = \sum_{i=1}^{N} d_i \quad \text{and} \quad v = \sum_{i=1}^{N} v_i$$

The predictive survival (5.8) reduces for a single interval $I_1 = [0, t_1]$ to

$$S(y_* | D_N) = \left[1 + \frac{y_*}{v}\right]^{-d}, \quad y_* \in I_1 \tag{5.12}$$

where $t_1 = \max_j y_j$ limits the interval of practical interest. If the model is left with the same N intervals then $[\lambda_i | D_N] \sim G(d, v)$ and the predictive survival is

$$S(y_* | D_N) = \left[1 + \frac{y_* - t_{i-1}}{v}\right]^{-d} \prod_{j=1}^{i-1} \left[1 + \frac{b_j}{v}\right]^{-d}, \quad y_* \in I_i, \; i = 1, 2, \ldots, N \tag{5.13}$$

For the two groups of the leukemia data, the survival (5.12) was compared to (5.13) with the grid τ_{ODT}. The expressions differed by less than .01 for all points, and the majority coincided up to two decimal places.

TABLE 8

(a) Marginal likelihood for treatment group

Data: 6,6,6,6*,7,9*,10,10*,11*,13,16,17*,19*,20*,22,23,25*,32*,32*,34*,35*

Q	τ		
	τ_{ODT}	{5,10,15,20,25,30,35}	{1,2,3,....,34,35}
0	-48.11	-49.26	-49.26
0.05	-48.71	-49.90	-49.75
0.10	-49.13	-50.29	-50.54
0.20	-49.78	-50.85	-50.54

*- censored

(b) Marginal likelihood for control group

Data: 1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23

Q	τ		
	τ_{ODT}	{4,8,12,16,20,24}	{1,2,3,...,22,23}
0	-73.86	-73.86	-73.86
0.05	-74.92	-74.68	-74.54
0.10	-75.68	-75.26	-75.11
0.20	-76.82	-76.09	-76.01

5.2 - Life Tables

The area of life tables can be also treated with the material developed in the previous section. Apart from some specific features of the data structure to be discussed below, life table analysis can be considered as an application of random sample models. Nevertheless, the introduction of life tables is used here to illustrate the systematic exploration of the baseline hazard. This is all there is to be modelled in life table analysis because the individuals are not identified but for their death age interval. In this area, it

is common to refer to the hazard function as the force of mortality. The main objective of this section is to explore the assumptions of a specific force of mortality for a given population. Examples of different parametric forces of mortality presented are:

(a) Gompertz force of mortality- $\mu(t) \quad \alpha \exp(\beta t)$, $t>0$, $\alpha,\beta>0$

(b) Weibull force of mortality- $\mu(t) = \alpha \beta t^{\beta-1}$, $t>0$, $\alpha>0$, $\beta>-1$

(c) Rayleigh force of mortality- $\mu(t) = \alpha+\beta t$, $t>0$, $\alpha,\beta>0$

(d) log-logistic force of mortality- $\mu(t) = \beta / (1+exp(-\alpha-\beta t))$, $t>0$, $\beta>0$

(a)-(d) show that the novelty in the analysis is the use of a specific parametric structure. Since there are no other covariates directly associated with each individuals, the only covariates are the functions of time specifying the form adopted for the force of mortality. On top of that, full use of the structure developed so far can be made by allowing for a piecewise parametric force of mortality where the dependence on time remains the same, though with the parameters changing along the intervals. This can be also extended to different laws in different intervals.

After a short introduction to life tables, inference is outlined for a general force of mortality. Finally, examples are provided for the forces of mortality (a)-(d).

5.2.1- Introduction to life tables

The material of a life table is the lifetime of an individual from a (generally human) population. The first pertinent observation is that the size of the population can be of order of 10^8 in the case of a country's population and therefore some summarisation, preferably in terms of sufficient statistics, is vital. There are two types of life tables:

(a) cohort life tables- reflecting the mortality experience of a group of individuals through their whole lifespan. The only information is the number of deaths and recorded or unrecorded withdrawals in every period of time considered. It could well take 100 years or so to construct such a table. They have a limited application for some controlled populations but it is very difficult to obtain one for a human popula-

tion and even so they are likely to be unreliable.

(b) current life table- reflecting the mortality experience of a group of individuals during one short period of their lifetime, e.g. one year. These are much faster to obtain and far more reliable. The main objective is to give the expected life of an individual at any age, with obvious applications to demography and actuarial science.

Cohort tables are longitudinal studies, whilst current tables are cross-sectional studies. By far the most used are current life tables and these are considered here. They can be presented in two forms: complete life tables where the age intervals have 1 year of length, or abridged life tables where the age intervals are greater than one year. Generally, they have intervals of 5 years apart from the first 5 years divided as 0-1, 1-5 or specifying each of the 5 years separately. While the tables are the final products, what is of interest here is the raw material upon which they are constructed. This is generally set in a table with the following columns:

(I) Age interval- of 1 year for complete tables and generally of 5 for abridged tables with the risky first year of life generally isolated.

(II) Number of deaths in each interval.

(III) Midyear population in each interval- or equivalently, the population at risk in each interval.

(IV) Fraction of last age interval of life- is the average fraction of each interval lived by an individual dying in that interval.

The above fraction is very important for the analysis and it has been observed to be close to .5 for a number of populations in complete and abridge life tables from 5 years on (Chiang, 1968). This shows that on average people die symmetrically within age intervals. In the first 5 years there is a shift towards around .1 for the first year of life and towards .4 for the remaining years representing a strong tendency to die in the beginning of the interval.

Another characteristic of life tables in that summarisation using the age intervals automatically induces the grid $\mathcal{T}$. Also, standard life tables analysis usually specifies intervals open to the right and closed to the left. This has to be modified here to conform with the notation of this thesis but the differences introduced by this change are negligible.

5.2.2- Estimation procedures

Let b_i, c_i, d_i be the length, the midyear population and the number of deaths in interval I_i, $i = 1, 2, ..., N$. Once again, t_N is set to a value above which the number of individuals in the population is very small and the analysis thereafter is of negligible practical interest. For this population there is a total of

$$\sum_{j=i}^{N} (c_j + d_j) = d_i + x_i \quad , \quad x_i = c_i + \sum_{j=i+1}^{N} (c_j + d_j)$$

individuals at risk at the beginning of each interval I_i, $i = 1, 2, ..., N$. Of those, d_i die in I_i and x_i survive through I_i. Finally, let e_i be the fraction of last age interval of life in I_i. In most tables, this information is available. When it is not, it has to be estimated perhaps with the values suggested above. In any case, the total survival time in I_i (TST_i) is

$$\begin{aligned} TST_i &= \text{total survival time of the individuals living through } I_i \\ &\quad + \text{total survival time of individuals dying in } I_i \\ &= x_i\, b_i + d_i\, e_i\, b_i \\ &= b_i\, (x_i + e_i\, d_i) \end{aligned} \tag{5.14}$$

As already mentioned, the main feature of life tables is that individuals are unidentifiable within each age interval. They have therefore the same covariates related to the particular force of mortality model assumed. Before proceeding with the analysis, a result has to be established. It shows that the individual contributions to the likelihood for each interval analysed in any order lead to the same result. It guarantees that there is no order dependence in the estimation procedure if the covariates are the same for all individuals. More important than that, it shows that all the observations can be analysed at once within each interval. This not only saves what may be an enormous amount of computing time, but also makes the analysis possible since rarely are the individual death

times available.

Lemma 5.1. Let $H_1, \ldots, H_k$ be k conditionally independent pieces of survival information yielding the likelihood

$$\prod_{j=1}^{k} \lambda_j^{\chi_j} \exp(-\lambda_j v_j)$$

and guide relations $g(\lambda_j) \simeq \mathbf{z}_j \boldsymbol{\beta}_j$, $j = 1, \ldots, k$ with χ_j = indicator of death in H_j, v_j = survival time in H_j and $\mathbf{z}_j$ = vector of covariates associated with H_j, $j = 1, \ldots, k$. If the system is static ($\boldsymbol{\beta}_j = \boldsymbol{\beta}_{j-1} = \boldsymbol{\beta}$, $j = 1, \ldots, k$) with prior distribution $[\boldsymbol{\beta} \mid H_0] \sim [\mathbf{a}, \mathbf{R}]$ and $[\lambda \mid H_0] \sim G(\alpha, \gamma)$ and the covariates are all the same (leading to a common guide relationship $g(\lambda) = g(\lambda_j) = \mathbf{z}\boldsymbol{\beta}$, $j = 1, \ldots, k$) then

$$[\lambda \mid H_0, H_{j_1}, \ldots, H_{j_k}] \sim [\lambda \mid H_0, H] \sim G(\alpha + d, \gamma + v), \quad d = \sum_{j=1}^{k} \chi_j, \quad v = \sum_{j=1}^{k} v_j$$

and $[\boldsymbol{\beta} \mid H_0, H_{j_1}, \ldots, H_{j_k}] \sim [\boldsymbol{\beta} \mid H_0, H] \sim [\mathbf{m}, \mathbf{C}]$ with

$$\mathbf{m} = \mathbf{a} + \frac{\mathbf{s}}{q}(g - f) \tag{5.15a}$$

$$\mathbf{C} = \mathbf{R} - \frac{\mathbf{s}\mathbf{s}^T}{q}\left(1 - \frac{p}{q}\right) \tag{5.15b}$$

where $H = \{H_1, \ldots, H_k\}$, H_0 stands for previous information, $\{j_i, i = 1, \ldots, k\}$ is any given permutation of $\{1, \ldots, k\}$ denoting the order under which the observations were analysed and g and p are the mean and variance of $[g(\lambda) \mid H_0, H]$.

Proof: If the result is valid for $k = 2$ then it is valid for any finite k by induction. Using the permutation $\{j_1, j_2\}$ and standard results from chapter 5,

$$[\lambda \mid H_0, H_{j1}] \sim G(\alpha + \chi_{j_1}, \gamma + v_{j_1}) \text{ and } [\beta \mid H_0, H_{j_1}] \sim [\mathbf{m}_1, \mathbf{C}_1] \tag{5.16}$$

with

$$\mathbf{m}_1 = \mathbf{a} + \frac{\mathbf{s}}{q}(g_1 - f) \tag{5.17a}$$

$$\mathbf{C}_1 = \mathbf{R} - \frac{\mathbf{s}\mathbf{s}^T}{q}\left(1 - \frac{p_1}{q}\right) \tag{5.17b}$$

where g_1 and p_1 are the mean and variance of $[g(\lambda)|H_0,H_{j_1}]$. Since the model is static, the prior for updating the information from H_{j_2} is (5.16). Let $f_1 = \mathbf{z}\mathbf{m}_1$, $\mathbf{s}_1 = \mathbf{C}_1\mathbf{z}^T$ and $q = \mathbf{z}\mathbf{s}_1$. Then, by Bayes' theorem,

$$[\lambda \mid H_0, H_{j_1}, H_{j_2}] \sim G([\alpha + \chi_{j_1}) + \chi_{j_2}, (\gamma + v_{j_1}) + v_{j_2}) \tag{5.18}$$

and, by linear Bayesian methods, $[\boldsymbol{\beta} \mid H_0, H_{j_1}, H_{j_2}] \sim [\mathbf{m}_2, \mathbf{C}_2]$ with

$$\mathbf{m}_2 = \mathbf{m}_1 + \frac{\mathbf{s}_1}{q_1}(g_2 - f_1) \tag{5.19a}$$

$$\mathbf{C}_2 = \mathbf{C}_1 - \frac{\mathbf{s}_1\mathbf{s}_1^T}{q_1}\left(1 - \frac{p_2}{q_1}\right) \tag{5.19b}$$

where g_2 and p_2 are the mean and variance of $[\lambda \mid H_0, H_{j_1}, H_{j_2}]$.

It is obvious from (5.18) that $[\lambda \mid H_0, H_{j_1}, H_{j_2}] \sim G(\alpha + d, \gamma + v)$ for $d = \chi_{j_1} + \chi_{j_2}$ and $v = v_{j_1} + v_{j_2}$. Also, from (5.17b), $\mathbf{s}_1 = \xi\mathbf{s}$ and $q_1 = \xi q$ with $\xi = p_1/q$. Using this, and replacing the value of $\mathbf{m}_1$ and $\mathbf{C}_1$ in (5.19a) and (5.19b), leads to

$$\begin{aligned}\mathbf{m}_2 &= \mathbf{a} + \frac{\mathbf{s}}{q}(g_1 - f) + \frac{\mathbf{s}}{q}\left\{g_2 - \mathbf{z}\left[\mathbf{a} + \frac{\mathbf{s}}{q}(g_1 - f)\right]\right\} \\ &= \mathbf{a} + \frac{\mathbf{s}}{q}[g_1 - f + g_2 - f - (g_1 - f)] \\ &= \mathbf{a} + \frac{\mathbf{s}}{q}(g_2 - f)\end{aligned} \tag{5.20a}$$

and

$$\begin{aligned}\mathbf{C}_2 &= \mathbf{R} - \frac{\mathbf{s}\mathbf{s}^T}{q}\left(1 - \frac{p_1}{q}\right) - \xi\frac{\mathbf{s}\mathbf{s}^T}{q}\left(1 - \frac{p_2}{\xi q}\right) \\ &= \mathbf{R} - \frac{\mathbf{s}\mathbf{s}^T}{q}\left(1 - \xi + \xi - \frac{p_2}{q}\right) \\ &= \mathbf{R} - \frac{\mathbf{s}\mathbf{s}^T}{q}\left(1 - \frac{p_2}{q}\right)\end{aligned} \tag{5.20b}$$

On the other hand, the lemma assumptions simplify the likelihood to

$$\lambda^{d} \exp(-\lambda v)$$

which updates the parametric distributions to $[\lambda \mid H_0, H] \sim G(\alpha + d, \gamma + v)$ and, via linear Bayesian methods, to $[\beta \mid H_0, H] \sim [\mathbf{m}, \mathbf{C}]$ with $\mathbf{m}$ and $\mathbf{C}$ in (5.15). So $[\lambda \mid H_0, H_{j_1}, H_{j_2}] \sim [\lambda \mid H_0, H]$ and consequently $g_2 = g$ and $p_2 = p$. Comparison of (5.15) with (5.20) shows that $\mathbf{m}_2 = \mathbf{m}$ and $\mathbf{C}_2 = \mathbf{C}$ and the proof is completed.

The above result can be extended to non-linear models with guide relationship $g(\lambda) \simeq h(\mathbf{z}, \beta)$. The only changes introduced are the replacement of f, $\mathbf{s}$ and q by f^*, $\mathbf{s}^*$ and q^* with $f^* = E[h(\mathbf{z}, \beta) \mid H_0]$, $\mathbf{s}^* = Cov[\beta, h(\mathbf{z}, \beta) \mid H_0]$ and $q^* = V[h(\mathbf{z}, \beta) \mid H_0]$. In practice, those values are difficult to obtain and can be approximated in a number of ways suggested in chapter 2. The extension of the lemma to NLM can be obtained following the same parallel derivations between linear and non-linear models developed in section 2.2.

Use of the lemma implies that the likelihood for each interval I_i depends on a single parameter λ_i and can be written with (5.14) as

$$\lambda_i^{d_i} \exp\{-\lambda_i [b_i (x_i + \lambda_i d_i)]\} \tag{5.21}$$

for a linear or a non-linear model.

The main feature of the models of this section is that within each interval the vector of covariates remains the same for all the individuals. They represent the particular dependence on time induced by the force of mortality assumed for the population in that interval. Since the intervals impose a discretisation of time, a suitable time point often chosen to represent an age interval I_i is its mid-interval point $\tau_i = (t_{i+1} + t_i)/2$. The guide relationship for a general non-linear model can be written as

$$g(\lambda_i) \simeq h(\mathbf{z}_i, \beta_i), \quad \mathbf{z}_i \equiv \mathbf{z}_i(\tau_i), \quad i = 1, 2, \ldots, N$$

Assuming that the same force of mortality to hold for the whole lifespan of a table, a full updating cycle goes like this:

(i) $[\boldsymbol{\beta}_{i-1} | D_{i-1}] \sim [\mathbf{m}_{i-1}, \mathbf{C}_{i-1}]$;

(ii) $[\boldsymbol{\beta}_i | D_{i-1}] \sim [\mathbf{a}_i, \mathbf{R}_i]$ with $\mathbf{a}_i = \mathbf{G}_i \mathbf{m}_{i-1}$ and $\mathbf{R}_i = \mathbf{G}_i \mathbf{C}_{i-1} \mathbf{G}_i^T + \mathbf{W}_i$.

(iii) $[\lambda_i | D_{i-1}] \sim G(\alpha_i, \gamma_i)$ with α_i and γ_i such that $E[g(\lambda_i) | D_{i-1}] \simeq f_i$ and $V[g(\lambda_i) | D_{i-1}] \simeq q_i$ with $f_i = E[h(\mathbf{z}_i, \boldsymbol{\beta}_i) | D_{i-1}]$, $\mathbf{s}_i = Cov[\boldsymbol{\beta}, h(\mathbf{z}_i, \boldsymbol{\beta}_i) | D_{i-1}]$ and $q_i = V[h(\mathbf{z}_i, \boldsymbol{\beta}_i) | D_{i-1}]$.

(iv) $[\lambda_i | D_i] \sim G(\alpha_i - d_i, \gamma_i - b_i(z_i + e_i d_i))$, via Bayes' theorem.

(v) $[\boldsymbol{\beta}_i | D_i] \sim [\mathbf{m}_i, \mathbf{C}_i]$ with

$$\mathbf{m}_i = \mathbf{a}_i + \frac{\mathbf{s}_i}{q_i}(g_i - f_i)$$

$$\mathbf{C}_i = \mathbf{R}_i - \frac{\mathbf{s}_i \mathbf{s}_i^T}{q_i}\left(1 - \frac{p_i}{q_i}\right)$$

where $g_i = E[g(\lambda_i) | D_i]$ and $p_i = V[g(\lambda_i) | D_i]$.

The evolution specified in (ii) is general and in practice it can undergo substantial simplifications. If a specific force of mortality is assumed for the whole table, then a static model with $\mathbf{G}_i = \mathbf{I}$ and $\mathbf{W}_i = \mathbf{0}, \forall i$ is appropriate. Sometimes, there may be reasons to assume the same law to hold only locally throughout the table (Elandt-Johnson & Johnson, 1980). In this case, $\mathbf{G}_i = \mathbf{I}, \forall i$ but $\mathbf{W}_i \neq \mathbf{0}$ must be specified to represent changes between successive age groups. The system evolution can be set in a number of ways, in particular, using a discounting approach. All this is already presented in some detail in chapter 3 and put into practice in chapter 4. Also, a non-linear system equation can be used as in chapter 2. The linear case where $h(\mathbf{z}_i, \boldsymbol{\beta}_i) = \mathbf{z}_i \boldsymbol{\beta}_i$ leads to $f_i = \mathbf{z}_i \mathbf{a}_i$, $\mathbf{s}_i = \mathbf{R}_i \mathbf{z}_i^T$ and $q_i = \mathbf{z}_i \mathbf{R}_i \mathbf{z}_i^T$. Specific features of the above estimation cycle for some commonly used forces of mortality are described below.

The analysis proceeds thereafter with smoothing, model checking and prediction for a new individual exactly as before. One of the main interests of a life table is to predict, for any given age, the remaining lifetime of an individual. This is obtained from the

Predictive distribution for a new individual after rescaling the probabilities for the interval of interest, namely from the required time onwards. A summary such as the mean of this conditional distribution can be evaluated. As pointed out at the end of chapter 3, this is highly dependent on the evolution assumptions for the interval I_{N+1}. It is much easier and at least as informative to summarise this information, if at all necessary, through the median of this series of conditional distributions for each age interval.

5.2.3- Examples

Consideration is given to the forces of mortality (a)-(d) either with a static or a piecewise model. The only difference between those two approaches is the fact that $\mathbf{W}_i = \mathbf{0}, \forall i$, in the former. A change of the law governing the force of mortality in the middle of the table is also possible but the corresponding evolution for that interval is entirely subjective. For any force of mortality, analysis of the life table is carried out by discretising the continuous hazard function within the age intervals. Time is generally represented for each age group by the mid-point interval but if one has reasons to believe it not to be the best time point to represent the interval, this could be changed accordingly.

(a) Gompertz and Weibull forces of mortality

In both cases, $g(\lambda_i) = \log \lambda_i$ and $h(\mathbf{z}_i, \boldsymbol{\beta}_i) = \mathbf{z}_i \boldsymbol{\beta}_i$, with $\boldsymbol{\beta}_i^T = (\beta_{i,0}\ \beta_{i,1})$, $i = 1, 2, \ldots, N$. The only difference is that for the Gompertz model, $\mathbf{z}_i = (1\ ,\ \tau_i)$ and for the Weibull model, $\mathbf{z}_i = (1\ ,\ \log \tau_i)$ with τ_i = mid-point of interval I_i, $i = 1, 2, \ldots, N$. The guide relationship $\log \lambda_i \simeq \mathbf{z}_i \boldsymbol{\beta}_i$ has already been extensively studied implying that if $[\boldsymbol{\beta}_i \mid D_{i-1}] \simeq [\mathbf{a}_i, \mathbf{R}_i]$ then $[\boldsymbol{\beta}_i \mid D_i] \simeq [\mathbf{m}_i, \mathbf{C}_i]$ with

$$\mathbf{m}_i = \mathbf{a}_i + \frac{\mathbf{s}_i}{q_i} \log \left[\frac{1 + q_i d_i}{1 + b_i q_i \exp(f_i)(x_i + e_i d_i)} \right] \tag{5.22a}$$

$$\mathbf{C}_i = \mathbf{R}_i - \frac{d_i}{1 + d_i q_i} \mathbf{s}_i \mathbf{s}_i^T \tag{5.22b}$$

The above equations can be seen as a generalisation of (3.20) and are obtained in the exactly the same way. The linear structure for both models is obtained after suitable

reparametrisation in the log scale.

(b) Rayleigh force of mortality

Now, $g(\lambda_i) = \lambda_i$ and $h(\mathbf{z}_i, \boldsymbol{\beta}_i) = \mathbf{z}_i\boldsymbol{\beta}_i$, $\mathbf{z}_i = (1, \tau_i)$, $\boldsymbol{\beta}_i^T = (\beta_{i,0}\ \beta_{i,1})$, $i = 1, 2, \ldots, N$. Given $[\boldsymbol{\beta}_i \mid D_{i-1}] \simeq [\mathbf{a}_i, \mathbf{R}_i]$, the guide relation gives $E[\lambda_i \mid D_{i-1}] = f_i = \mathbf{z}_i\mathbf{a}_i$ and $V[\lambda_i \mid D_{i-1}] = q_i = \mathbf{z}_i\mathbf{R}_i\mathbf{z}_i^T$. So, $[\lambda_i \mid D_{i-1}] \sim G(\alpha_i, \gamma_i)$ with $\alpha_i = f_i^2/q_i, \gamma_i = f_i/q_i$. After updating with (5.21), $[\lambda_i \mid D_i] \sim G[\alpha_i + d_i, \gamma_i + b_i(x_i + e_i d_i)]$ and $[\boldsymbol{\beta}_i \mid D_i] \sim [\mathbf{m}_i, \mathbf{C}_i]$ with

$$\mathbf{m}_i = \mathbf{a}_i - \frac{d_i - f_i\, b_i(x_i + e_i d_i)}{f_i - q_i\, b_i(x_i + e_i d_i)}\,\mathbf{s}_i \tag{5.23a}$$

$$\mathbf{C}_i = \mathbf{R}_i - \frac{2 f_i\, b_i(x_i - e_i d_i) + q_i\, b_i^2 (x_i + e_i d_i)^2 - d_i}{[f_i + q_i\, b_i(x_i + e_i d_i)]^2}\,\mathbf{s}_i\,\mathbf{s}_i^T \tag{5.23b}$$

(c) Log-logistic force of mortality

As a final example, an illustration of a case with no possible linearisation is presented. The best one can do is to assume $g(\lambda_i) = \lambda_i$ and $h(\mathbf{z}_i, \boldsymbol{\beta}_i) = \beta_{i,0}/[1 + \exp(-\mathbf{z}_i\boldsymbol{\beta}_i)]$, $\mathbf{z}_i = (1, \tau_i)$, $\boldsymbol{\beta}_i^T = (\beta_{i,0}\ \beta_{i,1})$, $i = 1, 2, \ldots, N$. Given $[\boldsymbol{\beta}_i \mid D_{i-1}] \sim [\mathbf{a}_i, \mathbf{R}_i]$, the guide relationship induces, as in (b) above, $[\lambda_i \mid D_{i-1}] \sim G(\alpha_i, \gamma_i)$ with $\alpha_i = f_i^2/q_i$ and $\gamma_i = f_i/q_i$. Linear approximations (2.29) lead to $f_i = \dot{h}(\mathbf{z}_i, \mathbf{a}_i)\,\mathbf{a}_i$, $\mathbf{s}_i = \mathbf{R}_i[\dot{h}(\mathbf{z}_i, \mathbf{a}_i)]^T$ and $q_i = \dot{h}(\mathbf{z}_i, \mathbf{a}_i)\mathbf{R}_i$ where $\dot{h}$ is the derivative of h with respect to $\boldsymbol{\beta}$. The values of $\mathbf{m}_i$ and $\mathbf{C}_i$ obtained after updating are the same as in (5.23) but for the values of $f_i, \mathbf{s}_i$ and q_i.

5.3 - Competing Risks

Up until now, it has always been assumed that the only observed random element associated with each individual is its survival time. There are a number of ways that this can be extended. Some of them are presented in the next chapters and involve a generalisation of the theoretical framework developed in this thesis. There is an important practical case, however, that after suitable manipulations can be shown to be nothing but a

special case of survival models. It is the case of competing risks where the extra source of randomness comes from the observation of the cause of death. The observation for each individual involves its (possibly censored) survival time and the single factor that led to the death. This classification factor is one of K possible types of death. There are two ways of analysing this problem. The first involves the assumption of K hypothetical or latent survival times of which only the first one is to be observed while the second one models only the observed quantities, the survival time and its cause. The latter method seems more reasonable and leads to a satisfactory analysis without the unnecessary extra assumptions needed otherwise. The classical analysis for both methods along with some criticism of the latent variables approach are presented in Kalbfleisch & Prentice (1980). It is relevant to this comparison that it is not possible from the data to assert any sort of independence between the different failure types, as will be seen below. Modelling through observables avoids this problem but the latent variable approach invariably ends up having to assume independence. The competing risks approach provides useful models to a number of areas including life tables where interest lies in understanding the different patterns associated with death by cardiovascular diseases, cancer and accidents, say.

Let λ be the (overall) hazard function for the survival time and define

$$\lambda_k(y;z) = \lim_{\Delta_k \to 0} \frac{Pr(Y \in \Delta Y, \delta = k \mid Y > y; z)}{\Delta_k} \quad , k = 1, \ldots, K \tag{5.24}$$

where ΔY is an interval of length Δ_k including y. λ_k can be called, by comparison, the cause-specific hazard function and has the same interpretation of an ordinary hazard function for the particular cause of death considered. The vector of covariates is assumed to be constant through time at this stage. The analysis for time-dependent covariates is a straightforward extension following the material of section 4.2. In association with λ_k, one can also define the cause-specific (sub)density functions as

$$f_k(y) = \lim_{\Delta_k \to 0} \frac{P(Y \in \Delta Y, \delta = k)}{\Delta_k} \quad , k = 1, \ldots, K \tag{5.25}$$

The overall and cause-specific distributions are easily related. In fact, from (5.24),

$$\sum_{k=1}^{K} \lambda_k(y;\mathbf{z}) = \sum_{k=1}^{K} \lim_{\Delta_k \to 0} \frac{P(Y \in \Delta Y, \delta = k \mid Y > y;\mathbf{z})}{\Delta_k}$$

$$= \lim_{\Delta_k \to 0} \frac{\sum_{k=1}^{K} P(Y \in \Delta Y, \delta = k \mid Y > y;\mathbf{z})}{\Delta_k}$$

$$= \lim_{\Delta_k \to 0} \frac{P(Y \in \Delta Y \mid Y > y;\mathbf{z})}{\Delta_k}$$

$$= \lambda(y;\mathbf{z}) \tag{5.26}$$

and, using (1.5),

$$S(y) = \exp\left\{-\sum_{k=1}^{K} \int_0^y \lambda_k(t;\mathbf{z})\,dt\right\} = \prod_{k=1}^{K} \exp\left\{-\int_0^y \lambda_k(t;\mathbf{z})\,dt\right\} \tag{5.27}$$

The subdensity functions cannot be regarded as common densities because for any fixed k they integrate to $P(\delta = k) \neq 1$. They are only densities in the case of one cause of death with probability one. This reduces the problem to the previous set-up. Nevertheless, expression (5.25) gives the exact contribution to the likelihood of an individual observed to die at y of cause k and can be written in terms of the cause-specific hazard functions $\lambda_l, l = 1, \ldots, K$ as

$$S(y) \cdot \lim_{\Delta_k \to 0} \frac{P(Y \in \Delta Y, \delta = k)}{\Delta_k\, P(Y > y)} =$$

$$= \prod_{l=1}^{K} \exp\left\{-\int_0^y \lambda_l(t;\mathbf{z})\,dt\right\} \lim_{\Delta_k \to 0} \frac{P(Y \in \Delta Y, \delta = k \mid Y > y)}{\Delta_k}$$

$$= \lambda_k(y;\mathbf{z}) \prod_{l=1}^{K} \exp\left\{-\int_0^y \lambda_l(t,\mathbf{z})\,dt\right\} \tag{5.28}$$

Also, if the individual is censored at y its contribution to the likelihood is $S(y)$. This can all be put into a general individual contribution

$$\prod_{k=1}^{K} [\lambda_k(y;\mathbf{z})]^{\chi_k} \exp\left\{-\int_0^y \lambda_k(t;\mathbf{z})\,dt\right\} \tag{5.29}$$

where $\chi_k = 1$ if the individual is observed to die of cause k and 0, otherwise. Once again, a time factorisation $\bar{\tau}$ is performed and it is assumed that each cause-specific hazard λ_k is piecewise constant along $\bar{\tau}$, that is,

$$\lambda_k(y;\mathbf{z}) = \lambda_{i,k} \quad , y = I_i \, , \, i = 1,2,...,N \, , \, k = 1,...,K$$

This implies a $PE(\bar{\tau},\boldsymbol{\lambda})$ distribution with $\boldsymbol{\lambda} = (\lambda_1,...,\lambda_{N+1})$, $\lambda_i = \sum_{k=1}^{K} \lambda_{i,k}$, $i = 1,2,...,N,N+1$. Assume a sample of m individuals is observed, each with piecewise constant cause-specific hazard functions $\lambda_k^{(j)}$, $j = 1,...,m$, $k = 1,...,K$. Applying the TFL to the individual contributions (5.29) leads to the likelihood factors l_i given by

$$\prod_{j=1}^{r_i} \prod_{k=1}^{K} \lambda_{i,k}^{(j)\chi_{ijk}} \exp\left\{-\lambda_{i,k}^{(j)}(y_{ij} - t_{i-1})\right\} \quad , i = 1,2,...,N \tag{5.30}$$

where $\chi_{ijk} = 1$ if individual j dies in I_i of cause k and 0, otherwise. (5.30) is a direct generalisation of (3.19). Observe that although (5.30) and (5.29) are factorised in terms of the different causes of death, no independence assumption has been made nor has been imposed any relation between the different λ_k's. The model is completed after association of the cause specific hazard functions with the relevant covariates in each case. Suppose that for each individual a set of p_k covariates $\mathbf{z}_k$ is assumed relevant for the hazard related to the cause $k, k = 1,...,K$. Let $\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K$ be the respective parameters modelling the effect of $\mathbf{z}_1,...,\mathbf{z}_K$. In order to put into the structure of chapter 3, the parameter $\boldsymbol{\beta}_i$ of dimension $p = p_1 + ... + p_K$ is formed by joining all the $\boldsymbol{\beta}_{i,k}$'s for all k and $\mathbf{z}_{j,k}$ is extended accordingly to a vector of K components each having p_k 0's apart from the k^{th} component filled with the appropriate values of the corresponding covariates. The guide relationship generalises to

$$\log \lambda_{i,k}^{(j)} \simeq \mathbf{z}_{j,k}\boldsymbol{\beta}_i \quad , i = 1,2,...,N \, , \, j = 1,...,r_i \, , \, k = 1,...,K \tag{5.31}$$

The model is completed as before with a partial prior for $\boldsymbol{\beta}_i$ based on its mean, variance and Gamma conjugate prior for the $\lambda_{i,k}^{(j)}$'s. The relationships between different causes is established through the sequential estimation for the $\boldsymbol{\beta}_i$'s.

Apart from the usual interest in relating covariates to the relative risks of death, an important issue in competing risks is the study of relationships between different causes of death. In many cases, interest lies in asserting the influence that one cause of death has in the other ones. Even though no relation between the cause-specific hazard functions was made, the likelihood factors (5.30) factorise across causes. No information about the relationship between different causes can therefore be obtained solely from data. Once again, the Bayesian approach can prove useful in tackling this problem by suitable specification of subjective inputs. Suppose that the prior specification sets $V[(\beta_{0,1}^T \cdots \beta_{0,K}^T) \mid D_0] = diag(\mathbf{C}_{0,1}, \ldots, \mathbf{C}_{0,K})$ representing a prior judgement of no correlation between the different causes of death. If, in addition, the evolution is specified similarly with $\mathbf{W}_t = diag(\mathbf{W}_{t,1}, \ldots, \mathbf{W}_{t,K})$ where each $\mathbf{W}_{t,k}$ is a $p_k \times p_k$ matrix, $k = 1, \ldots, K$ then the parameters will keep the uncorrelated structure through time. If, however, the evolution is specified in another way, typically non-block diagonal, then interrelation between the different sets of parameters will appear. Even though the likelihood still factorises for different causes, the relationship implied on the parametric structure by the system evolution adopted will interfere and influence the whole inference procedure. Comparison between an independence model and a general one can be made and assessed as before with the Bayes factor.

This assessment is made here with the breast cancer data introduced by Boag (1949) and reproduced in Table 9. It is the simplest of cases with only 2 causes considered but gives a useful insight into the problem.

There are a number of ways to model this data set but since the only information given is the state of the patient at its last observed time, the two, somewhat arbitrary, categories emerge. It is assumed therefore that the other causes group constitutes a different risk pattern even though some patients in this group had at some stage suffered from breast cancer. No independent variables are provided to explain the survival pattern hence there are only two parameters, each modelling one cause-specific hazard. For clarity of exposition, it is assumed that $\mathcal{T}$ = {observed death times} and that the system disturbances are

TABLE 9

Survival times (in months) of breast cancer patients

Cancer present	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, 111*, 112*, 113*, 114, 114*, 114*, 117*, 121*, 123*, 126, 129*, 131, 131*, 133*, 134*, 134*, 136*, 141*, 143*, 167*, 174, 177*, 179*, 189*, 201*, 203*, 203*, 213*
Other causes	0.3, 4, 7.4, 15.5, 23.4, 46, 46, 51, 65, 68, 83, 88, 96, 110, 111, 112, 132, 162, 228*

*- censored.

driven by a constant diffusion **Q**. The system equation is then

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}_i = \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}_{i-1} - \begin{pmatrix}w_1\\ w_2\end{pmatrix}_i \quad , \quad \begin{pmatrix}w_1\\ w_2\end{pmatrix}_i \sim \left[\begin{pmatrix}0\\ 0\end{pmatrix}, b_i \begin{pmatrix}q_1 & q_{12}\\ q_{12} & q_2\end{pmatrix}\right]$$

and analysis starts with $[(\beta_1\ \beta_2)_0^T \mid D_0] \sim [\mathbf{O}, 10^3 \mathbf{I}_2]$. As previously pointed out, $q_{12} = 0$ implies no relation between the two groups or causes of death. This class of models has only 2 parameters to be specified, q_1 and q_2 while the general class including relationship between the two failure patterns requires the specification of an extra element, q_{12}. Considering first the model with no relation between the two causes, the diffusion with $q_1 = q_2 = .01$ gives a predictive likelihood of $\exp(-576.08)$. This is an improvement on the model with no evolution ($q_1 = q_2 = 0$) for which this value was $\exp(-582.89)$. The model assuming correlation between the two causes allow for q_{12} to be different from 0. A number of values were tried and the results are given in Table 10.

TABLE 10
Predictive likelihood values
for models with relation between the causes

q_{12}	-.01	-.008	-.005	0	.005	.01
log l	-575.60	-575.54	-575.71	-576.08	-576.62	-578.84

Although the evidence is not conclusive, probably due to the small number of deaths in the second group, it suggests negative correlation between w_1 and w_2. There is not an overwhelming indication of relation between the causes for this data set but this could well be the case for some other large data sets. This sort of analysis can also prove useful when, in the presence of several causes of death, interest lies in establishing groups within each one cause affects the outcome of the others but outside which the different causes have no relation, acting independently.

References

Boag, J. M. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy (with discussion). *J. R. Statist. Soc., Ser. B,* **11**, 15-53.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, **2**, 1152-1174.

Chiang, C. L. (1968). *Introduction to Stochastic Processes in Biostatistics.* Wiley: New York.

Elandt-Johnson, R. C. and Johnson, N. L. (1980). *Survival Models and Data Analysis.* Wiley: New York.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209-230.

Gamerman, D. (1987). Dynamic inference on survival functions. In *Probability and Bayesian Statistics* (ed. R. viertl). Plenum: New York.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika,* **52**, 203-223.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley: New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Ass.*, **53**, 457-481.

Martz, H. F. and Waller, R. A. (1982). *Bayesian Reliability Analysis*. Wiley: New York.

Mashhoudy, H. (1985). Bayesian Reliability Growth Analysis. Unpublished Ph.D. thesis, University of Warwick.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *J. Amer. Statist. Ass.*, **71**, 897-902.

CHAPTER SIX

ANALYSIS OF POINT PROCESSES

Up until now, the nature of the observation process studied has been that of standard survival analysis. Data have consisted of many observation units (called individuals) for each of which a single occurrence (called death) takes place with nothing further to be observed thereafter for that individual. Apart from the possibility that some observations could be censored, the only change considered was in the previous section where the cause of the occurrence had some importance.

In this chapter, a dual observation process is considered. Namely, that in which there is only one observation unit from which many successive occurrences are observed. The use of terms from medicine to describe it is no longer useful (since only one death can take place for an individual!). The area to be used as illustration now is reliability and the labels individual, death and survival time are replaced by component, failure and failure time. In particular, software reliability is considered where the component is a software programme (or system) being tested for correction of errors or bugs.

The chapter is outlined as follows. The next section provides a brief introduction to the point processes and presents the main results to be used after which the relation to the previous chapters is clear. The next sections comment on particular cases in different directions. The first one deals with the analysis in the absence of covariates and presents the derivation of the extension to continuous time estimation. Then, marked point processes are introduced. They constitute processes where there is an additional random variable associated with every failure taking place. Finally, a selection of models that have been put forward for the analysis of software reliability is presented and each is shown to belong to the class of dynamic models of this thesis.

Generalisations to the observation of many components and towards multiple failures with competing risks are left for the final chapter.

6.1 - Introduction

Point process is the name given to the observation process where successive occurrences take place. These occurrences can take place in any space but the processes considered here are the semi-infinite line representing time. They will be considered as the failures taking place in the life of a component. In this context, such a process can be represented as the above line with superimposed points at the times these failures take place. The process is then completely characterised by the times $Y_1, Y_2, \cdots$ between successive failures or by the cumulative times $T_1 = Y_1$, $T_2 = Y_1 + Y_2$, $\cdots$ of the failures. By emphasising the counts of the numbers of failures in different intervals, one has an alternative way of characterising these process. When viewed in this way the process is called a counting process. Defining $N(s,t]$ as the number of points in the interval $(s,t]$, one can easily establish the equivalence of the two approaches through

$$N(t) = N(t_0, t) = \begin{cases} 0\,, & t \in [t_0, T_1) \\ 1\,, & t \in [T_1, T_2) \\ \cdot & \\ i\,, & t \in [T_{i-1}, T_i) \\ \cdot & \end{cases} \tag{6.1}$$

where t_0 is the time origin as usual taken as 0 unless otherwise stated.

A convenient starting point to the study of the point processes is through Poisson processes. A non-homogeneous Poisson process (NHPP) can be defined as a counting process $N(t)$ such that:

a) $Pr(N(t_0) = 0) = 1$;

b) $N(s,t] \sim$ Poisson $(\Lambda(t) - \Lambda(s))$ where Λ is a non-negative, non-decreasing function of t ;

c) $N(t)$ has independent increments (Snyder, 75, pg.38).

A few points have to be made about the above conditions. The first one is introduced

to ensure no failures at the time origin. Condition (b) allows the possibility that Λ may be discontinuous, although, in most cases, reasonable descriptions of phenomena can be made with a differentiable Λ. Then one can obtain the function λ given by the derivative of Λ and called the intensity rate of the process.

It may be useful, however, to extend existence of λ to a larger class of functions. For continuous, non-differentiable functions Λ, λ can be defined as

$$\lambda(t) \quad \lim_{\Delta t \to 0^+} \frac{\Lambda(t) - \Lambda(t-\Delta t)}{\Delta t} \tag{6.2}$$

Observe that for the times where Λ is differentiable, this coincides with the derivative. Otherwise, any arbitrary value of λ can be used but retaining (6.2) makes it fall in line with the notation adopted in this thesis as will be made clear later in the section. Even for jump points of Λ, the Dirac delta function δ could be used conceptually so that if $\Lambda(u^+) = c + \Lambda(u), c > 0$, for some $u > 0$, then a factor $c\delta(t-u)$ is added to λ. Discontinuity of Λ can be useful for processes where more than one failure can occur at the same time. Observe from (ii) that if $\Lambda(t^+) - \Lambda(t) = c \neq 0$ then $Pr(N(t, t+\epsilon) > 1) > 0$, $\forall \epsilon > 0$.

In fact, condition (b) can be replaced by

(b') $N(s,t] \sim$ Poisson $\left(\int_s^t \lambda(u)du \right)$ where λ is a non-negative function of t.

From now on, I shall be using the intensity rate λ to define a process rather than Λ. The relation between intensity rates of processes and hazard functions is given below.

The final condition states that for $s_1 < s_2 \leq t_1 < t_2$, $N(s_1,s_2]$ and $N(t_1,t_2]$ are independent as can be recalled from the definition in section 2.3. Finally, the non-homogeneity comes from the fact that $\lambda(t)$ is not constant. In the homogeneous case, the intensity $\lambda(t) = \lambda$ is the same for all time points characterising an homogeneous process since the distribution of counts over intervals of equal length is the same.

Once again, the relation between counting and point processes can be established and NHPP be defined through the failure times. Condition (c) implies that the Y_i's are independent. From (b),

$$S_i(t) = Pr(Y_i > t) = Pr(N(T_{i-1}, t] = 0) = \exp\left\{ - \int_{T_{i-1}}^{t} \lambda(s)ds \right\} \quad , t > T_{i-1} \qquad (6.3)$$

Where $T_j = \sum_{k=1}^{j} Y_k$. Comparing (6.3) with (1.5), it is clear that λ is the hazard function of Y_i for $t > T_{i-1}$, $i = 1,2,\ldots$ So, a NHPP is a series of non-negative independent random variables with hazard functions $\lambda_i(t), t > 0$, $i = 1,2,\ldots$ The intensity rate of such process is

$$\lambda(t) = \begin{cases} \lambda_1(t), & t \in (t_0, Y_1] \\ \lambda_2(t), & t \in (Y_1, Y_1 + Y_2] \\ \cdot \\ \cdot \\ \cdot \end{cases}$$

A useful NHPP for the purpose of this thesis is the piecewise homogeneous Poisson process (PHPP). This is a NHPP with intensity rate given by

$$\lambda(t) = \begin{cases} \lambda_1, & t \in (t_0, t_1] \\ \lambda_2, & t \in (t_1, t_2] \\ \cdot \\ \cdot \\ \cdot \end{cases} \qquad (6.4)$$

The value of the grid $\tau = \{t_1, t_2, \cdots\}$ of non-decreasing times is completely arbitrary and need not coincide with the observed values $\{y_1, y_1 + y_2, \cdots\}$ but it may if so desired by the modeller. It should contain the points at which the rate of occurrence of failures changes.

The likelihood factors $\{l_i\}$ associated with the interval I_i must be obtained to perform a dynamic analysis. In the case of NHPP, this task is simplified by condition (c) stating that the events taking place in I_i do not depend on past information, D_{i-1}. The total likelihood is given by

$$\prod_{i=1}^{n} \lambda_i^{n_i} \exp\{-\lambda_i b_i\} \qquad (6.5)$$

Where n_i is the observed value of $N(t_{i-1},t_i]$ and $b_i = length\,(I_i)$, as usual. See Snyder (1975, pg. 63) for a rigorous derivation of (6.5). Definition (6.2) ensures that failures occurring at t_i contribute to the likelihood with a factor of λ_i associated with interval I_i. The factorisation of (6.5) in terms only involving the parameters associated with each interval is very appropriate. It reflects previous results from chapter 3 with the advantage now that only one parameter is associated with each likelihood term. The fact that Poisson processes are constructed with independent increments is now an unnecessary restriction to the models. The dynamic approach takes naturally account of past information and a larger class of models can now be introduced. This is the class of self-exciting processes, built over the structure of Poisson processes by dropping condition (c) and allowing the intensity rate λ to depend on previous information. Analogous to the PHPP are the piecewise self-exciting processes (PSEP) which have intensity rate (6.4) but with the λ_i's now explicitly dependent on the corresponding D_{i-1}'s. Observe that the PSEP cover a large class of models and are quite flexible through changes in τ. From now on, those are the processes to be considered in this chapter unless otherwise stated. It is immediate to see that they still lead to the likelihood (6.5).

If the time grid is determined by the successive failure times then the likelihood in (6.5) simplifies to

$$\prod_{i=1}^{r} \lambda_i^1 \exp\{-\lambda_i y_i\}$$

where the power 1 in λ_i could be replaced by $dn(t)$ where $dn(t)$ is the observed value of $N(t+\epsilon) - N(t)$ when $\epsilon \to 0^+$ if one allows for concomitant failures. When the λ's are related to covariates associated associated with each failure time, then the time grid must obviously contain the times failures occur.

In the next section, the analysis is performed for the case where no covariates are present. The equations for the case where the grid τ of observation points is extended to the continuum are obtained with the use of a limiting argument.

6.2 - The no-covariates case: discrete and continuous estimation

Reliability models commonly used to date rarely consider relevant additional information associated with the failure times. In such cases, the analysis can be performed in terms of the λ's without reference to any structured system equation as in previous chapters. This is because the main interest lies in the prediction of future failure times rather than in smoothing. Even in cases when there is some information available to build more elaborate models, the analysis based solely on failure times can prove revealing and may be useful as a preliminary analysis.

For these reasons, the no-covariates case seems a good starting point in the build-up to more complex models in the subsequent sections. Assuming a grid $\mathcal{T} = \{t_1, t_2, \cdots\}$ of times and a piecewise homogeneous self-exciting (or Poisson) process, the likelihood can be factorised into terms l_i where

$$l_i = \lambda_i^{n_i} \exp\{-\lambda_i b_i\} \tag{6.6}$$

as previously specified.

The system equation is much the same as in section 5.1. Assume that given the information up to t_{i-1}, λ_{i-1} has distribution $[\lambda_{i-1} | D_{i-1}] \sim G(\alpha_{i-1}, \gamma_{i-1})$. Then referring to an evolution in the log scale $\beta = \log \lambda$, one has $[\beta_{i-1} | D_{i-1}] \sim [m_{i-1}, C_{i-1}]$ with $m_i = \log(\alpha_{i-1} / \gamma_{i-1})$ and $C_i = \alpha_{i-1}^{-1}$. Assuming an evolution with constant diffusion, $[\beta_i | D_{i-1}] \sim [m_{i-1}, C_{i-1} + b_i Q]$ for some $Q > 0$ and $b_i = \text{length}(I_i)$, as usual. This implies, again as usual, that $[\lambda_i | D_{i-1}] \sim G((C_{i-1} + b_i Q)^{-1}, (C_{i-1} + b_i Q)^{-1} \exp(-m_{i-1}))$. Substituting the values of m_{i-1} and C_{i-1} and after simple manipulations,

$$[\lambda_i | D_{i-1}] \sim G(\phi_i \alpha_{i-1}, \phi_i \gamma_{i-1}) \tag{6.7}$$

with

$$\phi_i = (1 + b_i Q \alpha_{i-1})^{-1} \tag{6.8}$$

This leads to preservation of means and inflation of variances in the same way as obtained in section 5.1. Also, (6.7) can be updated through the likelihood (6.6) giving

$$\lambda_i \mid D_{i-1} \sim G(\phi_i \alpha_{i-1} + n_i, \phi_i \gamma_{i-1} + b_i) \tag{6.9}$$

from where the relevant equations for smoothing and model criticism are as in section 5.1. The attention here however is in predicting the next failure time Y from the present time, t_N. Assume that $\lambda_N \mid D_N \sim G(\alpha_N, \gamma_N)$ and $\tau = \{t_1, t_2, ..., t_N, t_{N+1}, t_{N+2}, ...\}$ with corresponding piecewise constant values of the intensity rate. Once again, one can factorise the distribution of Y as follows

$$p(y \mid D_N) = p(y \mid Y > t_{N+i-1}, D_N) \prod_{j=1}^{i-1} S(t_{N+j} \mid Y > t_{N+j-1}, D_N) \tag{6.10}$$

for $y \in I_{N+i} = (t_{N+i-1}, t_{N+i})$, $i = 1, 2, ...$. Each factor in (6.10) is obtained as in section 3.6.1 through the following cycle:

i) set $j = 1$;

ii) obtain, through evolution, $[\lambda_{N+j} \mid D^*_{N+j-1}] \sim G(\alpha_{N+j}, \gamma_{N+j})$ with $D^*_{N+j} = D_N \cap [Y > t_{N+j}]$ and $D^*_N = D_N \cap [Y > t_N] = D_N$;

iii) obtain, by integration, $p(y \mid D^*_{N+j-1})$ for $y \in I_{N+j}$ as

$$\int p(y \mid \lambda_{N+j}, D^*_{N+j-1}) f(\lambda_{N+j} \mid D^*_{N+j-1}) d\lambda_{N+j} =$$

$$= \left[\frac{\alpha_{N+j}}{\gamma_{N+j} + y - t_{N+j-1}} \right]^{\chi_{N+j}} \left[1 + \frac{y - t_{N+j-1}}{\gamma_{N+j}} \right]^{-\alpha_{N+j}} \tag{6.11}$$

where χ_{N+j} is the occurrence indicator (1 when the required quantity is the density and 0 when it is the survival function);

iv) update, by Bayes' theorem,

$$f(\lambda_{N+j} \mid D^*_{N+j}) \propto f(\lambda_{N+j} \mid D^*_{N+j-1}) S(t_{N+j} \mid \lambda_{N+j}, Y > t_{N+j-1})$$

to obtain $[\lambda_j \mid D^*_{N+j}] \sim G(\alpha_{N+j}, \gamma_{N+j} + b_{N+j})$;

v) add 1 to j and return to ii);

Substitution of (6.11) gives

$$p(y \mid D_N) = \left[\frac{\alpha_{N+i}}{\gamma_{N+i} + y - t_{N+i-1}} \right]^{\chi_{N+i}} \left[1 + \frac{y - t_{N+i-1}}{\gamma_{N+i}} \right]^{-\alpha_{N+i}} \prod_{j=1}^{i-1} \left[1 + \frac{b_{N+j}}{\gamma_{N+j}} \right]^{-\alpha_{N+j}} \tag{6.12}$$

for $y \in I_{N+1}$.

This procedure can be, of course, used retrospectively to already observed failure times providing yet another test for the model. One would expect a good model to have either its predicted mode or median around the observed failure time. The median and the mode can be easily (and visually) picked up in a graphical display of (6.12) for the density and the survival functions respectively. Adjustments to the model can be made in a number of ways like changing the grid or by allowing a non-constant diffusion Q. The simplest and most effective change is to assume a piecewise constant diffusion Q given by Q_i in each interval I_i, $i = 1,2,\ldots$. The only correction to be made by this change is in (6.8) given now by $\phi_i = 1 + b_i Q_i \alpha_{i-1}$.

Poisson processes and their generalisations have been widely studied in both applied and theoretical statistics. Of particular interest is the problem of obtaining good estimates of the intensity rate. The relation to Bayesian statistics and this thesis appeared in the study of doubly stochastic Poisson processes. This term was coined by Cox (1955) to designate NHPP whose intensity is assumed to be randomly generated. A review of recent work in this area can be found in Grandell (1976). In all cases, a continuous estimation of $\lambda(t)$ given information up to a time t is sought. This invariably leads to complicated integral equations that the above approach based on a discretisation avoids. It is a useful exercise, however, to show how it can be extended to cope with continuous estimation of the intensity rate. This will be done in two equivalent ways. The first one through a heuristic limiting derivation and the second is a constructive deduction as in Gill (1984).

The main idea is to take an interval, say $I_i = (t_{i-1}, t_i]$, and observe its behaviour as $t_i - t_{i-1} = \Delta t \to 0^+$. To be more specific, the limiting procedures will be carried out from the left to make full use of the nice continuity properties. In doing so, the notation will be changed from λ_j, D_i, α_k, and so forth to $\lambda(t_j), D(t_i), \alpha(t_k)$ so that the times associated with the quantities used are clear. One has from (6.7), (6.8) and (6.9) that for a small Δt

$$[\lambda(t_{i-1}) \mid D(t_{i-1})] \sim G(\alpha(t_{i-1}), \gamma(t_{i-1}))$$

$$[\lambda(t_i) \mid D(t_i)] \sim G(\alpha(t_i), \gamma(t_i))$$

Where

$$\alpha(t_i) = [1 + \Delta t \cdot Q(t_i)\alpha(t_{i-1}) + o(\Delta t)]^{-1}\alpha(t_{i-1}) + n(t_{i-1}, t_i] \tag{6.13a}$$

$$\gamma(t_i) = [1 + \Delta t \cdot Q(t_i)\alpha(t_{i-1}) + o(\Delta t)]^{-1}\gamma(t_{i-1}) + \Delta t \tag{6.13b}$$

The inclusion of the term $o(\Delta t)$ is due to

$$\int_{t_{i-1}}^{t_i} Q(s)\,ds \quad \Delta t \cdot Q(t_i) + o(\Delta t)$$

Where the diffusion Q is assumed to be at least left continuous. Equations (6.13) can be rewritten as

$$\alpha(t_i) - \alpha(t_i - \Delta t) = -\frac{\Delta t\, Q(t_i)\alpha^2(t_{i-1})}{1 - \Delta t\, Q(t_i)\alpha(t_{i-1}) + o(\Delta t)} + n(t_{i-1}, t_i] + o(\Delta t) \tag{6.14a}$$

$$\gamma(t_i) - \gamma(t_i - \Delta t) = -\frac{\Delta t\, Q(t_i)\alpha(t_{i-1})\gamma(t_{i-1})}{1 - \Delta t\, Q(t_i)\alpha(t_{i-1}) + o(\Delta t)} + \Delta t + o(\Delta t) \tag{6.14b}$$

Before going into formal calculations, one should observe from (6.14) that, as $\Delta t \to 0^+$, $\lim \gamma(t_i - \Delta t) = \gamma(t_i)$ but $\lim \alpha(t_i - \Delta t) = \lim n(t_i - \Delta t, t_i] = dn(t_i)$ so that γ is continuous but α jumps every time a failure occurs hence the notation $\alpha(t) = \alpha(t^-) + dn(t)$. Formal derivatives of α are not available then without use of Dirac delta functions. Taking limits as $\Delta t \to 0^+$ in (6.14) gives the differential forms

$$d\alpha(t) = -Q(t)[\alpha(t)]^2\, dt + dn(t) \tag{6.15a}$$

$$d\gamma(t) = -Q(t)\alpha(t)\,\gamma(t)\,dt + dt \tag{6.15b}$$

The solution to (6.15) gives the parameters of the Gamma distribution of $[\lambda(t) \mid D_t]$, for every t.

Another simple way of obtaining the same result is as follows:

i) Assume that the intensity rate $\lambda(t)$ of a process has stochastic evolution in time given by

$$[\lambda(t) \mid D(t^-)] \sim G(\alpha(t^-), \gamma(t^-))$$

where $D(t^-)$ is the information collected up to time $t-dt$, that is, just before t and α and γ satisfy the differential from

$$d\alpha(t) = -Q(t)\alpha^2(t)\,dt \tag{6.16a}$$

$$d\gamma(t) = -Q(t)\alpha(t)\gamma(t)\,dt \tag{6.16b}$$

for some positive function Q.

ii) Obtain the likelihood for interval $(t-dt, t]$ conditional on $D(t^-)$ as

$$\lambda(t)^{dn(t)}\ \exp(-\lambda(t)\,dt)$$

iii) Update the distribution of $[\lambda(t)\,|\,D_t] \sim G(\alpha(t),\gamma(t))$ via Bayes' theorem with α and γ now given by

$$d\alpha(t) = -Q(t)[\alpha(t)]^2\,dt + dn(t) \tag{6.17a}$$

$$d\gamma(t) = -Q(t)\alpha(t)\gamma(t)\,dt + dt \tag{6.17b}$$

The evolution adopted in (i) was, of course, tailor-made to make the resulting equations match (6.15). Other forms could be used as well leading to a variety of other differential forms. One may try to introduce a discount function β in the λ scale such that $[\lambda(t)\,|\,D(t^-)] \sim G(\alpha(t^-),\gamma(t^-))$ where

$$d\alpha(t) = \alpha(t)-\alpha(t-dt) = [\beta(t)]^{dt}\ \alpha(t-dt) = [1 + dt\log\beta(t) + o(dt)]\,\alpha(t-dt)$$

$$d\gamma(t) = \gamma(t)-\gamma(t-dt) = [\beta(t)]^{dt}\ \gamma(t-dt) = [1 + dt\log\beta(t) + o(dt)]\,\gamma(t-dt)$$

The likelihood factor in (ii) was obtained using the fact that $N(t-dt, t]$ conditional on $D(t-dt) = D(t^-)$ has distribution Poisson $(\lambda(t)dt)$, from condition (b').

Assuming that $t_1, t_2, \cdots$ are the observed cumulative times of failure, equations (6.15) can be solved recursively within each interval (t_{i-1}, t_i) and then adding the jump $dn(t_i)$ before moving to the next interval. In the case Q is constant, these are

$$\alpha(t) = [\psi_i(t)]^{-1}\ \alpha(t_{i-1}) + dn(t) \tag{6.18a}$$

$$\gamma(t) = [\psi_i(t)]^{-1}\ [\gamma(t_{i-1}) + t - t_{i-1} + \alpha(t_{i-1})\,Q(t-t_{i-1})^2/2] \tag{6.18b}$$

where

$$\psi_i(t) = 1 + \alpha(t_{i-1})\,Q\,(t-t_{i-1})$$

for $t \in (t_{i-1}, t_i]$, $i = 1,2,....$ So, α is a strictly decreasing function of time but for the jump points. Taking the coefficient of variation which in the Gamma case is given by $\alpha^{-1/2}$ as a measure of information in a distribution, one can see that the only times information arrives are the failure times. Also, this continuous approach can be compared with a discrete one when t reaches the end point of the interval, t_i. In this case, comparison between (6.18a) and (6.9) shows that the increment in the value of α is the same. Substitution of $t_i = b_i - t_{i-1}$ in (6.18b) however gives

$$\gamma(t_i) = \frac{\gamma(t_{i-1})}{1 + \alpha(t_{i-1})\, Q\, b_i} + \left[\frac{1 + \alpha(t_{i-1})\, Q\, (b_i/2)}{1 + \alpha(t_{i-1})\, Q\, b_i} \right] b_i$$

which is slightly different from the equivalent value of γ_i in (6.9). The ratio between (6.18a) and (6.18b) provides the on-line mean trajectory of the intensity rate through time. This trajectory can be plotted providing a useful display.

Finally, it should be stressed that the differential equations obtained through this continuous estimation procedure can be analytically solved even for arbitrary non-constant diffusion Q as long as the product $Q\,\alpha$ is analytically integrable. This is a great advantage over other methods that have been proposed in the past.

6.3 - Marked point processes

Another extension for Poisson and self-exciting processes is towards the so called marked point processes (MPP). These are points processes that have associated with each failure an additional random variable called mark. The marks need not be independent from each other nor from the underlying point process $N(t)$.

In this section, I will assume, as usual, that the basic point process is a PSEP and that the time grid $\mathcal{T}$ consists of the cumulative failure times. This choice of $\mathcal{T}$ is adopted only to avoid unnecessary notational complications and the analysis can be easily extended to other grids.

In performing the analysis for MPP, it is assumed that there is additional information available associated with the underlying PSEP and with the marks themselves. This

information is conveyed in the form of covariates and it could well be the case of some covariates being explanatory variables for both PSEP and the marks.

Assume the marks to be $X_1, X_2, \ldots$ and each of them with density $f(x_i \mid \theta_i, D_{i-1}, n_i)$ where θ_i is the single parameter summarising the distribution of X_i, $i = 1,2,\ldots$ and the distribution is conditional not only on previous information but also on the events observed for the point processes through I_i.

The TFL can again be evoked to factorise the likelihood l into $\prod l_i$. Each likelihood factor l_i provides information on the underlying PSEP and on the mark X_i taking place in I_i. It is given by

$$\lambda_i^{n_i} \exp(-\lambda_i b_i) \cdot f(x_i \mid \theta_i, D_{i-1}, n_i) \tag{6.20}$$

where n_i = observed value of $N(t_{i-1}, t_i] = 1$, by construction of $\mathcal{T}$ and $b_i = y_i$, $i = 1,2,\ldots$. As previously mentioned, the intensity rate can be related to a set of variables, say $\mathbf{z}_I$, but also the parameter of the distribution of the marks can be related to another set of covariates, say $\mathbf{z}_{II}$. These two sets of covariates may also have some intersection for it is feasible that some variables may be explanatory for the underlying process and the mark variable. One then proceeds by establishing the guide relationships

$$\lambda_i \simeq g_I(\mathbf{z}_{I,i}, \boldsymbol{\beta}_{I,i})$$

$$\theta_i \simeq g_{II}(\mathbf{z}_{II,i}, \boldsymbol{\beta}_{II,i})$$

for $i = 1,2,\ldots$. In view of the fact that there may be intersection between $\mathbf{z}_I$ and $\mathbf{z}_{II}$, it is advisable to merge $\mathbf{z}_I$ and $\mathbf{z}_{II}$ into $\mathbf{z}$ and, equivalently, $\boldsymbol{\beta}_I$ and $\boldsymbol{\beta}_{II}$ into $\boldsymbol{\beta}$. The model is completed by a system equation like (3.18). With the system equation, evolution can be performed to construct the prior moments for $\boldsymbol{\beta}$. Then suitable prior distributions for λ and θ are constructed using the corresponding guide relationships and updated as before. A couple of examples are provided below to clarify the estimation cycle.

Example 1 - *Discrete marks with Poisson distribution and Logarithmic link*

Assume $X_i \sim \text{Poisson}(\theta_i)$ and guide relation $\log \theta_i \simeq \mathbf{z}_{II,i}\boldsymbol{\beta}_i$, $i = 1,2,\ldots$. Also, as usual, the guide relation for the intensity rate is $\log \lambda_i \simeq \mathbf{z}_{I,i}\boldsymbol{\beta}_i$. The likelihood factor l_i is

proportional to

$$\lambda_i \exp(-\lambda_i y_i) \cdot \theta_i^{x} \exp(-\theta_i)$$

with prior for $\beta_i | D_{i-1} \sim \mathbf{a}_i, \mathbf{R}_i$, this can be updated first to $[\beta_i | D_{i-1}, n_i = 1] \sim \mathbf{m}_{i-1}, \mathbf{C}_{i-1}$ where

$$\mathbf{m}_{i,1} = \mathbf{a}_i + \frac{\mathbf{s}_0}{q_0} \log \left[\frac{1 + q_0}{1 + q_0 y_i \exp(f_0)} \right] \tag{6.21a}$$

$$\mathbf{C}_{i,1} = \mathbf{R}_i - \frac{\mathbf{s}_0 \mathbf{s}_0^T}{1 + q_0} \tag{6.21b}$$

where $f_0 = \mathbf{z}_{I,i} \mathbf{a}_i$, $\mathbf{s}_0 = \mathbf{R}_i \mathbf{z}_{I,i}^T$ and $q_0 = \mathbf{z}_{I,i} \mathbf{s}_0$. Then, it can be updated with the information from X_i to $\beta_i | D_i \sim \mathbf{m}_i, \mathbf{C}_i$ where $D_i = \{D_{i-1}, [N_i = 1], [X_i = x_i]\}$ to

$$\mathbf{m}_i = \mathbf{m}_{i,1} + \frac{\mathbf{s}_1}{q_1} \log \left[\frac{1 + q_1 x_i}{1 + q_1 \exp(f_1)} \right] \tag{6.22a}$$

$$\mathbf{C}_i = \mathbf{C}_{i,1} - \frac{x_i}{1 + q_1 x_i} \cdot \mathbf{s}_1 \mathbf{s}_1^T \tag{6.22b}$$

where $f_1 = \mathbf{z}_{II,i} a_{i,1}$, $S_1 = \mathbf{C}_{i,1} \mathbf{z}_{II,i}^T$ and $q_1 = \mathbf{z}_{II,i} \mathbf{s}_1$. The vectors of covariates $\mathbf{z}_{I,i}$ and $\mathbf{z}_{II,i}$ have been extended with corresponding zero entries to have the same length. The updated moments (6.21) and (6.22) have been obtained in the same way as in chapter 2. From the prior distribution for β_i, a Gamma distribution for λ_i is constructed and updated with the information received from the PSEP in I_i. This is transformed back to β_i using linear Bayesian methods. This distribution is then used to construct a Gamma prior for θ_i. This is updated with the information from the mark and fed through to β_i as with λ_i. Once the posterior for β_i is available, the evolution to interval I_{i+1} is performed and another estimation cycle can be started.

Example 2 - *Continuous marks with Gamma distribution and Logarithmic Link*

Assume $X_i \sim G(\chi_i, \theta_i)$, χ_i known and guide relation $\log \theta_i \simeq \mathbf{z}_{II,i} \beta_i$ and the guide relation $\log \lambda_i \simeq \mathbf{z}_{I,i} \beta_i$ for the intensity rate, $i = 1,2,...$ Then, l_i is proportional to

$$\lambda_i \exp(-\lambda_i y_i)\ \theta_i^{\chi_i} \exp(-\theta_i x_i)$$

Then, the updating for λ_i is the same as in example 1 leading to $[\beta_i \mid D_{i-1}, n_i = 1] \sim m_{i-1}, C_{i-1}$ with moments given by (6.21). Updating for the information from X_i gives $\beta_i \mid D_i \sim \mathbf{m}_i, \mathbf{C}_i$ where

$$\mathbf{m}_i = \mathbf{m}_{i-1} - \frac{\mathbf{s}_1}{q_1} \log\left[\frac{1-\chi_i q_1}{1-q_1 x_i \exp(f_1)}\right] \tag{6.23a}$$

$$\mathbf{C}_i = \mathbf{C}_{i-1} - \frac{\chi_i}{1-\chi_i q_1} \mathbf{s}_1 \mathbf{s}_1^T \tag{6.23b}$$

with $f_1, \mathbf{s}_1$ and q_1 are as in example 1. The updating cycle is then completed as in example 1 with system equation (3.18) and moved to the next interval. Observe that this example covers the case of exponential distributions ($\chi_i = 1$) and censored marks ($\chi_i = 0$).

6.4 - Special reliability models

The study of point processes has led to the adoption of many specific forms to model reliability problems. In this section, some of those forms are presented and, in particular, those suggested to software reliability problems. It will be shown how these special models can be fitted or at least adapted to belong to the class of models suggested in this chapter.

6.4.1- Jelinski & Moranda (1972) model

This was one of the first models suggested in the area of software reliability. Assume that there is a total of N bugs (undetected at $t_0 = 0$) in a piece of software that could either be a long programme or a whole system.

The system is put to the test and every time a failure occurs, it is tracked down, corrected and the test continues. Assume that the observation process $N(t)$ is a PSEP with intensity rate $\lambda(t)$ given by (6.4) with

$$\lambda_i = \Phi\,[N-(i-1)]\ ,\ i = 1,2,... \tag{6.24}$$

This model reflects the (optimistic) view that as time passes, more and more bugs are found and, therefore, the rate of occurrence of failures decreases as there are less and less bugs to be found. The speed with which the intensity rate decreases is determined by the constant Φ.

Before showing how this model can be fitted into the structure of this chapter, a couple of points have to be made. First, it is quite feasible that Φ (the speed parameter) changes its value as bugs are detected. Also, although the testing stage is designed to remove eventual bugs from the system, there is no guarantee that this actually takes place. It often happens that in the process of correcting detected errors, new ones are inadvertently thrown into the system. All this amounts to saying that rather than (6.24), one should have a model allowing for both Φ and N to explicitly depend on time, or in this case, on i. It is worth remembering that if one insists on having static parameters, this can be arranged by specification of no evolution.

After those considerations, the model can be completed by a system equation

$$\begin{pmatrix}\Phi\\ N\end{pmatrix}_i = \begin{pmatrix}\Phi\\ N\end{pmatrix}_{i-1} + \mathbf{w}_i \quad ,\mathbf{w}_i \sim [O, y_i \mathbf{W}_i] \tag{6.25}$$

where $\mathbf{W}_i = \mathbf{O}$ if Φ and N are to be constant and by the guide relation

$$\lambda_i \simeq f(\beta_i) = \Phi_i [N_i - (i-1)] \text{ with } \beta_i^T = (\Phi_i \; N_i),\ i = 1,2,... \tag{6.26}$$

(6.26) presents us with a non-linear form albeit one of the simplest one. This non-linearity can be removed via the parametric transformation $\theta_i = (\theta_{i,1} \; \theta_{i,2})^T$ with $\theta_{i,1} = \Phi_i N_i$ and $\theta_{i,2} = \Phi_i$ or by a logarithmic link. The above is retained here though so that the inference can be made directly over the meaningful parameters Φ and N.

From (6.25) and a posterior $[\beta_{i-1} | D_{i-1}] \sim [\mathbf{m}_{i-1}, \mathbf{C}_{i-1}]$ one can obtain $[\beta_i | D_{i-1}] \sim [\mathbf{a}_i, \mathbf{R}_i]$ where $\mathbf{m}_{i-1} = \mathbf{a}_i = (a_{i,1} \; a_{i,2})^T$ and $\mathbf{C}_{i-1} + Y_i \mathbf{W}_i = \mathbf{R}_i = (\mathbf{r}_{i,1} \; \mathbf{r}_{i,2})$ with $\mathbf{r}_{i,j} = (r_{i,1j} \; r_{i,2j})^T$, $j = 1,2$. Then, results that

$$f_i = E[f(\beta_i) | D_{i-1}] = r_{i,12} + f(\mathbf{a}_i) \tag{6.27a}$$

$$\mathbf{s}_i = Cov[\beta_i, f(\beta_i) | D_{i-1}] = (\mathbf{r}_{i,2} \; \mathbf{r}_{i,1}) \mathbf{a}_i - (i-1) \mathbf{r}_{i,1} \tag{6.27b}$$

$$q_i = V[f(\beta_i)\,|\,D_{i-1}] \qquad r_{i,11}r_{i,22} - r_{i,22}^2 + a_{i,1}^2 r_{i,22} + a_{i,2}^2 r_{i,11}$$

$$- (i-1)^2 r_{i,11} - 2(i-1)(a_{i,2}r_{i,11} + a_{i,1}r_{i,22}) \qquad (6.27c)$$

with the help of multivariate normal moments as in Migon (1984). The guide relationship is now used to equate the mean and variance of $[\lambda_i\,|\,D_{i-1}]$ to f_i and q_i leading to $[\lambda_i\,|\,D_{i-1}] \sim G(\alpha_i, \gamma_i)$ with $\alpha_i = f_i^2/q_i$ and $\gamma_i = f_i/q_i$. Updating through Bayes' theorem leads to $[\lambda_i\,|\,D_i] \sim G(\alpha_i - 1, \gamma_i - y_i)$ which implies that $\beta_i|D_i \sim [\mathbf{m}_i, \mathbf{C}_i]$ with

$$\mathbf{m}_i = \mathbf{a}_i - \frac{\mathbf{s}_i}{q_i}\log\left|\frac{1-q_i}{1-q_i y_i \exp(f_i)}\right| \qquad (6.28)$$

$$\mathbf{C}_i = \mathbf{R}_i - \frac{\mathbf{s}_i\mathbf{s}_i^T}{1-q_i} \qquad (6.28b)$$

From this point, evolution for next cycle can start and the whole estimation cycle continue as usual. Smoothing and prediction can be performed as before using equations (6.27) to obtain the distribution of the λ's given that of the corresponding β's.

Apart from the prediction of the next failure time, it is important to have accurate measurements on the value of N at each time. It is often impracticable to wait to find all the errors in the system. This not only could take a very long time but can also make the system being released when it is already becoming obsolete. On the other hand, an early release could be prejudicial if the product keeps breaking and this can mean long-term damage for the manufacturers. A dynamic decision problem imposes itself for the study of the cost/benefit relation of the launching time of the product to the users. One suggestion is to specify a loss function $L(N, t)$ taking into account the profit to be gained by the release of the system against the loss incurred when the system is released with errors. The expected loss can be evaluated and decisions based upon it.

6.4.2- Littlewood & Verrall (1973) model

This is a similar but static version of the models proposed in this chapter. It assumes a PHPP for the observations with each λ_i having a Gamma prior with parameters α and $\psi_i(\theta)$. The model is completed with a prior distribution for α and θ. Observing that

$E[\lambda_i] = \alpha/\psi_i(\boldsymbol{\theta})$, one could incorporate α into ψ and write $E[\lambda_i] = \psi_i^{-1}(\boldsymbol{\beta}_i)$ for some parameter $\boldsymbol{\beta}$ as a function of α and $\boldsymbol{\theta}$. This can be adapted into the models of this chapter by assuming a guide relation of the form

$$g(\lambda_i) \sim \psi_i(\boldsymbol{\beta}_i)\ ,\ i = 1,2,... \tag{6.29}$$

where $g(x) = x^{-1}$ is the inverse function commonly used in standard generalised linear models (McCullagh & Nelder, 1983, pg 154).

As an illustration, suppose as in Littlewood & Verrall (1973) that $\psi_i(\boldsymbol{\beta}_i) = \beta_{i,2} + \beta_{i,2}\, i$. Then, after evolving from $[\boldsymbol{\beta}_{i-1}|D_{i-1}] \sim [\mathbf{m}_{i-1},\mathbf{C}_{i-1}]$ to $[\boldsymbol{\beta}_i|D_{i-1}] \sim [\mathbf{a}_i,\mathbf{R}_i]$ and obtaining $\phi_i = a_{i,1} + a_{i,2}\, i, \mathbf{s}_i = \mathbf{R}_i(1,\ i)^T$ and $q_i = (1,\ i)\mathbf{s}_i$, the prior distribution for λ_i can be constructed with the help of (6.29) by equating the moments of λ_i^{-1} and $\psi_i(\boldsymbol{\beta}_i)$. This leads to $[\lambda_i|D_{i-1}] \sim G(\alpha_i,\gamma_i)$ with $\alpha_i = 2 + f_i^2/q_i$ and $\gamma_i = f_i(1+\phi_i^2/q_i)$. Then, it can be updated as before leading to $[\boldsymbol{\beta}_i|D_i] \sim [\mathbf{m}_i,\mathbf{C}_i]$ with $\mathbf{m}_i$ and $\mathbf{C}_i$ as in (6.28) but with the above values of $f_i, \mathbf{s}_i$ and q_i. The analysis then follows as before.

6.4.3- Goel & Okumoto (1979) and Musa (1975) model

The approaches above basically model the observation process by assuming independent exponential failure times with specific forms for the parameters. The underlying Poisson processes are a mere consequence and are in fact not used. The models in this subsection make full use of the Poisson process structure to obtain a general continuous form for the intensity rate of the process. They are, however, based on the same ideas that led to Jelinski & Moranda model. There, the expected number of failures was taken as proportional to the number of remaining errors. In the models of this section, it is proportional to the *expected* number of remaining errors. This can be mathematically expressed in the difference form

$$\lambda(t)\Delta t = \Phi\left[N - \int_0^t \lambda(s)ds\right]\Delta t + o(\Delta t) \tag{6.30}$$

The expression for λ can be obtained after solving the resulting differential form. This model could also be conceivably implemented here after performing a discretisation of the intensity rate. In this case, (6.30) can be rewritten as the guide relation

$$\lambda_i \simeq \Phi_i \left[N_i - \Sigma_{j<i} \lambda_j b_j \right], \; i = 1,2,...$$

with a dynamic parameter $\beta = (\Phi \; N)^T$ and $b_j = \text{length}(I_j), \forall j$. Assuming that $[\beta_{i-1}|D_{i-1}] \sim [\mathbf{m}_{i-1}, \mathbf{C}_{i-1}]$, smoothing procedures can be applied to obtain the moments and covariances of $[\lambda_j \,|\, D_{i-1}], j \leq i-1$. Then, evolving to $[\beta_i|D_{i-1}] \sim [\mathbf{a}_i, \mathbf{R}_i]$, one can obtain

$$f_i = E[\Phi_i [N_i - \Sigma_{j<1} \lambda_j b_{N+j}] \,|\, D_{i-1}] \tag{6.31a}$$

$$\mathbf{s}_i = Cov[\beta_i, \Phi_i [N_i - \Sigma_{j<i} \lambda_j b_{N+j}] \,|\, D_{i-1}] \tag{6.31b}$$

$$q_i = V[\Phi_i [N_i - \Sigma_{j<i} \lambda_j b_{N+j}] \,|\, D_{i-1}] \tag{6.31c}$$

as in subsection 6.4.1. The presence of a function of the parameters $\lambda_1, \ldots, \lambda_{i-1}$ instead of the constant $(i-1)$ makes the above calculations that much harder. Alternatively, the values of the λ's in (6.31) could be replaced by $\hat{\lambda}_j = E[\lambda_j \,|\, D_{i-1}], j \leq i-1$ making the values of $f_i, \mathbf{s}_i$ and q_i the same as in (6.27) but for the replacement of $(i-1)$ by $\Sigma_{j<i} \hat{\lambda}_j b_{N+j}$. The remainder of the estimation cycle is as in 6.4.1. Even with the above approximation, the fact that smoothing is needed at every time interval may be prohibitive if there is a large number of time intervals.

Alternatively, the solution to the differential equation (6.30) can be used. This is given by

$$\lambda(t) = \Phi \, N \, exp(-\Phi t) \tag{6.32}$$

The way to proceed is to discretise (6.32) into piecewise constant values determined by the grid $\mathcal{T}$ as on section 5.2. In that section, the mid-point of each interval is used to represent the value of the hazard function in the interval. Assuming $\mathcal{T} = \{t_1, t_2, ...\}$, the same can be done here by a piecewise intensity rate with guide relation

$$\lambda_i \simeq \Phi_i \, N_i \, exp(-\Phi_i \tau_i) \quad , \; i = 1,2,...$$

where $\tau_i = (t_i + t_{i-1})/2$ and the parameters Φ and N are made dynamic. One can proceed by either retaining the meaningful parameters Φ and N in a non-linear model or by transforming them to $\beta = (\beta_1\ \beta_2)^T$ where $\beta_1 = log(\Phi N)$ and $\beta_2 = -\Phi$. With the use of a logarithmic link, this reduces the problem to a linear form with $z_i = (1\ ,\ \tau_i)$. The analysis is then performed as in the previous cases.

6.4.4- Aalen (1978, 1980) model

Another class of models suggested for a continuous intensity rate λ was introduced by Aalen (1978) and later extended by Aalen (1980). It simply assumes that

$$\lambda(t) = z(t)\,\beta$$

where $z(t)$ is a row vector of positive covariates affecting the intensity rate of the process. If it is assumed as before that the values of z only change with the change of time intervals and β is allowed to change with the same interval, then the analysis proceeds as before. In fact, estimation for a general model where both parameters and the corresponding covariates are allowed to change continuously with time is outlined in the next chapter. The derivation follows the same limiting procedures adopted in section 6.2.

This class of models is very rich and has been shown by Koch & Spreij (1985) to cover all previous examples when $z(t)$= number of remaining errors at time t.

6.4.5- Seasonal models

The occurrence of failures or breakdowns in a system can be attributed to a number of factors and can happen in any pattern. One possibility is to have failures displaying a repeated or, if you like, cyclical pattern. In it, failure times could go through a period where they are generally long to a subsequent period when they seem to be short, suggesting that failures occur in bunches or clusters.

This set-up is taken up by Crow & Singpurwalla (1984). They assume that the mean time between successive failures can be described in a Fourier series transformation of the form

$$\alpha_0 + \sum_{j=1}^{q} \alpha_j \cos\left(\frac{2\pi j}{n}\right) + \sum_{j=1}^{q} \beta_j \sin\left(\frac{2\pi j}{n}\right)$$

Where $n =$ number of failures observed and $q = [n/2]$. This cyclical representation is commonly used in time series. In some cases, only a few of the q terms are relevant for the analysis of the series and the models can be substantially reduced. This series representation must be used with caution with positive variables because they can easily lead to models fitting negative failure times as in the above paper. A logarithmic transformation can help to solve this problem.

In the simple case where the Fourier representation is deemed to depend on time only through the number of failures, this problem can be tackled with the structure of dynamic GLM presented in chapter 2. This will typically consist of exponential observations, a logarithmic link and a system equation with

$$\mathbf{G}_i = diag(1, \mathbf{H}_1, \cdots, \mathbf{H}_q), \quad \mathbf{H}_j = \begin{pmatrix} \cos(2\pi j/n) & \sin(2\pi j/n) \\ -\sin(2\pi j/n) & \cos(2\pi j/n) \end{pmatrix} \quad , j = 1,\ldots,q$$

Other forms of seasonal evolutions should be sought so that they can depend on time rather than on the number of failures. This involves a generalisation of the discrete form of seasonal evolution to continuous ones.

6.4.6- Autoregressive models

These models are also commonly known in time series and have been recently suggested by Singpurwalla & Soyer (1985) to the study of software reliability. The basic idea is to impose on the model the assumption that the outcome of the present failure time depends on the previously observed failure times. This dependence is generally taken in the first step to be linear and the number of past failure times affecting the outcome of the present is as small as possible for parsimony.

This can be implemented here by assuming a PSEP for the observations with each of the λ's related to the previous failure times via

$$\log \lambda_i \simeq \beta_{0,i} + \sum_{j=i}^{q} \beta_{j,i} Y_{i-j} \quad , i = 1,2,\ldots$$

The value of q is set by the modeller or can be chosen according to some optimality criterion or through the predictive likelihood as suggested in chapter 3.

Finally, the structure of the models of this chapter enable the superposition of effects as in DLM. This means that an autoregressive model can be added to a seasonal model or any other regression model with a logarithmic link.

References

Aalen, O. O. (1978). Non-parametric inference for a family of counting processes. *Ann. Statistics,* **6**, 701-726.

----- (1980). A model for non-parametric regression analysis of counting processes. *Springer Lect. Notes in Statist.,* **2**, 1-25.

Cox, D. R. (1955). Some statistical methods connected with series of events.*J. R. Statist. Soc., Ser. B,* **17**, 129-164.

Crow, L. H. & Singpurwalla, N. D. (1984). An empirically developed Fourier series model for describing software failures. *IEEE Trans. Rel.,* **R-33**, 176-183.

Gill, R. D. (1984). Understanding Cox's regression model: a martingale approach. *J. Amer. Statist. Ass.,* **79,** 441-447.

Goel, A. L. & Okumoto, K. (1979). Time-dependent error-detection rate model for software reliability and other performance measures. *IEEE Trans. Rel.,* **R-28**, 206-211.

Grandell, J. (1976). *Doubly Stochastic Poisson Processes.* Berlin: Springer-Verlag.

Jelinski, Z. & Moranda, P. B. (1972). Software reliability research. In *Statistical Computer Performance Evaluation, (ed.: W. Freiberger). New York: Academic.*

Koch, G. & Spreij, P. J. C. (1983). Software reliability as an application of martingale & filtering theory. *IEEE Trans. Rel.,* **R-32**, 342-345.

Littlewood, B. & Verrall, J. L. (1973). A Bayesian reliability growth model for computer software. *J. R. Statist. Soc., Ser. A,* **22**, 332-346.

McCullagh, P. & Nelder, J. A. (1983). *Generalized linear Models.* London: Chapman & Hall.

Migon, H. S. (1984). An approach to Bayesian non-linear forecasting problems with applications. Unpublished Ph.D. thesis. Univ. of Warwick.

Musa, J. D. (1975). A theory of software reliability and its application. *IEEE Trans. Software Engineering,* **SE-1**, 312-327.

Singpurwalla, N. D. & Soyer, R. (1985). Assessing (software) reliability growth using a random coefficient autoregressive process and its ramifications. *IEEE Trans. Software Engineering,* **11**, 1456-1464.

Snyder, D. L. (1975). *Random Point Processes.* New York: John Wiley.

CHAPTER SEVEN

GENERALISATIONS AND FURTHER TOPICS

The main results of this thesis were presented in previous chapters. There are a number of ways however that they can be generalised or extended. Also, there are some points that are not central to the analysis and were not fully exploited but are important for a complete theory. All this is briefly treated on this chapter.

The first generalisation was already hinted at the start of the previous chapter. As survival data deals with many observations with a single occurrence and point processes with one observation with many occurrences. it is no surprise that there is interest in the study of many observations with many occurrences. This can be extended even further by introducing the cause of occurrence as an extra randomness. These processes are studied in the first section.

It has been said throughout this thesis that the discretisation of the observation points is an artifice to provide tractable analysis. The more appropriate inference with a continuum of observation times is outlined in the second section. The derivations follow the same lines as the ones performed in section 5.2.

The inference procedure itself has a number of points needing further discussion, among them, the problem of order dependence. Alternatives are suggested and shown to lead to seemingly greater problems mainly with respect to computations. Also, the prior specification that has been somewhat neglected is given some attention. The important problems of analysis of residuals and robustness are also discussed. Finally, some words are drawn into the areas of model selection (particularly the specification of the evolution) and simulation.

7.1 - Multivariate point processes and multivariate failure data

A multivariate point process (MVPP) is a collection of point processes. They are generally assumed independent and are often presented in the literature without previous

reference to (univariate) point processes. This was not done here to emphasise the dual aspect of point processes and because a single point process is sufficient to the study of software reliability.

In the present context, MVPP are constructed to study similarities among point processes. To be more precise, one can think of factors affecting the intensity rate of the processes and study the effect they have on each process. The basic situation to be considered is of a series of similar components put into operation. The data consists in observing the failure times of the components. Typically, one would have the components operating in different conditions and the main problem is to establish the way those conditions affect the performance. The analysis however is very similar to the previous ones.

Let $N_1(t),...,N_m(t)$ be independent PSEP given their intensity rates $\lambda^{(1)}(t),...,\lambda^{(m)}(t)$. The intensity rates are assumed to be defined by the same common grid $\tau=\{t_1,t_2,...\}$ which can include the cumulative failure times from all the processes, if needed. The total likelihood for each process $N_j(t)$ is, from (6.5),

$$\prod_{i=1}^{\infty}\left[\lambda_i^{(j)}\right]^{n_{ij}}\exp\left(-\lambda_i^{(j)}b_i\right) \tag{7.1}$$

where $\lambda_1^{(j)},\lambda_2^{(j)},...$ are the piecewise constant values of λ_j in each interval and n_{ij} is the observed value of $N_j(t_{i-1},t_i]$, $j=1,..,m$. Due to the conditional independence, the overall likelihood from all processes is just the product of the factors like (7.1). Once again, the TFL can be performed giving the factors

$$l_i=\prod_{j=1}^{m}\left[\lambda_i^{(j)}\right]^{n_{ij}}\exp\left(-\lambda_i^{(j)}b_i\right)\quad,\ i=1,2,... \tag{7.2}$$

One can relate the λ's to the relevant covariates via

$$g(\lambda_i^{(j)})\simeq h(\mathbf{z}_{i,j},\boldsymbol{\beta}_i)\quad,j=1,...,m\ ,\ i=1,2,... \tag{7.3}$$

where g and h are specified by the modeller and $\mathbf{z}_{i,j}$ is the value assumed by the covariates for process N_j at time interval I_i, $j=1,...,m$, $i=1,2,...$ The above set-up is exactly the same as for survival data analysis in chapter 3. If one assumes further that $g=log$ and h is linear than the inference procedure leads to the same equations with substitution of the

χ_{ij}'s by the n_{ij}'s and of the y_{ij}'s by the b_i's being the only change required.

The multivariate failure data with competing risks provides the largest generalisation envisaged at this stage. The analysis for competing risks has already been presented in section 5.3 but it was not considered multivariate because only a single failure (survival) time could be observed. Here, the different causes of failures are not entirely competing. Hence, the observation process can proceed beyond the first observed failure.

This observation process includes, as special cases:

(i) survival data - by not taking causes into account and restricting the observation process to the first survival time;

(ii) competing risk data - by restricting the observation process to the first survival time;

(iii) MVPP - by not taking causes into account;

Following section 5.3, define for a point process $N(t)$

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{Pr\,[N(t-\Delta t,t] = 1, \delta = k \mid D(t-\Delta t), \mathbf{z})}{\Delta t} \quad , k = 1,\ldots,K \tag{7.4}$$

where $D(t-\Delta t)$ is information up to time $t-\Delta t$, δ denotes the cause of failure out of a total of K and $\mathbf{z}$ denotes the vector of covariates, as usual. λ_k can be called the cause-specific intensity rate of the process $N(t)$. It can easily be related to λ, the intensity rate of $N(t)$ by observing that

$$\sum_{k=1}^{K} \lambda_k(t) = \sum_{k=1}^{K} \lim_{\Delta t \to 0} \frac{Pr(N(t-\Delta t,t] = 1, \delta = k \mid D(t-\Delta t), \mathbf{z})}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{\sum_{k=1}^{K} Pr(N(t-\Delta t,t] = 1, \delta = k \mid D(t-\Delta t), \mathbf{z})}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{Pr(N(t-\Delta t,t] = 1 \mid D(t-\Delta t), \mathbf{z})}{\Delta t}$$

$$= \lambda(t) \tag{7.5}$$

Where the last equality follows from the definition of point processes given in chapter 6.

Let now $N_1(t), \ldots, N_m(t)$ be independent PSEP given their intensity rates $\lambda^{(1)}(t), \ldots, \lambda^{(m)}(t)$ where each of them is given by

$$\lambda^{(j)}(t) \quad \sum_{k=1}^{K} \lambda_k^{(j)}(t) \quad , j = 1,\ldots,m$$

and each of the $\lambda_k^{(j)}$'s is piecewise constant along a common grid $\mathcal{T} = \{t_1, t_2, \ldots\}$. The total likelihood for each process $N_j(t)$ is

$$\prod_{i=1}^{r} \left\{ \prod_{k=1}^{K} \left[\lambda_{i,k}^{(j)} \right]^{n_{ijk}} \right\} \exp(-\lambda_i^{(j)} b_i) \tag{7.6}$$

where n_{ijk} is the number of occurrences in I_i of cause $k, k = 1, \ldots, K$, $j = 1,\ldots,m$. Again, the likelihood factors are easily obtained from (7.6) as

$$\prod_{j=1}^{m} \left\{ \prod_{k=1}^{K} \left[\lambda_{i,k}^{(j)} \right]^{n_{ijk}} \right\} exp(-\lambda_i^{(j)} b_i) \quad , i = 1,2,\ldots \tag{7.7}$$

Equation (7.6) is an extension of (6.5) using the ideas of section 5.3. One can obtain it using the following argument. Consider only the interval $I_i = (t_{i-1}, t_i]$ without loss of generality since the likelihood naturally factorises for PSEP. Suppose a failure of type k_1 is observed at $x \in I_i$. Then, the likelihood is

$$\lim_{\Delta \to 0} Pr(N_j(t_{i-1}, x-\Delta] = 0,\ N_j(x-\Delta, x] = 1,\ \delta = k_1,\ N_j(x, t_i] = 0 \mid D_{i-1}) =$$

$$= \lim_{\Delta \to 0} \exp\left[-\lambda_i^{(j)}(x-\Delta-t_{i-1})\right] \lambda_{i,k_1}^{(j)} \exp\left[-\lambda_i^{(j)}\Delta\right] \exp\left[-\lambda_i^{(j)}(t_i - x)\right]$$

$$= \lambda_{i,k_1}^{(j)} \exp\left[-\lambda_i^{(j)} b_i\right]$$

The above result can be generalised for a number of failures and intervals leading to (7.6). Observe that once (7.7) is obtained, the inference procedures of this thesis can be readily used for given forms of relationship between the cause-specific intensity rates and the relevant covariates.

7.2 - Survival analysis in continuous time

As it has been discussed in chapter 3, the use of a discrete grid $\mathcal{T}$ of observation

Points is appropriate for a tractable analysis but is not in fact reflecting the true observation process. In it, one has knowledge about the individual's survival at every new time and observation is in fact continuous. The inference procedure for continuous time can be carried out although the presence of many individuals will make the results harder to obtain when compared to the derivation of section 6.2.

Here, the estimation equations are obtained and shown to lead to continuous Kalman filter-like equation just as in the discrete case. The derivations are carried out as before via a limiting argument by making the length of a given interval shrink towards 0. In doing this $\beta_i, D_j, \mathbf{m}_k$ are changed to $\beta(t_i), D(t_j), \mathbf{m}(t_k), \cdots$ to conform with the notation.

Suppose that for a small interval $I_i = (t_{i-1}, t_i]$ of length $\Delta t = t_i - t_{i-1}$,

$$\beta(t_{i-1}) \mid D(t_{i-1})] \sim [\mathbf{m}(t_{i-1}), \mathbf{C}(t_{i-1})]$$

$$[\beta(t_i) \mid D(t_i)] \sim [\mathbf{m}(t_i), \mathbf{C}(t_i)]$$

The idea is to relate $\mathbf{m}(t_i)$ and $\mathbf{C}(t_i)$ to $\mathbf{m}(t_{i-1})$ and $\mathbf{C}(t_{i-1})$ as $\Delta t \to 0^+$. Assuming that β evolves according to (2.61), it implies through (2.66) that, due to the evolution,

$$\mathbf{m}(t_i) - \mathbf{m}(t_{i-1}) = \Delta t\, \mathbf{F}(t_i)\, \mathbf{m}(t_{i-1}) + o(\Delta t) \tag{7.8a}$$

$$\mathbf{C}(t_i) - \mathbf{C}(t_{i-1}) = \Delta t\, [\mathbf{F}(t_i)\mathbf{C}(t_{i-1}) + \mathbf{C}(t_{i-1})\mathbf{F}^T(t_i) + \mathbf{W}(t_i)] + o(\Delta t) \tag{7.8b}$$

where $W(s) = \Gamma(s)\mathbf{Q}(s)\Gamma^T(s)$. The inclusion of the term $o(\Delta t)$ is due to

$$\mathbf{G}(t_i, t_{i-1}) = \mathbf{I} + \mathbf{F}(t_i)\Delta t + o(\Delta t)$$

obtained from the properties of G, and

$$\int_{t_{i-1}}^{t_i} \mathbf{G}(t_i, s)\, \mathbf{W}(s)\, \mathbf{G}^T(t_i, s)\, ds = \Delta t\ \mathbf{W}(t_i) + o(\Delta t)$$

If there are $r(t_i)$ individuals present in I_i with death indicators $\chi_j(t_i)$ then the total change in $\mathbf{m}$ and $\mathbf{C}$ is given, from (3.20), by

$$\mathbf{m}(t_i) - \mathbf{m}(t_{i-1}) = \Delta t\, \mathbf{F}(t_i)\mathbf{m}(t_{i-1}) + \sum_{j=1}^{r(t_i)} \frac{\mathbf{s}_j(t_i)}{q_j(t_i)} \log \left\{ \frac{1 + q_j(t_i)\chi_j(t_i)}{1 + q_j(t_i)\, \Delta t \exp [f_j(t_i)]} \right\} + o(\Delta t)$$

$$\mathbf{C}(t_i)-\mathbf{C}(t_{i-1}) = \Delta t\{\mathbf{F}(t_i)\mathbf{C}(t_{i-1})-\mathbf{C}(t_{i-1})\mathbf{F}^T(t_i)+\mathbf{W}(t_i)\}-\sum_{j=1}^{r(t_i)}\frac{\chi_j(t_i)}{1+q_j(t_i)}\mathbf{s}_j(t_i)\mathbf{s}_j^T(t_i)+o(\Delta t)$$

where $f_j(t_i)=\mathbf{z}_j\mathbf{m}_{j-1}(t_i)$, $\mathbf{s}_j(t_i)=\mathbf{C}_{j-1}(t_i)\mathbf{z}_j^T$ and $q_j(t_i)=\mathbf{z}_j\mathbf{s}_j(t_i)$. $\mathbf{m}_j(t_i)$ and $\mathbf{C}_j(t_i)$ are the mean and variance of $\beta(t_i)$ given the information from the first j of the $r(t_i)$ individuals in interval I_i and covariates are assumed time-independent. Observing that for a small Δ,

$$\log\left(\frac{a}{1-b\Delta}\right) \quad \log a - b\Delta - o(\Delta)$$

one obtains, by letting $\Delta t \to 0^+$, that

$$d\mathbf{m}(t) = \mathbf{F}(t)\mathbf{m}(t)dt + \sum_{j=1}^{r(t)}\frac{\mathbf{s}_j(t)}{q_j(t)}\{\log[1+q_j(t)\chi_j(t)] - q_j(t)\exp[f_j(t)]dt\} \tag{7.9a}$$

$$d\mathbf{C}(t) = [\mathbf{F}(t)\mathbf{C}(t)+\mathbf{C}(t)\mathbf{F}^T(t)+\mathbf{W}(t)]dt - \sum_{j=1}^{r(t)}\frac{\chi_j(t)}{1+q_j(t)}\mathbf{s}_j(t)\mathbf{s}_j^T(t) \tag{7.9b}$$

The above equations were obtained using heuristic arguments in order to obtain an idea of the final forms. It can be, however, made more rigorous by using elements from martingale and filtering theory but this is beyond the scope of this thesis. The equations are not easy to solve now especially due to the summation terms. It is possible to see, however, that at the points at which deaths occur, the moments of β experience a jump with a decrement in the value of the variance confirming results from section 6.2.

7.3 - Models assumptions and inference procedures

The models were constructed in a versatile way but there are not many assumptions involved in the build-up process. Basically, they only need piecewise exponentiality (associated with the TFL), a system equation and a link function between those two. The first two are fairly simple and acceptable and the main problem of the models seems to be the order dependence in its inference procedure. This point is therefore discussed first.

The dependence of the analysis on the order the observations are processed is shown to be very small but it still contradicts the basic rules of probability. It arises from the

use of the conjugate inference of the dynamic GLM proposed by West, Harrison & Migon (1985) in a static environment. It provides the required answer in a computationally efficient way. The usual criticism of the use of conjugate distribution weakens in a time series application geared towards prediction where predictive measures can help in the choice of the more appropriate model. The particular way the conjugate inference treats discordance between prior and likelihood is also discussed in the next section.

It is the way the likelihood parameters ($\lambda's$) are related to the system parameters ($\beta's$) that brings in the order dependence. Many alternatives can be suggested:

(i) use of suitable version of a multivariate Gamma distribution;

(ii) specification of a full distribution for the $\beta's$;

(iii) correction of the updating procedure to remove order dependence.

The first point is only possible if rich enough families of multivariate Gamma distributions are found so that they can produce the covariance structure of the regression design. The second point is the natural Bayesian answer to the problem but, unless a conjugate distribution is used, one will be left with many different multivariate Bayesian analysis to perform and then relate them with some sort of evolution. Even the conjugate family adopted will have to be large enough to contain the distribution resulting from the evolution step. High-dimensional numerical integration procedures will still be needed to relate moments to densities of this conjugate distribution. Those integrations are to be repeatedly used for every time interval and computing facilities still seem a little way off in terms of routine application. Efforts towards this direction are therefore welcomed (see Naylor & Shaw (1985)). The final point has also been investigated without showing much progress.

The piecewise exponentiality was adopted having in mind a model that can, in practical terms, reproduce in a discrete way any form of hazard function. Even for moderate sized samples, a fair number of intervals is constructed providing a very good idea of the form of the hazard function. Also, in keeping with the idea of letting the data speak for

themselves, the exponential distribution is used in the piecewise approach. Any other distribution would not only imply extra assumptions in the form of the hazard but bring extra parameters to the analysis. So, unless there is a definite reason for the use of a different piecewise distribution (as in section 5.2), exponential distributions should be used.

Finally, system equations of the form (3.18) are again simple enough to be understood and controlled by the user. In simpler problems, when there is only a simple parameter evolving in time as in section 5.1 or 6.2, other forms can be used as well. In more structured problems, however, (3.18) should be used. The specification of the values of the W_i's is important and a number of suggestions are presented in chapter 2. Alternatively , the discount approach could be used (with caution) as suggested in chapter 2 or in blocks, as suggested by West & Harrison (1986). Each block contains the parameters deemed to be related and to evolve in the same way. Section 4.3 presents one use of this approach with a single block containing all the parameters. One sensible variation is to have 2 blocks: one with the baseline hazard parameters and the other one with the regression parameters. This updated version of the discount approach is implemented in the computer programmes presented in the next chapter.

7.4 - Robustness and analysis of residuals

One aspect of model fit is the numerical summary provided, as shown in chapter 3, by the predictive or marginal likelihood of the model. This is a comparative measure that assesses the goodness of the model in comparison with others models. It does not provide an idea of how reasonable the model is (on its own) to explain the data. The usefulness of the model in prediction is the final test but it would be nice if some check of structural or systematic inadequacy could be detected before obtaining poor predictive results. A good model (by the predictive likelihood approach) can still perform poorly in prediction if its goodness derives from comparisons with even poorer models.

The idea is to use the predictive approach simultaneous to the estimation process by predicting and comparing the prediction to each individual observed. One can detect systematic departure of the models or even influential observation to be singled out. In standard time series, the problem is approached as follows: at time $t-1$, predictions for time t are made based on the distribution F_t. The distribution of $F_t(Y_t)$ prior to observing Y_t is $U[0,1]$ if Y_t derives from F_t and so is the distribution of $S_t(Y_t)$, the predictive survival function of Y_t. Since these distributions are conditional on all previous information, it follows that $S_1(Y_1), S_2(Y_2),...$ form a random sample from a $U[0,1]$ distribution if the data actually derives from the model. After observing $Y_t = y_t$, one has $S_t(Y_t) = s_t$ and checks can be readily made to see if $\mathbf{s} = (s_1, s_2,...)$ correspond to observations from a random uniform sample.

In survival data, the problem is more complicated due to the presence of many observations to be predicted, most of which censored. At any time t_{i-1}, one has to predict the outcome of

$$Y_{ij} = min(Y_j, t_i) \quad , j = 1,...,r_i \ , i = 1,2,...,N$$

conditional on $\{[Y_j > t_{i-1}], j = 1,..,r_i\}$ because this is what is going to be observed in I_i. For every interval, most of the Y_{ij}'s are not going to be fully observed. So, new methods have to be sought for this problem. One possibility is to use the observed values of $S(Y_{ij})$ for the Y_{ij}'s that are fully observed but this still lacks sound justification. Some suggestions on the definition of residuals from censored observations have been put forward by Kay (1977), Crowley & Hu (1977) and Lagakos (1981) among others. It is still unclear how censored data should be treated specially in this factorised approach.

7.5 - Prior assessments

This is another topic that was admittedly neglected. Its importance is beyond question and, although the effect of an informative prior was assessed to some extent in section 4.1, no constructive procedures were outlined. The particular structure of survival models could be used however to tackle this problem.

The only prior input to the model is the distribution of $[\beta_0|D_0]$. The main idea is to use the intuitive appeal of hazard functions to set the moments of β. In particular, if a logarithmic link is used, the effects of the covariates can be isolated. This is, in fact, another good reason for the repeated use of this link function. The prior can be set by the use of comparisons, as follows:

(i) assume two individuals with the same characteristics but for one, say z_k;

(ii) assume that the value of that covariate differ between the individuals by one unit of measurement (or any other amount one feels comfortable about comparisons). The ratio of the two hazards is $\exp(\beta_k)$;

(iii) using the notion of risk, set upper and lower bounds to the logarithm of the relative risk (represented by the ratio of the hazards) between the individuals at the time origin;

(iv) take those bounds as $\pm r$ standard deviations around the mean and obtain mean and variance of β_k, for $r > 0$.

This procedure is repeated for all covariates leading to a mean vector and diagonal covariance matrix for the regression parameters. One can even go further and consider covariance terms by repeating the above procedure with unit increments in two covariates. The diagonal covariance structure is a sensible procedure to start the analysis and is implemented in the computer system of the next chapter with a default value of $r = 2$. The above procedure does not apply for the baseline hazard parameter and other methods should be sought.

7.6 - Simulation

The emphasis of this work is in the detection of changes in the structure of dependence of survival times on some covariates. Also, the models are based on a discretisation that was admitted not to be the most accurate representation of the data but the most efficient one. For these reasons, use of simulation was scarce in this thesis.

There is no denying of the importance of simulation procedures specially to investigate the ability of the models to pick up the correct generating structures of the data. This work was done in the analysis of random samples (section 5.1) but there are many other areas it could be used. Some further work is needed in the determination of the link between hazard and regression coefficients and of the form the coefficients change in time. Simulation could also help in the determination of the time it takes the model to understand the mechanism generating the data.

References

Crowley, J. & Hu, M. (1977). Covariance analysis of heart transplant survival data. *J. Amer. Statist. Ass.*, **73**, 27-36.

Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Appl. Statist.*, **26**, 227-237.

Lagakos, S. W. (1981). The graphical evaluation of explanatory variables in proportional hazard regression models. *Biometrika*, **68**, 93-98.

Naylor, J. C. & Shaw, J. E. H. (1985). Bayes four - user guide. Nottingham Statistics Group, Department of Mathematics, University of Nottingham, Nottingham, U.K.

West, M. & Harrison, P. J. (1986). Monitoring and adaptation in Bayesian forecasting models. *J. Amer. Statist. Ass.*, **81**, 741-750.

West, M., Harrison, P. J. & Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Amer. Statist. Ass.*, **80**, 73-97.

CHAPTER EIGHT

MICROCOMPUTER SOFTWARE FOR THE ANALYSIS

This chapter presents details of the initial version of a program developed for implementation of the models of Chapters 3 and 4. The software is written in APL using the microcomputer version of APL*PLUS/PC available for OLIVETTI M24, IBM-PC and similar machines. No knowledge of APL is needed by the user. The program comprises a set of functions in the workspace named SURVIVAL. It is intended to be a self-contained package, providing facilities for data input, data analysis, displays and graphical outputs. The functions have been written in an interactive way explaining the different options available and with guiding instructions when required. Planned developments not yet implemented include extensions for the analysis of competing risk and point processes.

The workspace can be roughly divided into the three groups cited above. The management of the data input is particularly important in the case of time-dependent covariates. The data analysis group contains the selection of prior moments, time grid and type of evolution with consequent estimation and smoothing of the parameters. It also includes a number of scattered plots for preliminary assessments of the data. The final group provides the output of the analysis with time trajectories of the parameters, predictive features and retrospective assessment of the form of the hazard function as well as final estimates, useful, in particular, in static models.

The next section provides a description of the main functions contained in SURVIVAL. The following section provides an example of their use.

8.1 - SURVIVAL functions

The functions available are divided in the three groups above for expository purposes. In addition, there are a number of functions necessary for the operation of the workspace gathered in a control group. The functions are listed alphabetically within each group and a brief description of their use is given. The functions are presented in bold

characters with their arguments, if any, in italic characters. The functions that are not directly accessed by the user are preceded by an asterisk. All commands are made effective by simply typing the function name, with arguments as necessary, and pressing the return key.

DATA MANAGEMENT FUNCTIONS

All data sets to be analysed with SURVIVAL have to be input using the data management functions. This procedure ensures that the data is always stored in files in the same format. If there are time-dependent covariates, the data must be input again for all those individuals whose covariates change with time, specifying the times of the change. The changed data is then put into another component of the same file. This gives rise to the nomenclature initial data referring to the data at the initial time taken as 0. The initial data coincides with the data when all covariates are time-dependent, of course. This is referred to as (initial) data to cover both cases.

1. **AMEND** *X*

   Allows amendment to the (initial) data of a file. The left argument *X* is a vector with the indices of the individuals whose data is to be amended. APL expression are also allowed. Previous values of the data for the individual concerned is displayed to help the operations.

   Examples: i) AMEND 3 - amends the data for the third individual;

   ii) AMEND 1,4,5 or AMEND 1 4 5 - amends the data for the first, fourth and fifth individuals;

   iii) AMEND ¯1+2×⍳10 - amends data for the first 10 odd indexed individuals.

2. (*)**CHANGE**

   Allows input of changed values of the covariates. This fuction is automatically called in INPUT after indication of presence of time-dependent covariates.

3. **DATASET**

Provides a summary description of the data currently in use. It is also automatically called in READ.

4. **INPUT**

Allows input of a new data set into a file from the keyboard. It asks for a file title, a brief description, name of the time units, number of individuals, number and name of covariates and for the presence of time-dependence of the covariates. Data is then inputed individual by individual (automatically indexed) in the order: survival time, death indicator and covariates (ordered as in the list given) and all this information is stored in a file.

5. **LOOK** *X*

Displays data from a file. The left argument *X* is a vector as in AMEND. The examples are as in AMEND with replacement of amend by display.

6. **LOOKALL** *X*

Displays data from a file. The left argument *X* is a vector as in AMEND but typically of dimension 2. It differs from LOOK by displaying data for all the indices contained in the range of *X*.

Examples: i) LOOKALL 2 5 - displays data for all individuals indexed from 2 to 5;

ii) LOOKALL 6, 3, 11 - displays data for all individuals indexed from 3 to 11.

7. **(*)PREPARE**

Gives names to data vectors associating TIMES with the survival times, DEATHS with death indicators and the covariates with respective names (specified in INPUT). It is automatically called in READ and SELECT.

8. **READ**

Reads data from file and calls DISKDRIVE, PREPARE and DATASET. It asks for

the file title.

9. **WRITE**

Writes data into a file. This function is automatically called in AMEND, CHANGE and INPUT.

DATA ANALYSIS FUNCTIONS

1. **ANALYSE**

Defines the model and performs estimation and smoothing cycles. It prompts for the covariates to be included in the model, if any, and calls the functions specifying prior moments, evolution and time grid and fitting the model. Function READ must have been previously used to read data set from file.

2. **(*)ARRAY**

Prepares data set to be used in each estimation cycle. It removes individuals without contribution to the time interval and change value of the covariates for that interval in case of time-dependence of covariates. It is automatically called in ESTIMATION.

3. **(*)ESTIMATION**

Performs the estimation cycle. It is automatically called in ANALYSE.

4. **(*)EVOLUTION**

Performs evolution to next time interval in the estimation cycle. It is automatically called in ESTIMATION and uses the evolution previously set in SETEV. The discount approach uses the discount by blocks suggested in West & Harrison (86) and described in Chapter 7.

5. **(*)GRID**

Specifies the time grid to be used in the analysis. It allows for a choice of grids based on groups of intervals with equal lengths, death times or on an equal number of

death times per interval, or any other arbitrary time configuration. It is automatically called in ANALYSE .

6. (*)**PRIOR**

Specifies the prior moments for the analysis. The function asks if user wants to initiate with vague moments (0 mean and $10^3 \mathbf{I}_{p+1}$ variance). If not, takes upper and lower limits provided by the user for each parameter as 2 s.d. limits around means and constructs prior accordingly assuming a diagonal variance matrix. It is automatically called in ANALYSE.

7. *Y* **SELECT** *X*

Selects a subset of the data. The left argument *Y* is the name of the variable upon which the selection is to be based and the right argument *X* is a vector containing the limits of *Y* for the selection. It is useful for subsequent analyses and plots.

Examples: i) 0 SELECT *i* 50 - selects the first 50 observations in the data set;

ii) TIMES SELECT 100 200 - selects the individuals with survival time between 100 and 200 time units;

iii) DEATHS SELECT 1 - selects the individuals that die (consequently removing the censored ones).

8. (*)**SETEV**

Sets the evolution pattern for the analysis. It asks for the selection of either the additive variance matrix (AVM) or discount (D) approach and for the number of time units used as base for the evolution specification. The AVM approach offers the choice of arbitrary or principal components (see chapter 2, pg. 16) variance setting. Thus, if the setting is based on *b* time units then:

i) in the AVM approach, with a specified matrix $\mathbf{W}^*$,the analysis is based on $\mathbf{W} = \mathbf{W}^* / b$;

ii) in the D approach, with a specified vector of discounts $\mathbf{d}^*$, the analysis is based

on discounts $\mathbf{d} = ((d_0^*)^{1-b}, (d_1^*)^{1/d}, \ldots, (d_p^*)^{1-b})$ to conform with the treatment of unequal time intervals (see chapter 2, pg. 15).

This function is automatically called in ANALYSE.

(*)SMOOTHING

Performs the respective smoothing cycles. This function is automatically called in ESTIMATION.

DISPLAY AND GRAPHICS OUTPUT FUNCTIONS

The output functions are mostly graphical. To elaborate graphical displays, a user may interactively edit and amend plots by using the cursor. This can be moved using arrows as usual. If the Ctrl key is pressed simultaneously, the cursor moves 10 locations and if the Alt key is pressed simultaneously, it moves 50 locations in the direction indicated by the arrow. The options available are:

A to draw arrow from last cursor position to current position.

D to draw a line from last to current position.

E to erase rectangle whose corners are last and current position.

F to obtain the decile (when displaying survival functions) given by next character typed (1 = 10% decile, 5 = median, etc...). This moves the cursor to the required decile in the y axis. The horizontal arrows can then be used to place the cursor on the survival curve and, after pressing return, the decile is displayed on the screen.

H to provide a menu of the options available.

K to save the graphic display for later use.

L to change line type to that define by next character typed (0 = solid, 1 = dotted, etc...).

M to mark or remember current cursor position.

P to print screen content on Epson or similar.

on discounts $\mathbf{d} = ((d_0^*)^{1/b}, (d_1^*)^{1/d}, \ldots, (d_p^*)^{1/b})$ to conform with the treatment of unequal time intervals (see chapter 2, pg. 15).

This function is automatically called in ANALYSE.

9. (*)**SMOOTHING**

Performs the respective smoothing cycles. This function is automatically called in ESTIMATION.

DISPLAY AND GRAPHICS OUTPUT FUNCTIONS

The output functions are mostly graphical. To elaborate graphical displays, a user may interactively edit and amend plots by using the cursor. This can be moved using arrows as usual. If the Ctrl key is pressed simultaneously, the cursor moves 10 locations and if the Alt key is pressed simultaneously, it moves 50 locations in the direction indicated by the arrow. The options available are:

A to draw arrow from last cursor position to current position.

D to draw a line from last to current position.

E to erase rectangle whose corners are last and current position.

F to obtain the decile (when displaying survival functions) given by next character typed (1 = 10% decile, 5 = median, etc...). This moves the cursor to the required decile in the y axis. The horizontal arrows can then be used to place the cursor on the survival curve and, after pressing return, the decile is displayed on the screen.

H to provide a menu of the options available.

K to save the graphic display for later use.

L to change line type to that define by next character typed (0 = solid, 1 = dotted, etc...).

M to mark or remember current cursor position.

P to print screen content on Epson or similar.

Q to quit graphics.

S to draw symbol defined by next character typed (0 = . , 1 = ×, 2 = o, etc...).

T to write text with NW corner at current position ending with Return.

W to keep record of the current position.

Z to zoom text size defined by next character typed (0, 1 up to 9).

Before any graphics is displayed, the query PLOTFILE? appears on the screen. Entering *Y* produces a character array required for later production of high resolution, colour plots on an HP7475A pen plotter.

1. **ESTIMATES**

   Displays the final (given all the data) means of the parameters along with the covariance and correlation matrices. This is a useful summary, particularly if the model is static.

2. **HAZARD**

   Plots smoothed mean hazard functions for given individuals. It asks for the covariates of the individuals, if there is any.

3. **LLIK**

   Displays the logarithm of the predictive likelihood for a given model.

4. ***Y* LOGPLOT *X***

   Displays a scatter plot of the logarithm of *Y* versus *X*. The arguments *X* and *Y* are vectors of the same length. It produces a + or a 0 depending on whether the individual considered died or was censored. This function is unnecessary for APL speakers (see PLOT below).

   Example: i) TIMES LOGPLOT AGE - plot log of the survival times against AGE (where AGE is a covariate);

5. **PERCENTILES**

Displays the percentiles for a survival function recently plotted on screen. In addition to the option F for the deciles, this function produces any percentile of the survival function after interactively striking the key W over the curve on the required percentage levels. It is useful if, for instance, the 1st and 3rd quartiles are required.

6. *Y* **PLOT** *X*

Displays a scattered plot of *Y* versus *X* . The arguments *X* and *Y* are as in LOGPLOT so are the symbols + and 0 for death and censoring. APL expressions are allowed.

Examples: i) AGE PLOT DEATHS - plots AGE (where age is a covariate) against the corresponding death indicator for each individual,

ii) TIMES PLOT AGE+10xDEATHS - plots the survival times against AGE + 10 x DEATHS.

7. **PLOTTER**

Produces high resolution, colour plots on a Hewlett Packard 7475A colour plotter.

8. **PREDICT**

Plots predictive features of given individuals. This function asks for the covariates of the individuals, if any, and the predictive feature required. The options are survival, hazard or density functions.

9. **REPLAY**

Displays graphics previously saved interactively with H command.

10. **TRAJECTORY**

Plots the smoothed mean with 2 s.d. limits, and on-line mean trajectories over time of a parameter. Prompts for the name of the covariate associated with the parameter, if there are any covariates in the model. Produces the plot of the baseline hazard parameter by entering the name BASELINE for the covariate or automatically if there are no covariates in the model.

CONTROL FUNCTIONS

1. (*)APLCHARS

Downloads APL characters to an Epson or similar printer for printing of APL functions. Automatically called in PRINTALL.

2. (*)DISKDRIVE

Selects disk drive for data input and/or output. Automatically called in READ and WRITE.

3. END

Finishes APL section and returns to DOS after asking for confirmation of end of session.

4. (*)GRCONFIG

Identifies graphics configuration. It asks for the type of machine being used.

5. LISTFILES

Displays a list of all APL data files on disk.

6. PRINTALL

Prints all the APL functions on Epson FX80 printer, or similar.

7. REMDATA

Erases data variables from memory, ready for reading or input of a new dataset.

8. RESET

Erases working variables from memory, ready for new analysis.

8.2 - SURVIVAL in practice

In this section, the main features of SURVIVAL are briefly illustrated through details of an APL session. The main data set used comes from Feigl & Zelen (1965) and is presented in Table 11. The wide range of variation of the white blood count suggests the

TABLE 11

Survival times (weeks) of leukaemia patients

(Feigl & Zelen, 1965)

AG positive		AG negative	
Survival time	White blood count	Survival time	White blood count
65	2300	56	4400
156	750	65	3000
100	4300	17	4000
134	2600	7	1500
16	6000	16	9000
108	10500	22	5300
121	10000	3	10000
4	17000	4	19000
39	5400	2	27000
143	7000	3	28000
56	9400	8	31000
26	32000	4	26000
22	35000	3	21000
1	100000	30	79000
1	100000	4	100000
5	52000	43	100000
65	100000		

use of a logarithmic transformation for this variable and this is how this variable is stored and used. The demonstration data files DEMO3 and DEMO4 are also used to show the data management functions. The latter one contains time-dependent covariates so that the difference to the input of data files with time-dependent covariates can be best appreciated. The two data analyses performed use a variety of different options available to users. The second is an analysis of a random sample showing the changes that automatically take place in the functions for the case of no covariates. Also note that, although the grid was deliberately specified to include time points well beyond the time span (65 weeks), this was cut down to the first 10 intervals containing the data. Figures 15-22 are exact reproductions of what appears on the screen but for the fact that in the monochrome system each colour would have been replaced by a different line type (dashed, dotted, etc.). The Figures were in fact obtained with the function PLOTTER after saving the plot file.

EXAMPLE OF INTERACTIVE SESSION WITH SURVIVAL
==========

```
IDENTIFY YOUUR GRAPHICS CONFIGURATION
INPUT EITHER A=IBM COLOUR, B=HERCULES, C=OLIVETTI M24:C


     INPUT

    IDENTIFY DISK DRIVE FOR DATA INPUT AND/OR OUTPUT
INPUT EITHER A=DEFAULT, B=SECOND FLOPPY, C=HARD DISK :A

    INPUT TITLE FOR DATA FILE : DEMO3
    INPUT DATA SET DESCRIPTION : xxxxx
    INPUT NAME OF TIME UNITS : DAYS
    INPUT NUMBER OF OBSERVATIONS : 5
    INPUT NUMBER OF COVARIATES : 2
    INPUT NAME OF COVARIATES :
NAME OF COVARIATES No. 1 : SEX
NAME OF COVARIATES No. 2 : AGE
ARE THERE ANY TIME-DEPENDENT COVARIATES ? [Yes or No] N

INPUT DATA   Y, D, Z1, Z2, ..., ZP WHERE
      Y = SURVIVAL TIME
      D = 1, WHEN DEATH IS OBSERVED AND 0, OTHERWISE
      Z1, Z2, ..., ZP = VALUES OF THE COVARIATES
OBSERVATION 1 : 10 1 1 45
OBSERVATION 2 : 37 1 1 38
OBSERVATION 3 : 26 1 0 56
OBSERVATION 4 : 78 0 0 30
OBSERVATION 5 : 5 1 1 72

INPUT A TO CREATE A NEW FILE, B TO ENTER ADDITIONAL VALUES
                          OR RETURN TO OVER-WRITE OLD FILE : A

WRITING DATA TO FILE NAMED DEMO3
DATA INPUT COMPLETE


     LOOK 2 5

   2 37 1  1 38
   5  5 1  1 72

     LOOKALL 2 5

   2 37 1  1 38
   3 26 1  0 56
   4 78 0  0 30
   5  5 1  1 72
```

EXAMPLES OF INTERACTIVE SESSION WITH SURVIVAL (Contd.)

```
     AMEND 4

OBSERVATIONS TO BE AMENDED ARE :

   4 78 0  0 30

INPUT DATA   Y, D, Z1, ..., ZP WHERE

        Y = SURVIVAL TIME
        D = 1, WHEN DEATH IS OBSERVED AND 0, OTHERWISE
        Z1, Z2, ..., ZP = VALUES OF THE COVARIATES

        OBSERVATION 4 : 7.8 0 0 30

INPUT A TO CREATE NEW FILE, B TO ENTER ADDITIONAL VALUES
                      OR RETURN TO OVER-WRITE OLD FILE :

WRITING DATA TO FILE NAMED DEMO3
DATA INPUT COMPLETE


     INPUT

    IDENTIFY DISK DRIVE FOR DATA INPUT AND/OR OUTPUT
INPUT EITHER A=DEFAULT, B=SECOND FLOPPY, C=HARD DISK : B

    INPUT TITLE FOR DATA FILE : DEMO4
    INPUT DATASET DESCRIPTION : xxxxx
    INPUT NAME OF TIME UNITS : MONTHS
    INPUT NUMBER OF OBSERVATIONS : 4
    INPUT NUMBER OF COVARIATES : 1
NAME OF COVARIATE No. 1 : SCORE
ARE THERE ANY TIME-DEPENDENT COVARIATES ? [Yes or No] Y

    INPUT DATA   Y, D, Z1, Z2, ..., ZP WHERE
            Y = SURVIVAL TIME
            D = 1, WHEN DEATH IS OBSERVED AND 0, OTHERWISE
            Z1, Z2, ..., ZP = INITIAL VALUES OF THE COVARIATES
OBSERVATION 1 : 20 0 2.67
OBSERVATION 2 : 9 1 5.40
OBSERVATION 3 : 25 1 4.78
OBSERVATION 4 : 17 1 3.42

INPUT A TO CREATE NEW FILE, B TO ENTER ADDITIONAL VALUES
                      OR RETURN TO OVER-WRITE OLD FILE : A

WRITING DATA TO FILE NAMED DEMO4
INITIAL DATA INPUT COMPLETE
```

EXAMPLE OF INTERACTIVE SESSION WITH SURVIVAL (Contd.)

```
ENTER DATA IND, TC, Z1, Z2, ..., ZP WHERE
              IND = INDEX OF INDIVIDUAL
               TC = TIME OF CHANGE
      Z1, ..., ZP = CHANGED VALUES OF THE COVARIATES
(INDICATE END OF CHANGES BY TYPING A 0 (ZERO) CHARACTER)
CHANGE No 1 : 2 5 2.89
CHANGE No 2 : 1 15 3.79
CHANGE No 3 : 4 16 2.92
CHANGE No 4 : 0
HAVE ALL CHANGES BEEN ENTERED? [Yes or No] Y

INPUT A TO CREATE NEW FILE, B TO ENTER ADDITIONAL VALUES
                    OR RETURN TO OVER-WRITE OLD FILE : B

WRITING DATA TO FILE NAMED DEMO4
FINAL DATA INPUT COMPLETE


     LISTFILES
        0 DEMO1
        0 DEMO2
        0 DEMO3
        0 WBC
        1 DEMO4


     READ

    IDENTIFY DISK DRIVE FOR DATA INPUT AND/OR OUTPUT
INPUT EITHER A=DEFAULT, B=SECOND FLOPPY, C=HARD DISK :A

 ENTER NAME OF DATA FILE TO BE USED : WBC

DATA FILE : WBC
SURVIVAL TIMES OF LEUKAEMIA PATIENTS FROM FEIGL AND ZELEN
    (BIOMETRICS, 1965, PG. 826-838). THE TWO COVARIATES ARE AN AG
    INDICATOR (1, FOR + AND 0 FOR -) AND THE LOGARITHM OF WHITE
    BLOOD COUNT.
NUMBER OF OBSERVATIONS : 33
NUMBER OF DEATHS : 33
NUMBER OF COVARIATES :2
NAMES OF COVARIATES : AG       LOGWBC
COVARIATES ARE TIME-INDEPENDENT
TOTAL TIME SPAN : 156 WEEKS


     TIMES PLOT LOGWBC
INPUT TITLE FOR Y AXIS:TIMES
INPUT TITLE FOR X AXIS:LOG (WBC)
PLOTFILE?                                   n.b. Produces Figure 15
```

EXAMPLE OF INTERACTIVE SESSION WITH SURVIVAL (Contd.)

```
 ANALYSE

 ENTER NUMBER OF COVARIATES IN THE MODEL : 2

 ENTER NAME OF COVARIATE No. 1 : AG
 ENTER NAME OF COVARIATE No. 2 : LOGWBC

     IDENTIFICATION OF THE TYPE OF GRID CONSTRUCTION

ENTER EITHER A=EQUAL LENGTHS, B=DEATH TIMES, C=ARBITRARY: B
ENTER (EQUAL) NUMBER OF DEATH TIMES PER INTERVAL: 1

     INITIAL PRIOR MOMENTS SETTING

DO YOU WANT TO INITIATE WITH VAGUE PRIOR MOMENTS? [Yes or No] N

 ENTER LOWER AND UPPER LIMITS FOR PARAMETERS
(n.b. Equal limits implies parameter fixed at time origin)

 BASELINE HAZARD : -1 1
 AG       EFFECT : -1 1
 LOGWBC   EFFECT : -1 1

     SELECTION OF THE TYPE OF EVOLUTION

ENTER A FOR ADDITIVE VARIANCE MATRIX OR B FOR DISCOUNT : B
ENTER NUMBER OF WEEKS USED AS BASE FOR EVOLUTION SETTING : 1

ENTER DISCOUNT FACTORS OVER INTERVALS OF 1 WEEKS

 BASELINE HAZARD : .98
 AG       EFFECT : .99
 LOGWBC   EFFECT : .99

     ESTIMATION PROCEDURES UNDERWAY

 STARTING CYCLE No. 1 OUT OF A TOTAL OF 23

 ESTIMATION COMPLETED

 SMOOTHING COMPLETED


 TRAJECTORY

ENTER NAME OF COVARIATE : BASELINE
PLOTFILE?                                  n.b. Produces Figure 16
```

EXAMPLE OF INTERACTIVE SESSION WITH SURVIVAL (Contd.)

```
     TRAJECTORY

ENTER NAME OF COVARIATE : AG
PLOTFILE?                                   n.b. Produces Figure 17

     ESTIMATES

PARAMETERS ARE : BASELINE HAZARD
                 AG       EFFECT
                 LOGWBC   EFFECT

MEAN VECTOR          : ¯1.442 ¯0.9963 ¯0.1276

STANDARD ERRORS      : 0.9461 0.6102 0.1042

VARIANCE MATRIX      :
 0.8952    ¯0.1676    ¯0.0519
¯0.1676     0.3724    ¯0.02251
¯0.0519    ¯0.02251    0.01085

CORRELATION MATRIX   :
 1         ¯0.2903   ¯0.5267
¯0.2903     1        ¯0.3542
¯0.5267    ¯0.3542    1

     PREDICT

ENTER NUMBER OF CASES : 4

ENTER VALUES OF THE COVARIATES
    CASE No. 1 : 0,850000
    CASE No. 2 : 0,850000
    CASE No. 3 : 1,850000
    CASE No. 4 : 1,850000
    SELECT PREDICTIVE FEATURE TO BE DISPLAYED
ENTER H FOR HAZARD, S FOR SURVIVAL OR F FOR DENSITY FUNCTION: S
PLOTFILE?                                   n.b. Produces Figure 18

WOULD YOU LIKE TO SEE ANOTHER PREDICTIVE FUNCTION? [Yes or No] N

     HAZARD

ENTER NUMBER OF CASES : 2

ENTER VALUES OF COVARIATES
    CASE No. 1 : 0,820000
    CASE No. 2 : 1,820000
PLOTFILE?                                   n.b. Produces Figure 19
```

EXAMPLE OF INTERACTIVE SESSION WITH SURVIVAL (Contd.)

```
     AG SELECT 0

     DATASET

DATA FILE : WBC
SURVIVAL TIMES OF LEUKAEMIA PATIENTS FROM FEIGL AND ZELEN
    (BIOMETRICS, 1965, PG. 826-838). THE TWO COVARIATES AR AN AG
    INDICATOR (1, FOR + AND 0, FOR -) AND THE LOGARITHM OF WHITE
    BLOOD COUNT.
NUMBER OF OBSERVATIONS : 16
NUMBER OF DEATHS : 16
NUMBER OF COVARIATES : 2
NAMES OF COVARIATES : AG        LOGWBC
TOTAL TIME SPAN : 65 WEEKS

     ANALYSE

    ENTER NUMBER OF COVARIATES IN THE MODEL : 0

        IDENTIFICATION OF THE TYPE OF GRID CONSTRUCTION

ENTER EITHER A=EQUAL LENGTHS, B=DEATH TIMES, C=ARBITRARY: A
ENTER EQUAL LENGTH OF INTERVAL: 5
DOES THIS INTERVAL LENGTH GO UP TO THE END? [Yes or No] N
ENTER NUMBER OF INTERVALS WITH THIS LENGTH: 6

ENTER NEXT INTERVAL LENGTH: 10
ENTER NUMBER OF INTERVALS WITH THIS LENGTH: 30

        INITIAL PRIOR MOMENTS SETTING

DO YOU WANT TO INITIATE WITH VAGUE PRIOR MOMENTS? [Yes or No] N

    ENTER LOWER AND UPPER LIMITS FOR PARAMETERS
(n.b. Equal limits implies parameter fixed at time origin)

    BASELINE HAZARD : -1  1

        SELECTION OF THE TYPE OF EVOLUTION

ENTER A FOR ADDITIVE VARIANCE MATRIX OR B FOR DISCOUNT : A
ENTER NUMBER OF WEEKS USED AS BASE FOR EVOLUTION SETTING : 1

    SPECIFICATION OF THE ADDITIVE VARIANCE MATRIX
ENTER A FOR ARBITRARY OR B FOR COMPONENTS SETTING : B
ENTER NUMBER OF PRINCIPAL COMPONENTS(PC) TO BE USED : 1
ENTER 1-DIMENSIONAL PC VECTOR No 1 : .1

        ESTIMATION PROCEDURES UNDERWAY
```

EXAMPLE OF INTERACTIVE SESSION WITH SURVIVAL (Contd.)

```
STARTING CYCLE No. 1 OUT OF A TOTAL OF 10

ESTIMATION COMPLETED

SMOOTHING COMPLETED

 LLIK

THE VALUE OF THE LOG-LIKELIHOOD OF THIS MODEL IS -68.48


 TRAJECTORY
PLOTFILE?                                   n.b. Produces Figure 20


 PREDICT
SELECT PREDICTIVE FEATURE TO BE DISPLAYED
ENTER H FOR HAZARD, S FOR SURVIVAL OR F FOR DENSITY FUNCTION: S
PLOTFILE?                                   n.b. Produces Figure 21

WOULD YOU LIKE TO SEE ANOTHER PREDICTIVE FUNCTION? [Yes or No] Y
SELECT PREDICTIVE FEATURE TO BE DISPLAYED
ENTER H FOR HAZARD, S FOR SURVIVAL OR F FOR DENSITY FUNCTION: F
PLOTFILE?                                   n.b. Produces Figure 22

WOULD YOU LIKE TO SEE ANOTHER PREDICTIVE FUNCTION ? [Yes or No] N


 PERCENTILES

  PROB. OF SURVIVAL          WEEKS

          0.25               20.7
          0.50                8.9
          0.75                3.3


 REMDATA

 RESET

 END
DO YOU REALLY WANT TO END? [Yes or No] Y


A>
```

Figure 15

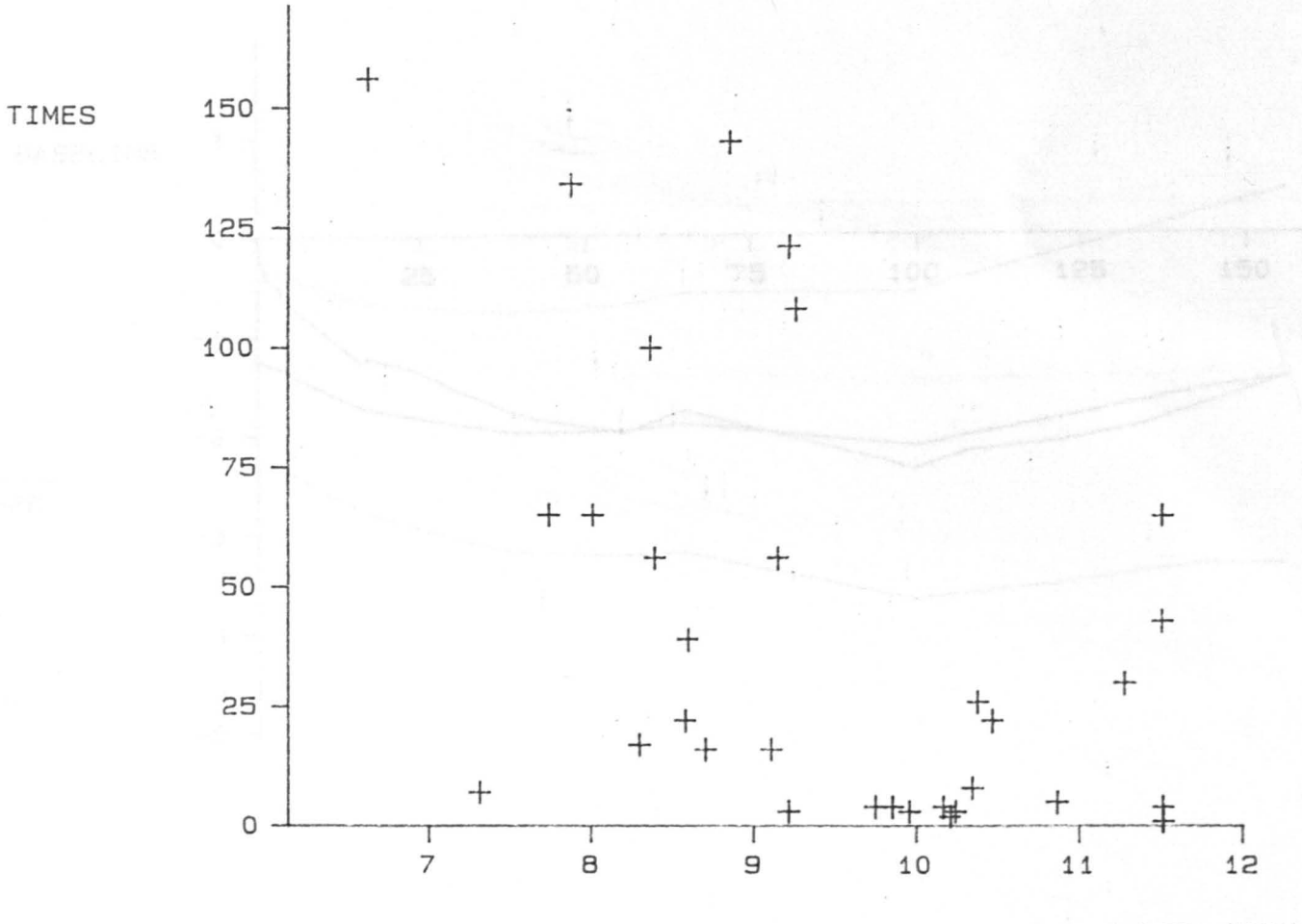


Figure 16

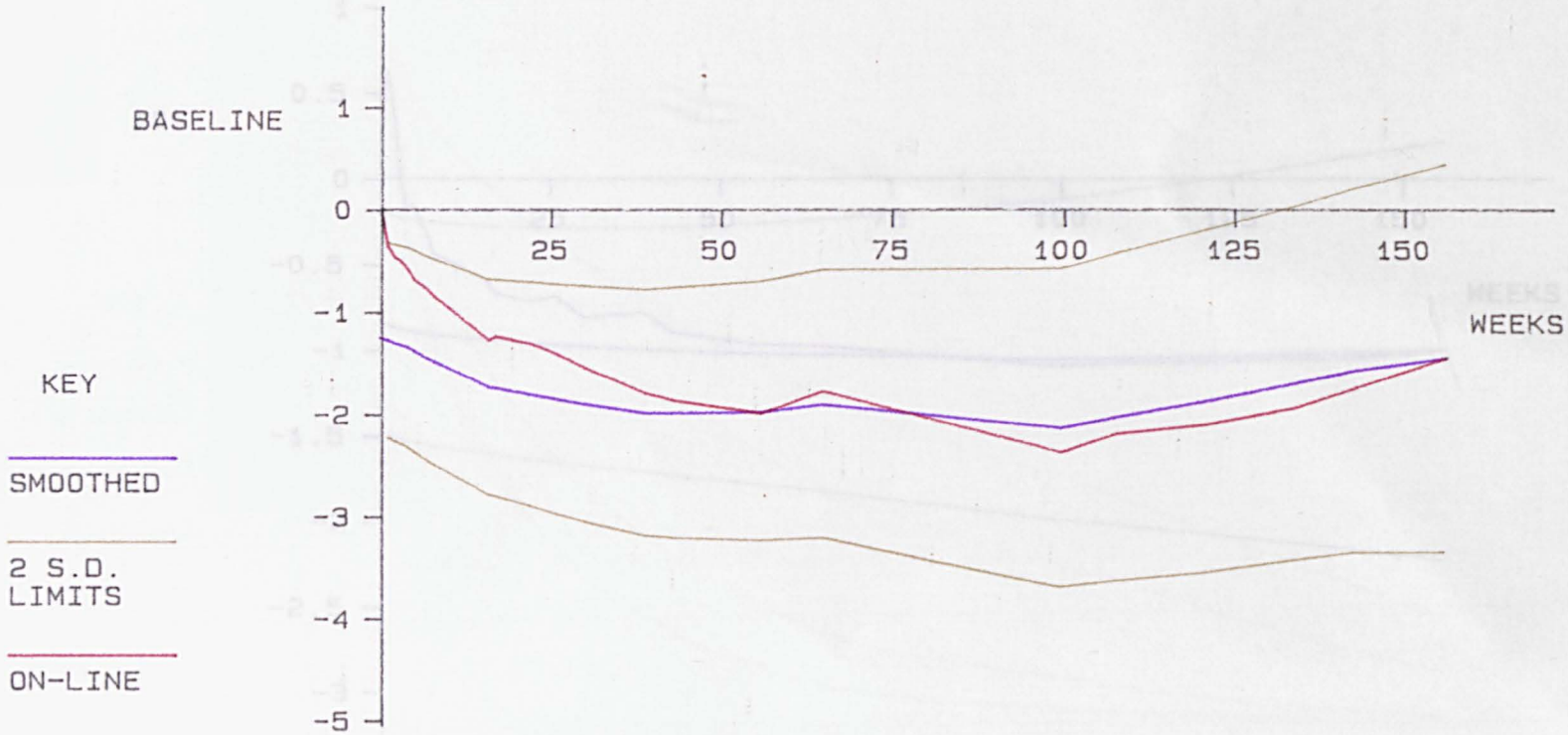

BASELINE
1
0
-1
-2
-3
-4
-5
25
50
75
100
125
150
WEEKS
KEY
SMOOTHED
2 S.D. LIMITS
ON-LINE

Figure 17

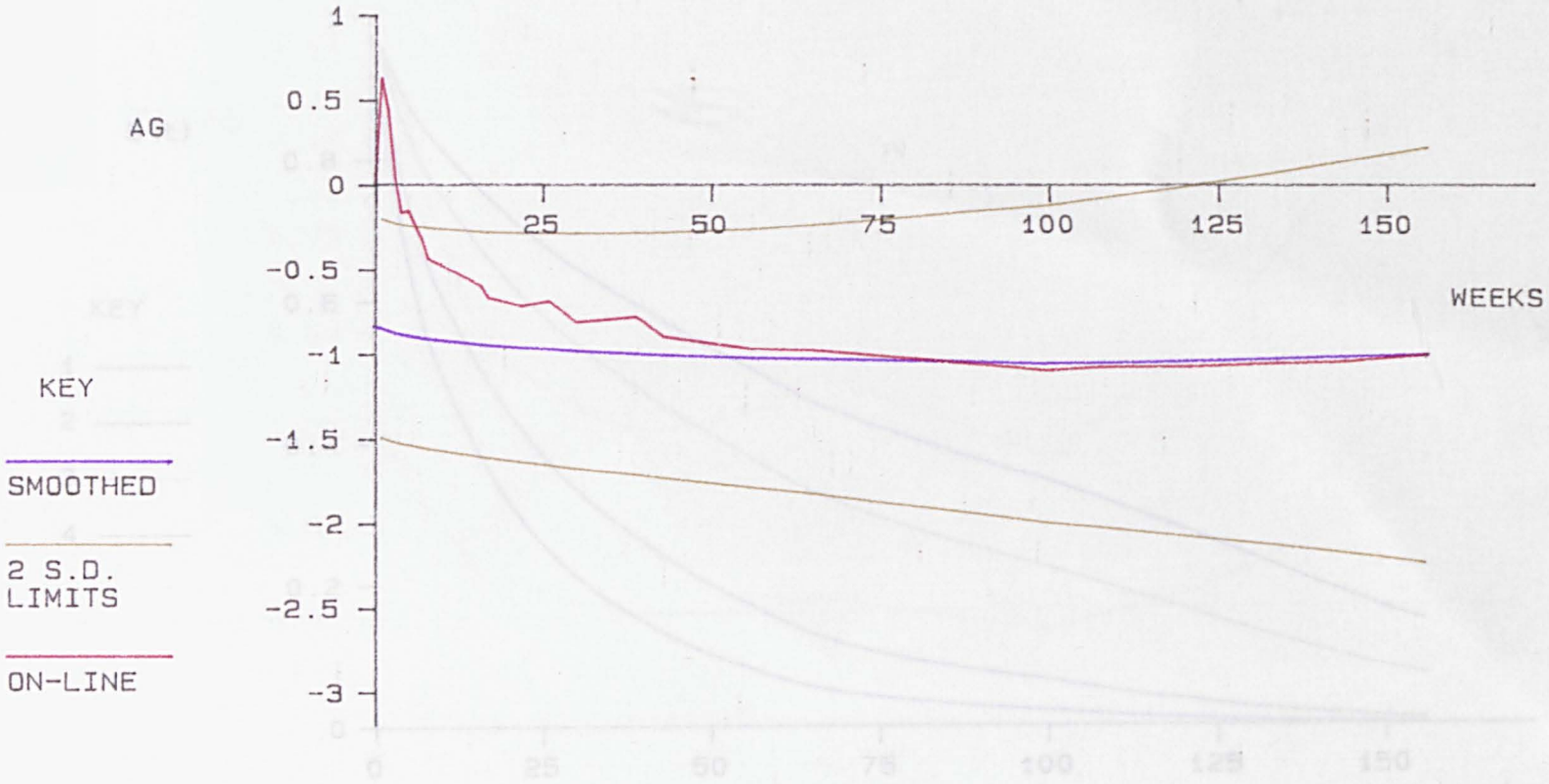

AG
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
25
50
75
100
125
150
WEEKS
KEY
SMOOTHED
2 S.D. LIMITS
ON-LINE

Figure 18

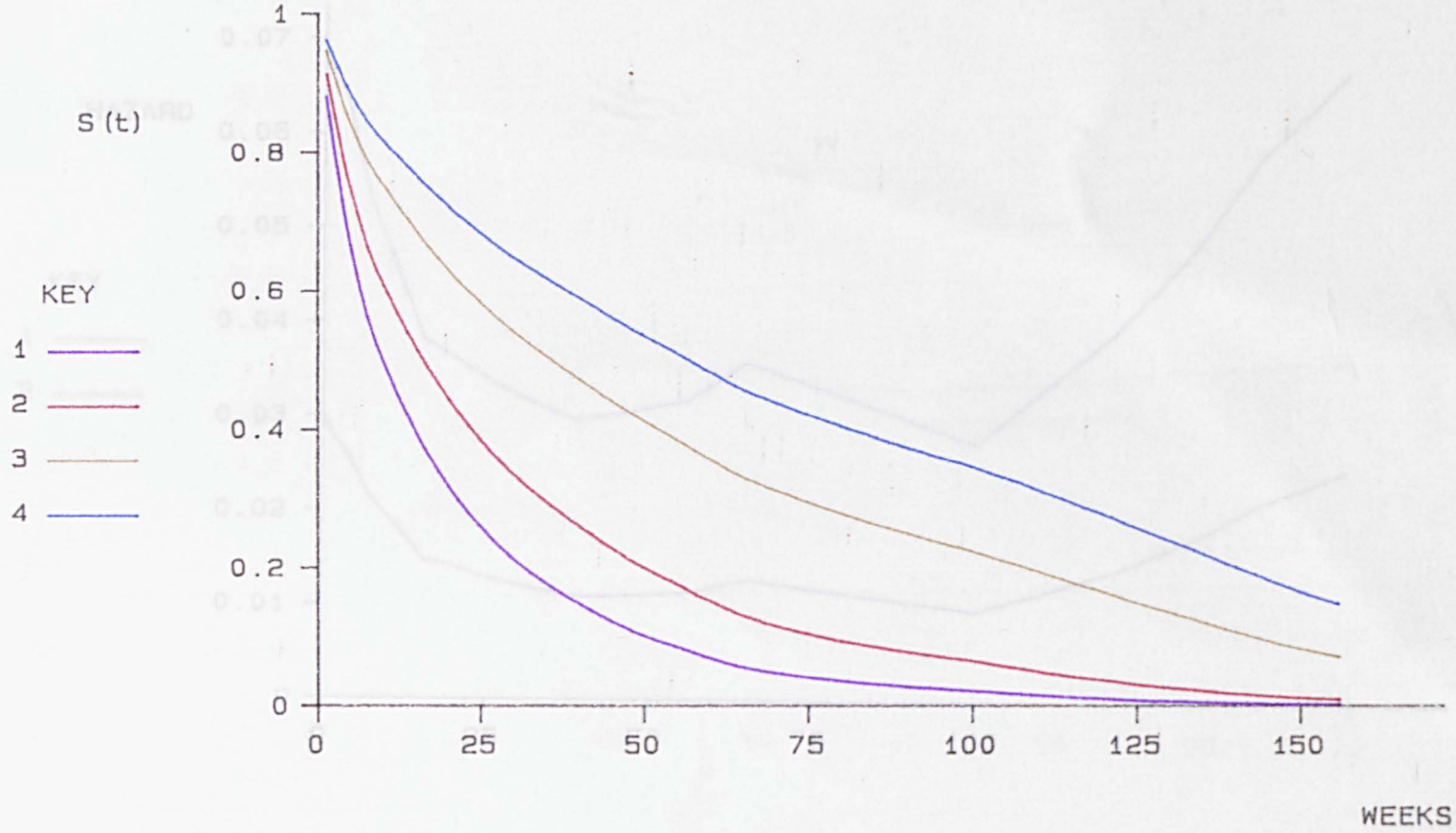


Figure 19

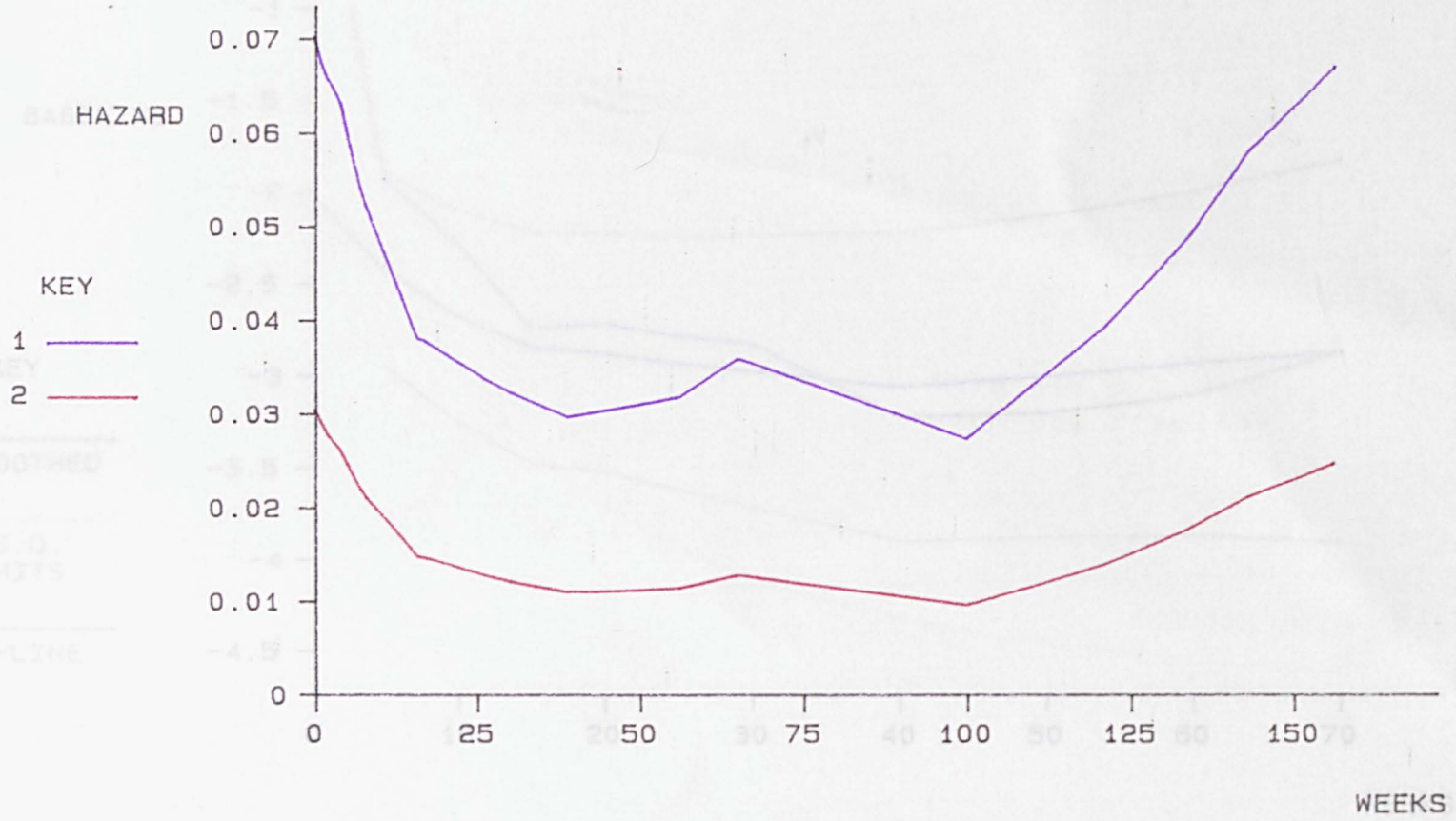

HAZARD
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
KEY
1
2
0
25
50
75
100
125
150
WEEKS

Figure 20

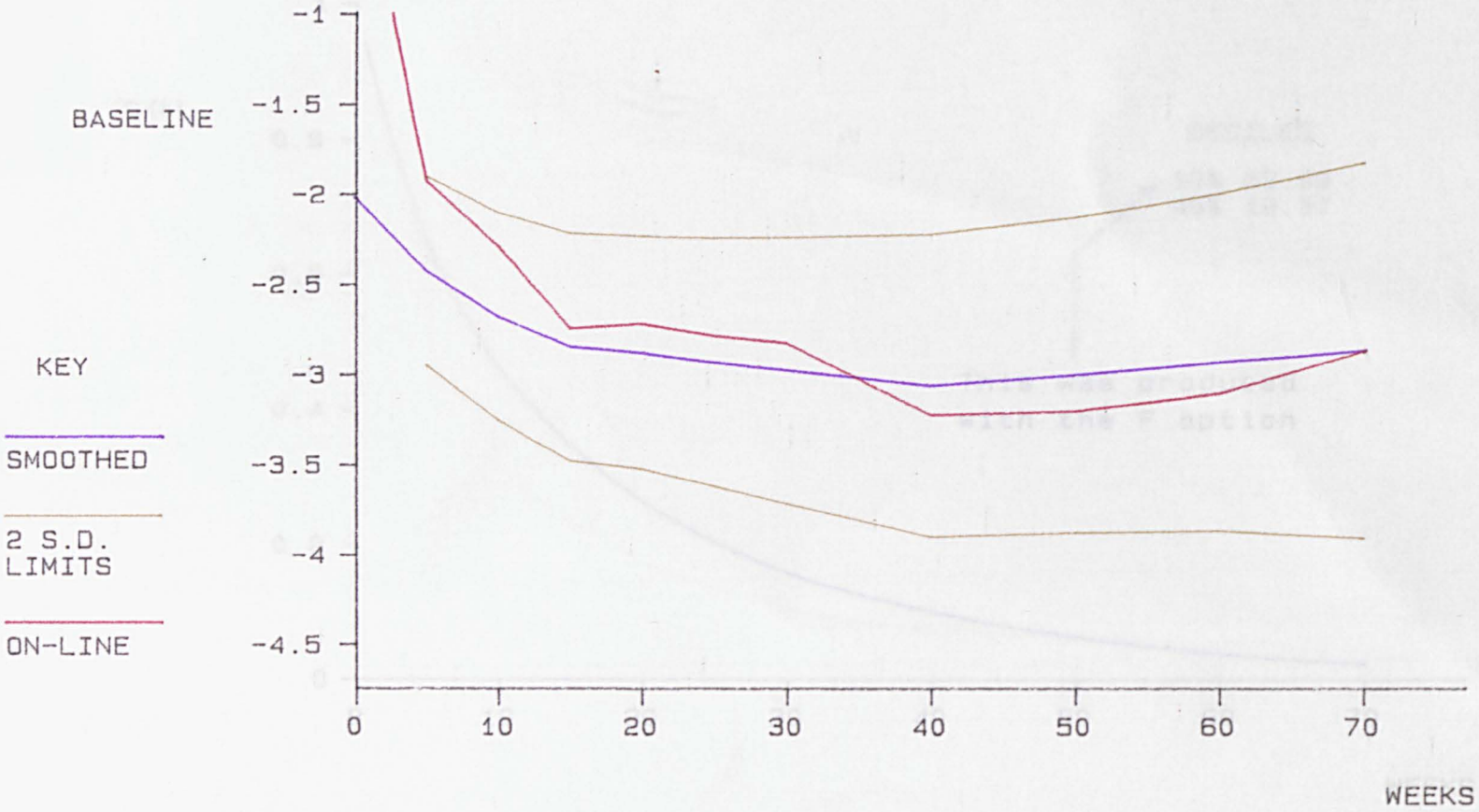

BASELINE
KEY
SMOOTHED
2 S.D. LIMITS
ON-LINE
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4.5
0
10
20
30
40
50
60
70
WEEKS

Figure 21

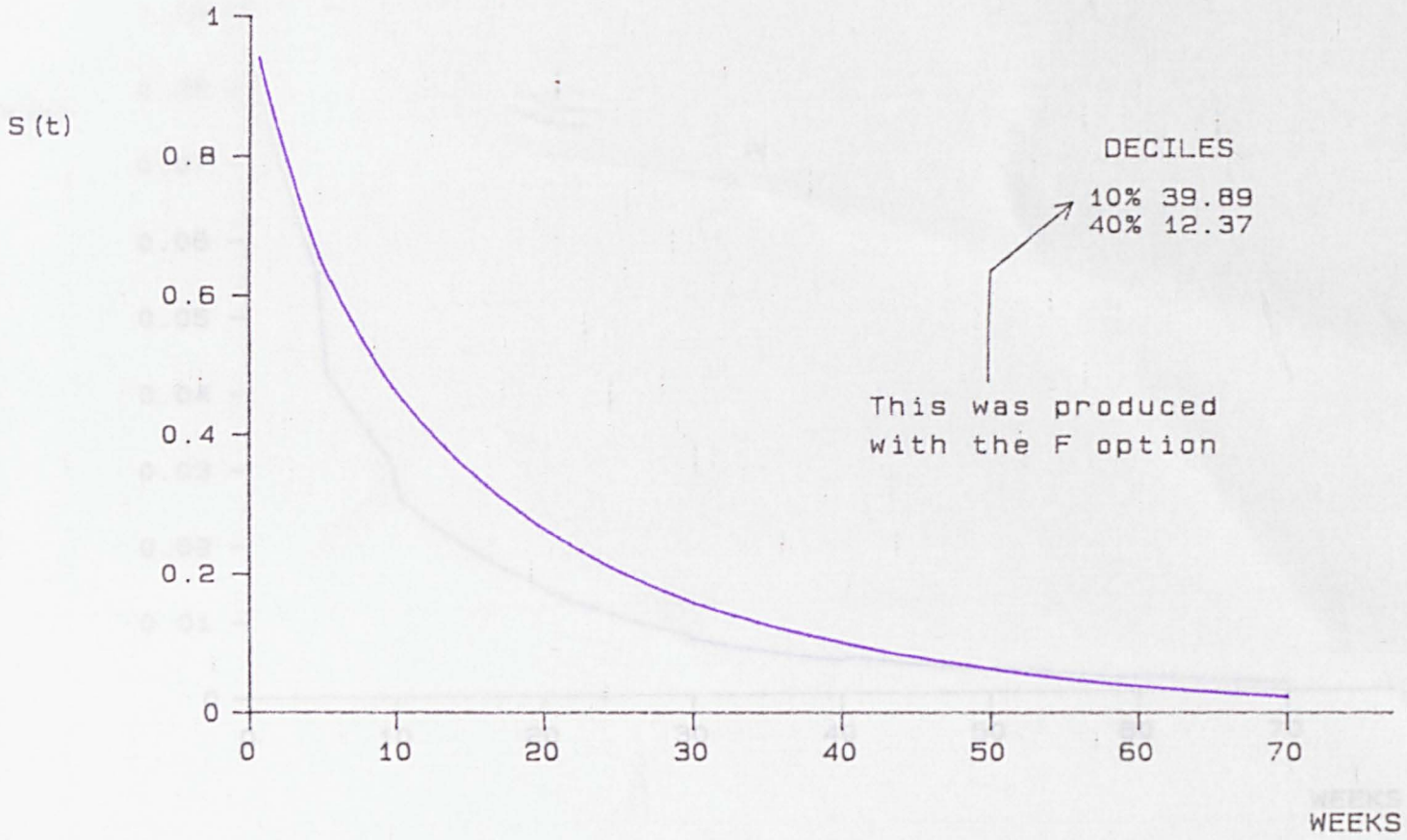
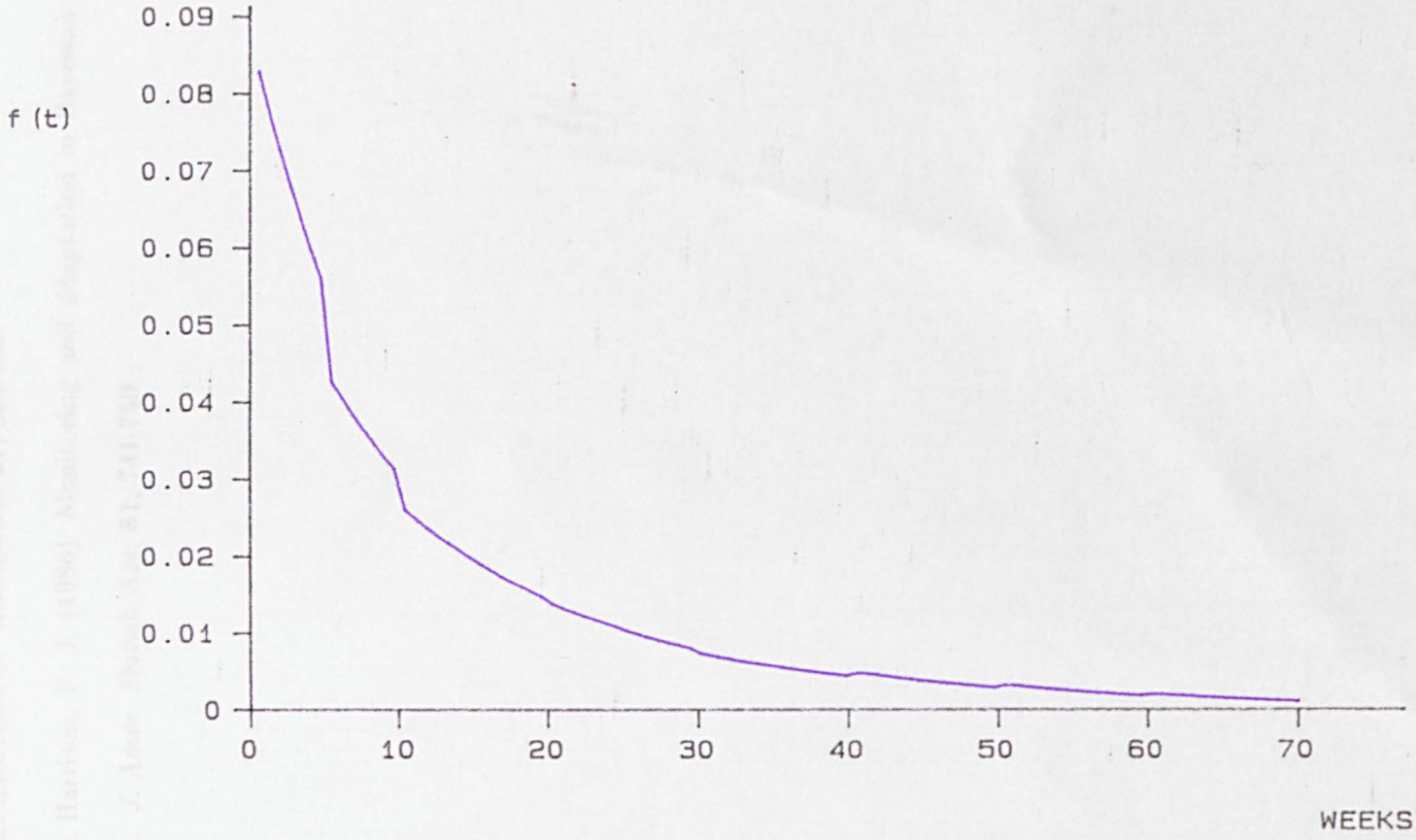

S(t)
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
WEEKS
DECILES
10% 39.89
40% 12.37
This was produced
with the F option

Figure 22


f (t)
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30
40
50
60
70
WEEKS

References

Feigl, P. & Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, **21**, 826-838.

West, M. & Harrison, P. J. (1986). Monitoring and adaptation in Bayesian forecasting models. *J. Amer. Statist. Ass*, **81**, 741-750.