RICE UNIVERSITY

A Dynamic Model for Survival Data with Longitudinal Covariates

by

Krzysztof Janusz Rudnicki

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Dr. Gary L. Rosner, Adviser
Professor of Biostatistics
UT MD Anderson Cancer Center

Dr. Katherine B. Enson, Chair
Professor of Statistics
Rice University

Dr. Javier Rojo
Professor of Statistics
Rice University

Dr. Robin C. Sickles
Professor of Economics
Rice University

HOUSTON, TEXAS

MAY, 2006
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 3216768
Copyright 2006 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P. O. Box 1346
Ann Arbor, MI 48106-1346
Abstract

A Dynamic Model for Survival Data with Longitudinal Covariates

by

Krzysztof Janusz Rudnicki

Analyses involving both longitudinal and time-to-event data are quite common in medical research. The primary goal of such studies may be to simultaneously study the effect of treatment on both the longitudinal covariate and survival, but secondary objectives, such as understanding the within-patients patterns of change of the time-dependent marker, or the relationship between the marker's profiles and the occurrence of the event of interest, are often considered.

Currently available methods of analyzing survival and longitudinal data usually introduce many undesirable and sometimes unreasonable assumptions. We introduce two flexible Bayesian hierarchical modeling approaches for analyzing these two types of data by use of dynamic models and survival analysis methods. In both approaches the longitudinal covariate is modeled via dynamic hierarchical models which allow the shapes of the longitudinal trajectories to be determined by the data rather than by the
assumed parametric model. The trajectories are patient specific, and the link between them is provided by the hierarchical structure of the model, allowing borrowing of strength across patients. In the survival part of the model, the first method, referred to as 2-stage model, uses the estimates of the longitudinal trajectories obtained in the first stage of the analysis, as a time-dependent covariate in the Cox PH model, to find the estimates of the corresponding survival model parameters. The second approach, called the joint model, assumes the piecewise exponential distribution for the event times of the patients and uses a discretized version of the Cox PH model. Some of the parameters of this survival model are also allowed to change over time, and again, dynamic models provide the description of the stochastic evolution of these parameters. A combination of various MCMC techniques is used to obtain a sample from the joint posterior distributions of all the model parameters. This distribution combines the likelihood of the longitudinal and survival data and the prior knowledge about the parameters. Simulation studies provide the measure of the quality of the method and both models are compared to one of the currently existing approaches.
Acknowledgements

First, I thank my adviser Gary Rosner, for his continuous support and guidance throughout the Ph.D. program. Gary was always there to listen, give advice and encourage. I thank him for his direction and patience, especially in responding to my numerous, sometimes lengthy, emails. His help in completing this project was invaluable. Gary also played a significant part in shaping my statistical interests and had confidence in me when I doubted myself. I feel truly privileged to have been able to work with him for the past few years.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Kathy Ensor, Dr. Javier Rojo, and Dr. Robin Sickles, for allowing me to pursue this project in the first place. Also, for providing insightful comments, good questions, and being available for guidance and reviewing my work on a very short notice. I would especially like to thank Dr. Ensor, for supporting both my research and teaching activities, and for guiding me through my first steps in teaching.

I would also like to say “thank you” to Dr. Marek Kimmel, without whose initiative I would have never left a little town in southeastern Poland and come to the large state of Texas to pursue a doctoral degree in statistics at such an outstanding school like Rice University. Only a Pole will ever fully understand how great, but
unlikely opportunity that was.

My special thanks go to my friends Adam and Beata Bobrowski, and Darek and Magda Skorczewski, who opened to me not only their homes, but also their refrigerators. I am very grateful for all their care, and both emotional and spiritual support during these years. I am especially indebted to Adam, who at the right moment mentioned my name to Dr. Kimmel, and thanks to whom, together with Dr. Kimmel, I was able to realize my growing interest in statistics. Thank you, Adam! For continuous help, guidance, and support, on both personal and scientific level.

I cannot forget about my numerous friends: Mike and Meredith, for all the fun we had together, and for their ever optimistic attitude, which made me smile even during the hardest moments; Steve and Brandee, Steve and Holly, Vinu and Sarah, for always being there for me and never forgetting about me; Clem, for priceless guidance and support; and many many others in many cities, countries and continents. You know who you are. Thank you all for your invaluable friendship and support over all these years.

I am eternally indebted to my entire family: my parents, Genowefa Rudnicka and Marian Rudnicki, for giving me life in the first place, for creating the best family I could have ever imagined to be born into, for passing on their values and faith to
me, for their unconditional love, support and encouragement to pursue my interests, even if they went beyond the boundaries of language and geography. My siblings and their families: my brother Robert, for always believing in me and shaping my musical taste; my sister Ewa, for her love, prayers and giving me a spiritual birth; and finally, my sister Kasia, who, together with her husband Philip, was an inexhaustible source of care, love and support on the foreign soil. I have no words to describe how much you all mean to me. Thank you!

Last, but definitely not least, I want to thank God, for being constantly with me on this long and not always easy journey, for His rich supply, for not giving up on me, for all His blessings, and for all these wonderful people He has put on my way during those years who continue to enrich my life every day.
Contents

Abstract ii

Acknowledgements iv

List of Figures xii

List of Tables xiv

1 Background 1

1.1 Introduction to Survival Analysis ............................ 1

1.1.1 Basic Mathematical Concepts ............................ 3

1.1.2 Cox Proportional Hazards Model ......................... 4

1.1.3 The Bayesian Approach ................................. 7

1.1.4 Remarks ........................................... 9

1.2 Introduction to Dynamic Models ........................... 11

1.2.1 The Dynamic Linear Model ............................ 13

1.2.2 Linear Growth Models ............................... 20
1.2.3 Dynamic Hierarchical Models ............................ 23

1.3 A Quick Overview of Simulation Methods ..................... 27
  1.3.1 Gibbs Sampling ........................................ 27
  1.3.2 Metropolis-Hastings Algorithm .......................... 28
  1.3.3 Forward Filtering Backward Sampling Algorithm ........... 29
  1.3.4 Auxiliary Particle Filters ............................... 30
  1.3.5 Remarks .............................................. 34

1.4 Previous Research on Modeling of Longitudinal and Time to Event Data 35
  1.4.1 Two-stage Method ....................................... 35
  1.4.2 Joint Modeling .......................................... 38
  1.4.3 Remarks .............................................. 47

1.5 Previous Research on Dynamic Models for Survival Data .......... 48

2 Preliminary Analysis ............................................ 55
  2.1 The Gastric Cancer Patients Data Analysis .................... 55
    2.1.1 The PE Model ......................................... 56
    2.1.2 The Lognormal Model .................................. 64

2.2 The Sales Data Analysis ....................................... 70

3 A Dynamic Model for Survival Data with Longitudinal Covariates 77
  3.1 Introduction .............................................. 77
  3.2 General Setup, Notation and Assumptions ...................... 78
3.3 The 2 stage Model ................................................. 81
  3.3.1 Longitudinal Data Model .............................. 81
  3.3.2 Survival Data Model .................................. 85
  3.3.3 The Joint Posterior Distribution ...................... 87
  3.3.4 Remarks .................................................. 89
3.4 The Joint Model ................................................. 90
  3.4.1 Longitudinal Data Model .............................. 90
  3.4.2 Survival Data Model .................................. 90
  3.4.3 The Joint Likelihood .................................. 91
  3.4.4 The Joint Posterior Distribution ...................... 92
  3.4.5 Remarks .................................................. 94

4 Simulation Study .................................................. 95
  4.1 The Data ..................................................... 95
    4.1.1 Generating the Data ................................ 96
  4.2 Computation ................................................. 99
    4.2.1 Handling the Ties in the Data ..................... 99
    4.2.2 The Priors ........................................... 99
    4.2.3 Simulation ........................................... 100
  4.3 Results ..................................................... 100
    4.3.1 Posterior Estimates of $\sigma^2, \sigma_\phi^2, \sigma_a^2, \eta, m_\phi, m_a, c_\phi^2,$ and $\beta$ ........ 100
4.3.2 Posterior Estimates of Θ and Γ: Overall and Patient Specific

Posterior Mean Trajectories ........................................ 102

4.3.3 Posterior Predictive Mean Trajectories ...................... 107

4.3.4 Baseline Hazard and Hazard .................................. 107

4.4 Sensitivity Analysis .................................................. 109

4.4.1 Varying the Amount of Noise in the Data ................. 109

4.4.2 Sensitivity to the Prior Moments ............................. 115

4.5 Comparison to an Existing Method .............................. 120

4.5.1 The Method ...................................................... 120

4.5.2 The Priors ....................................................... 123

4.5.3 Simulation ....................................................... 123

4.5.4 Results .......................................................... 124

4.6 Remarks ............................................................. 127

5 Conclusions and Future Considerations ......................... 129

5.1 Advantages ......................................................... 129

5.2 Future Considerations ............................................. 131

A Sampling from the the Joint Posterior Distribution .......... 134

A.1 2-stage Model ...................................................... 134

A.1.1 The Posterior Distribution of Φ ............................... 135

A.1.2 Full Conditionals for Φ ....................................... 135
A.2 Joint Model ................................................. 141
   A.2.1 The Posterior Distribution of $(\psi, \Phi)$ ................. 141
   A.2.2 Full Conditionals for $(\psi, \Phi)$ ......................... 141
A.3 Ibrahim et al. (2001) method ................................. 144
   A.3.1 Likelihood ............................................... 144
   A.3.2 Prior for $\Phi$ ........................................ 144
   A.3.3 The Posterior Distribution of $\Phi$ ....................... 145
   A.3.4 Full Conditionals for $\Phi$ .............................. 145

Bibliography ..................................................... 148
List of Figures

2.1 Estimated posterior mean trajectories for the treatment difference effect 60

2.2 Estimated posterior variance trajectories for the treatment difference effect ........................................ 63

2.3 Predictive survival according to treatment from PE and lognormal model 67

2.4 Company sales ($Y_t$) vs. total market sales ($F_t$) ................................................................. 71

2.5 One-step ahead forecast errors ................................................................. 72

2.6 On-line estimated trajectories of $\theta_t$ with 90% HPD ................................. 74

2.7 Ratio $Y_t/F_t$ vs. time, with on-line and smooth estimates .................. 74

4.1 Sample generated longitudinal trajectories ........................................ 97

4.2 Kaplan-Meier estimate of the sample generated data .......................... 98

4.3 Overall posterior mean trajectories: 2-stage and joint method ......... 103

4.4 Posterior mean slopes of the piecewise linear population longitudinal profile over time: 2-stage and joint method ........................................ 103

4.5 Patient specific posterior mean trajectories: 2-stage and joint method 106

4.6 Posterior predictive mean trajectories: 2 stage and joint method ... 106
4.7 Posterior means of log(baseline hazard) over time .................. 108
4.8 Estimated vs. "true" hazard: joint model .......................... 108
4.9 Overall posterior mean longitudinal trajectories ($\sigma_\epsilon^2 = 5$ vs. $\sigma_\epsilon^2 = 20$) 113
4.10 Posterior means of the slopes of the overall longitudinal trajectory

($\sigma_\epsilon^2 = 5$ vs. $\sigma_\epsilon^2 = 20$) ........................................ 113
4.11 Posterior predictive mean trajectories ($\sigma_\epsilon^2 = 5$ vs. $\sigma_\epsilon^2 = 20$) ........ 114
4.12 Posterior means of log(baseline hazard) ($\sigma_\epsilon^2 = 5$ vs. $\sigma_\epsilon^2 = 20$) ........ 114
4.13 Overall posterior mean trajectories: 2 stage and joint vs. Ibrahim

ct al. (2001) ............................................................... 125
4.14 Patient specific posterior mean trajectories: 2-stage and joint vs. Ibrahim

et al. (2001) ............................................................... 126
4.15 Posterior predictive mean trajectories: 2-stage and joint vs. Ibrahim

et al. (2001) ............................................................... 126
List of Tables

2.1 On-line and smoothed posterior means and variances of the treatment difference effect parameter (Gamerman (1991) method) ............... 59

2.2 On-line and smoothed posterior means and variances of the treatment difference effect parameter (auxiliary particle filter method) .... ... 62

2.3 On-line and smoothed posterior means of the parameters (auxiliary particle filter method) .................................................. 69

2.4 Sales data example results ................................................. 75

4.1 Simulation study results (2-stage and joint method, \( \sigma_2^2 = 20 \)) ....... 101

4.2 Simulation study results (joint method, \( \sigma_2^2 = 5 \) vs. \( \sigma_2^2 = 20 \)) .... 110

4.3 Sensitivity analysis results (\( \sigma_2^2 = 20 \)) .......................... 116

4.4 Simulation study results (Ibrahim et al. (2001) method) ................. 124
Chapter 1

Background

1.1 Introduction to Survival Analysis

Survival analysis, or in other words the analysis of time-to-event data, refers to a set of statistical methods for which the response variable of interest is the time from the beginning of follow-up of an individual until an event occurs. In biomedical setting, the event may be death, incidence of a disease or disease symptoms, relapse, recovery, etc. Let us from now on refer to the time variable as survival time, and to the event of interest as failure. (Kleinbaum and Klein (2005))

An important concept of survival analysis is the one of censoring. If for any reason we do not know the exact survival time of a patient, although we may have some information about it, censoring occurs. The most common kind of censoring is \textit{right-censoring}, in which the true, but unknown, survival time of a patient is only
known to be greater than or equal to a certain value. Kleinbaum and Klein (2005) list three main reasons for an observation to be right-censored: 1) a patient has not experienced the failure by the time of the analysis, 2) a patient is lost to follow-up during the data collection process, 3) a patient withdraws from the study for any other reason than the experience of the event of interest. Analogously to right-censoring, if the exact value of an observation is only known not to exceed a certain value, we say that the observation is left-censored. An example of this kind of censoring may be when the exact time of the entry to the study of an individual is not known. Both right and left-censoring may occur at the same time causing the observation to be doubly censored. This happens when the follow-up period started after the process of interest began, and it ended before the event occurred. Hence, all we know is that the time until the event is at least as long as the follow-up period. Another common type of censoring is interval censoring, which occurs when all we know about the true value of an observation is that it occurred within a specified time interval.

Typically, it is assumed that censoring is non-informative, i.e., the censoring time is independent of failure, meaning that it does not provide us with any extra information about the exact time of failure, besides the fact that it is greater (or less) than or equal to a certain value, or that it lies within a certain time interval, depending on the type of censoring in the data. This assumption is sometimes over-simplistic in that it may ignore some important information hidden in the censoring mechanism and may lead to biased results.
The following section illustrates some more specific aspects of survival analysis and is based on Ibrahim et al. (2001), section 1.4.

1.1.1 Basic Mathematical Concepts

Following the notation by Ibrahim et al. (2001), suppose $T$ is a continuous non-negative random variable corresponding to the survival times of a group of individuals. Let $f(t)$ and $F(t)$ denote the probability density function (pdf) and the cumulative distribution function (cdf) of $T$, respectively. One of the two most important functions of survival analysis, which gives us the probability that an individual will survive beyond time $t$, is called the survival function, and is defined by

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(u) \, du. \quad (1.1)$$

$S(t)$ has the following properties:

1. it is a monotone non-increasing function,

2. $S(0) = 1$, i.e., at the beginning of the study all individuals are alive,

3. $S(\infty) = \lim_{t \to \infty} S(t) = 0$, i.e., if the study were carried on indefinitely, no one alive would be left in the study.

The second important function, called the hazard function, is “the instantaneous rate of failure at time $t$” (Ibrahim et al. (2001), p. 14), and is defined by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}. \quad (1.2)$$
It should be noted that $\lambda(t)$ is not a conditional probability of failure, but a conditional rate of failure. It is greater than 0, but unlike the probability of any event, it is not bounded by 1 from above. However, $\lambda(t)\Delta t$ “is the approximate probability of failure in $(t, t + \Delta t]$, given survival up to time $t$” (Ibrahim et al. (2001), p. 14). Each of the four so far presented functions, i.e., $f(t), F(t), S(t)$, and $\lambda(t)$, uniquely specifies the distribution of $T$, and all of these specifications are mathematically equivalent. It is possible to express $S(t)$ and $f(t)$ in terms of $\lambda(t)$. From (1.2) we have

$$\lambda(t) = \frac{S'(t)}{S(t)} = -\frac{d}{dt} \log(S(t)).$$

(1.3)

Thus, it follows that

$$S(t) = \exp \left(- \int_0^t \lambda(u) \, du \right) = \exp(-\Lambda(t)).$$

(1.4)

where $\Lambda(t) = \int_0^t \lambda(u) \, du$ is called cumulative hazard function. By the third property of $S(t)$ we have that $\Lambda(\infty) = \infty$. Hence, $\lambda(t) \geq 0$ and $\int_0^\infty \lambda(t) \, dt = \infty$. Finally, combining (1.2) and (1.4) we get

$$f(t) = \lambda(t) \exp\left(- \int_0^t \lambda(u) \, du \right).$$

(1.5)

1.1.2 Cox Proportional Hazards Model

Typically, the data in the survival analysis context consist not only of the survival times of the patients, but also of a set of covariates, some of which may be time-dependent. The relationship between the covariates and the risk of failure is often of particular interest to the investigator. Thus, a model that allows us to incorporate
the covariate information is desirable. Cox (Cox (1972)) suggested a model that accommodates the covariate component in the following way. Following Ibrahim et al. (2001), let \( \mathbf{x} \) be the vector of covariate values of an individual, \( \mathbf{\beta} \) – a vector of regression coefficients, and \( \lambda_0(t) \) – the baseline hazard function at \( t \). Then the hazard at time \( t \) for the individual is given by

\[
\lambda(t|\mathbf{x}) = \lambda_0(t) \exp\{g(\mathbf{x}, \mathbf{\beta})\},
\]

where \( g(\cdot) \) is any function of \( \mathbf{x} \) and \( \mathbf{\beta} \). Usually \( g(\cdot) \) is taken to be a linear function, so that (1.6) can be written as

\[
\lambda(t|\mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}'\mathbf{\beta}).
\]

The model’s main assumption is that given no time-dependent covariates, the hazard ratio for two individuals is constant over time. In particular, if \( g(\cdot) \) is linear, then the hazard ratio for two individuals with covariate vectors \( \mathbf{x}_1 \) and \( \mathbf{x}_2 \), depends on the difference \( \mathbf{x}_1'\mathbf{\beta} - \mathbf{x}_2'\mathbf{\beta} \) at any time \( t \). (Ibrahim et al. (2001))

Assuming right-censoring, the likelihood function for the above model can be constructed as follows. Adopting the notation by Ibrahim et al. (2001), let \( n \) be the number of patient under study, and \( t_i \) and \( c_i \) – true (possibly unobserved) failure time and potential censoring time of the \( i^{th} \) individual, respectively. The observed survival data is thus given by \( n \) pairs \((y_i, \delta_i)\), where \( y_i = \min(t_i, c_i) \), and \( \delta_i \) is a censoring indicator equal to 1 if \( t_i \leq c_i \), and 0 otherwise, \( i = 1, \ldots, n \). Also, let \( \mathbf{x}_i \) denote the observed covariate vector for the \( i^{th} \) individual. Now, assuming that all \( t_i's \) are i.i.d.
with common density \( f(t) \), the likelihood function for \((\beta, \lambda_0(\cdot))\) can be written as:

\[
L(\beta, \lambda_0(t)|D) \propto \prod_{i=1}^{n} \left[ \lambda_0(y_i) \exp(x_i' \beta) \right]^{\delta_i} \exp \left( - \int_{0}^{y_i} \lambda_0(u) \exp(x'_i \beta) \, du \right), \tag{1.8}
\]

where \( D = (n, y, X, \delta) \), \( y = (y_1, y_2, \ldots, y_n)' \), \( \delta = (\delta_1, \delta_2, \ldots, \delta_n)' \), and \( X = (x_1, \ldots, x_n)' \).

Estimating the parameters of the model via the likelihood function (1.8) requires estimating the baseline hazard function \( \lambda_0(t) \), which is often not of primary interest to the investigator. Cox’s idea of partial likelihood (Cox (1972, 1975)), in which \( \lambda_0(t) \) “is allowed to take on arbitrary values since it does not enter into the estimating equations” (Ibrahim et al. (2001), p. 16), solves this problem. Namely, the partial likelihood for \( \beta \) is given by

\[
PL(\beta|D) = \prod_{j=1}^{d} \frac{\exp(x_{(j)}' \beta)}{\sum_{k \in \mathcal{R}_j} \exp(x_k' \beta)} = \prod_{i=1}^{n} \left( \frac{\exp(x_{(i)}' \beta)}{\sum_{k \in \mathcal{R}_i} \exp(x_k' \beta)} \right)^{\delta_i}, \tag{1.9}
\]

where \( d \) is the number of distinct event times, \( y_{(j)} \) is the \( j^{th} \) ordered survival time, \( j = 1, \ldots, d \), and \( \mathcal{R}_j \) is the risk set at time \( y_{(j)} \), defined as “the set of individuals who are event-free and uncensored at a time just prior to \( y_{(j)} \)” (Ibrahim et al. (2001), p. 16). For simplicity, we assume that there are not ties in the data.

It is worth noting that the product in (1.9) is taken over only those individuals who were observed to fail. Censored observations do not contribute to the numerator, but they do contribute to the denominator. The estimate of the parameter \( \beta \) can be obtained by maximizing (1.9) as a function of \( \beta \) (Ibrahim et al. (2001)).
1.1.3 The Bayesian Approach

Once again, following the notation by Ibrahim et al. (2001) in section 1.5, let $\theta$ be a vector of unknown parameters, and $D$ – the observed data with the corresponding likelihood function $L(\theta|D)$. The Bayesian approach assumes that $\theta$ is a random variable with a certain prior distribution $\pi(\theta)$ (based on our prior knowledge of $\theta$), which combined with the likelihood of the data via Bayes' theorem yields the posterior distribution of $\theta$ given by

$$
\pi(\theta|D) = \frac{L(\theta|D)\pi(\theta)}{\int_{\Theta} L(\theta|D)\pi(\theta) \, d\theta},
$$

where $\Theta$ is the parameter space of $\theta$. All the inferences about the parameter vector $\theta$ are then based on the posterior distribution $\pi(\theta|D)$. Let us note that the denominator of (1.10) does not depend on $\theta$ and enters into the equation merely as a normalizing constant $m(D)$, often called the marginal distribution of the data or the prior predictive distribution. Thus we can write

$$
\pi(\theta|D) \propto L(\theta|D)\pi(\theta).
$$

We see then that the observed data contributes to the posterior through the likelihood function $L(\theta|D)$ and the prior information enters through $\pi(\theta)$. If we have no, or little prior knowledge of $\theta$, it is a common practice to specify a non-informative (or sometimes called flat or vague) prior. Examples of such priors include distributions such as uniform or normal with a large variance. Often, it is useful and easier to specify such or, so called, improper prior (i.e., a prior that does not integrate to 1).
In many applications, however, non informative priors are not sufficient and great care is needed in carrying out the informative prior elicitation. One of the methods for choosing an appropriate informative prior, which will not be discussed here in detail, is, so called, power prior specification, developed for various types of models for survival data by Ibrahim and Chen (1998), Ibrahim et al. (1999), Chen et al. (1999), and Chen et al. (2000).

It is worth mentioning that in most cases $\pi(\theta|D)$ is not available in closed form, due to the fact that $m(D)$ usually does not have a closed form. This raises a question of how to sample from such an "incomplete" posterior distribution. This will be discussed briefly later in Section 1.3.

Since prediction is often one of the main goals of a statistical analysis, let us specify yet one more important quantity. If $f(y_{new}|\theta)$ is a sampling density of $y_{new}$, where $y_{new}$ is a future observation vector, then

$$\pi(y_{new}|D) = \int_{\Theta} f(y_{new}|\theta) \pi(\theta|D) d\theta; \tag{1.12}$$

is the posterior predictive distribution of $y_{new}$ given the data $D$. It is easy to see that (1.12) is just the posterior expectation of $f(y_{new}|\theta)$. Therefore, once samples from $\pi(\theta|D)$ are available, sampling from (1.12) is straightforward.

Having specified the basics of the Bayesian approach, applying the method to survival analysis is straightforward. Our data $D$ would then consist of the survival times for all the individuals under study, the sample size, the matrix of all the covariates
and the censoring indicators, and our parameter vector would include the vector of regression coefficients, $\beta$, in the Cox proportional hazards model, for example.

### 1.1.4 Remarks

We may ask what is the advantage of the Bayesian methods over the traditional frequentist approach in survival analysis. First, as Ibrahim et al. (2001) (p. 26) point out, “survival models ... [can be often] quite hard to fit, especially in the presence of complex censoring schemes”. The availability of various MCMC techniques and advanced statistical software makes the fitting and the implementation of such complex survival models much easier under the Bayesian paradigm. Also, in the frequentist approach we often need asymptotic results, which can be difficult to derive or simply not available, and the question whether the available sample size is large enough for the asymptotic approximations to be valid arises. The Bayesian paradigm bypasses this problem through MCMC techniques, thanks to which “we can make exact inference for any sample size without ... [using] asymptotic calculations”. (Ibrahim et al. (2001), p. 26)

Another advantage is that in the Bayesian approach we can easily incorporate prior information in the form of, for example, historical data, while the frequentist approach does not allow us to do this. We should mention here that “for many models, frequentist inference can be obtained as a special case of Bayesian inference with many types of non-informative priors”. (Ibrahim et al. (2001), p. 27)
Ibrahim et al. (2001) outline a few other areas in which Bayesian methods show advantages over frequentist methods. These include:

- "the availability and flexibility of model building and data analysis tools. For example, ..., model comparisons of nested or non-nested models are easily entertained via Bayes factors or model selection criteria" (Ibrahim et al. (2001), p. 27),

- calculation of some data analysis tools in survival models, such as predictive distributions and residuals (these are generally easier under the Bayesian paradigm),

- missing data (they are considered as parameters in the Bayesian approach, while complicated methods are often required in the frequentist framework),

- variance estimates in missing data problems ("in the Bayesian paradigm they are a by-product of the Gibbs sampler" (Ibrahim et al. (2001), p. 27)).
1.2 Introduction to Dynamic Models

Survival models may include many covariates, some of which can be time-dependent. These covariates are often measured only periodically with a substantial amount of variability which arises from measurement error as well as from true biologic differences. Also, since longitudinal covariates consist of multiple measurements of a marker on a single individual, some of the data might be missing and the issue may need to be addressed in the analysis. Thus, in order to produce unbiased results, the time-dependent covariates themselves need to be modeled in an appropriate fashion. Moreover, in many applications it is assumed that the relationship between covariates and time to failure does not change as time passes. Such assumption in some cases may be unreasonable and a tool that is able to reflect the possible time dependence is necessary. This leads us to consider a class of models called \textit{dynamic models}. The section is based on West and Harrison (1997).

The term \textit{dynamic models} refers to a class of mathematical methods used in mathematical and statistical modeling of time series processes that reflect the dynamic nature of such processes, and recognize the model uncertainty due to the passage of time. The most common "subclass [of dynamic models] is that of normal dynamic linear models, referred to simply as dynamic linear models, or DLMs" (West and Harrison (1997), p. 23), and it will be briefly described in section 1.2.1.

Following the notation by West and Harrison (1997), let $Y_t$, where $t = 1, 2, \ldots$ is the time index, denote the $t^{th}$ value of a series of real-valued quantities, $Y$, observed
over time. Suppose that the time origin is represented by $t_0$ and is set to 0, and the available information at this time is $D_0$. At any time $t$ all inference about the future is based on the current information set, $D_t$. The information updating is sequential in the sense that the available information evolves as time passes by, so that if the value of $Y_t$ is all the relevant information we obtain at time $t$, then $D_t = \{Y_t, D_{t-1}\}$, where $D_{t-1}$ is the information set at time $t-1$. If any additional relevant information can or needs to be incorporated into the model, then $D_t = \{I_t, D_{t-1}\}$, where $I_t$ represents this additional information, and obviously includes $Y_t$. (West and Harrison (1997))

Now, let $p(Y_t|\theta_t, D_{t-1})$ be the distribution that describes the one-step ahead evolution of the series, where $\theta_t$ is the parameter vector at time $t$. While in most cases the dimension of $\theta_t$ does not change over time, sometimes it is necessary either to increase or reduce the number of its elements, according to the current knowledge of the analyst. As always in the Bayesian framework, $\theta_t$ is a random variable, so that its prior distribution at time $t$ is given by $p(\theta_t|D_{t-1})$, which combined with the information from the data observed at time $t$, yields the posterior distribution of the model parameters $p(\theta_t|D_t)$. (West and Harrison (1997))

In the following next few sections let us take a look at some of the fundamental concepts of the dynamic model theory.
1.2.1 The Dynamic Linear Model

One of the simplest univariate dynamic linear models, which can be represented by the quadruple \( \{F_t, \lambda, V_t, W_t\} \), can be defined as (West and Harrison (1997)):

**Observation equation:** \( Y_t = F_t \mu_t + \nu_t, \) \( \nu_t \sim N(0, V_t) \)

**System equation:** \( \mu_t = \lambda \mu_{t-1} + \omega_t, \) \( \omega_t \sim N(0, W_t) \)

**Initial information:** \( (\mu_0|D_0) \sim N(m_0, C_0) \),

where \( \lambda \) and \( F_t \) are known, and \( \nu_t \) and \( \omega_t \) are internally and mutually independent error sequences, and are also independent of \( (\mu_0|D_0) \). \( V_t \) and \( W_t \) may be unknown, and then prior distributions for these parameters need to be specified. A particular example of the model defined above, in which \( F_t \) and \( \lambda \) are both equal to 1, is the first-order polynomial model, represented by \( \{1, 1, V_t, W_t\} \). West and Harrison (1997) in Theorem 2.1 show that for this model the one-step forecast and posterior distributions for any \( t > 0 \) can be obtained sequentially in the following way:

**Posterior for \( \mu_{t-1} \):** \( (\mu_{t-1}|D_{t-1}) \sim N(m_{t-1}, C_{t-1}) \).

**Prior for \( \mu_t \):** \( (\mu_t|D_{t-1}) \sim N(m_{t-1}, R_t) \),

where \( R_t = C_{t-1} + W_t \).

**1-step forecast:** \( (Y_t|D_{t-1}) \sim N(f_t, Q_t) \),

where \( f_t = m_{t-1} \) and \( Q_t = R_t + V_t \).
Posterior for $\mu_t$:  

$$(\mu_t|D_t) \sim N(m_t, C_t),$$

where $m_t = m_{t-1} + A_t e_t$, $C_t = A_t V_t$,

$$A_t = R_t / Q_t, \text{ and } e_t = Y_t - f_t.$$

The proof of this result is by induction and, since it is pretty straightforward, it will not be presented here. For more details, please refer to West and Harrison (1997), section 2.2.2.

West and Harrison (1997) provide a discussion of some of the elements of the distributions presented above. First, they note that $e_t$ is the one-step ahead forecast error, and "$A_t$ is the prior regression coefficient of $\mu_t$ upon $Y_t$, and, in this particular case, is the square of their correlation coefficient" (West and Harrison (1997), p. 37).

Also, they point out that $m_t$ can be alternatively written as $m_t = A_t Y_t + (1-A_t)m_{t-1}$, "showing that $m_t$ is a weighted average of the prior level estimate $m_{t-1}$ and the observation $Y_t$" (West and Harrison (1997), p. 37). The value of $A_t$, called the adaptive coefficient, is between 0 and 1, and is closer to 0 when $R_t < V_t$, i.e., the prior information outweighs the information from the data, and is closer to 1 when the opposite occurs. Also, it is worth noting that the posterior variance $C_t$, is smaller than $R_t$, reflecting "an increase in information about $\mu_t$ due to the additional observation $Y_t$". (West and Harrison (1997), p. 38)

Another level of simplification of the $\{1, 1, V_t, W_t\}$ dynamic linear model is the constant model, in which the observational and evolution variances are constant in time. The details regarding this particular type of the model can be found in West
and Harrison (1997), in section 2.3.

Let us now take a look at somewhat more general concepts of the dynamic linear models. West and Harrison (1997) point out that what enables us to carry out effective dynamic modeling is the conditional independence structure of such models. Namely, they state that "at time \( t \), given \( \theta_t \), the past, present, and future are mutually independent. Also, given just \( D_t \), all the information concerning the future is contained in the posterior \( \cdots \) distribution \( (\theta_t|D_t) \). Further, if this distribution is normal, \( N(m_t, C_t) \), then given \( D_t \), the pair \( \{m_t, C_t\} \) contains all the relevant information about the future, so that in the usual statistical sense, given \( D_t \), \( \{m_t, C_t\} \) is sufficient for \( \{Y_{t+1}, \theta_{t+1}, \ldots, Y_{t+k}, \theta_{t+k}\} \)" (West and Harrison (1997), p. 98).

For each \( t \), the general multivariate \( DLM \) \( \{F_t, G_t, V_t, W_t\} \) is defined as follows (West and Harrison (1997), def. 4.1):

Observation equation: \( Y_t = F_t\theta_t + \nu_t \), \( \nu_t \sim N(0, V_t) \)

System equation: \( \theta_t = G_t\theta_{t-1} + \omega_t \), \( \omega_t \sim N(0, W_t) \)

Initial information: \( (\theta_0|D_0) \sim N(m_0, C_0) \),

where \( Y_t \) is an \( (r \times 1) \) observation vector, \( F_t \) and \( G_t \) are known \( (n \times r) \) and \( (n \times n) \) matrices, respectively, \( V_t \) and \( W_t \) are not always known \( (r \times r) \) and \( (n \times n) \) variance matrices, respectively, \( \theta_t \) is a \( (n \times 1) \) parameter vector, and \( m_0 \) and \( C_0 \) are some prior moments. Also, as before, the observational and evolution error sequences are assumed to be internally and mutually independent, and independent of \( (\theta_0|D_0) \).

Consider for now a special case of the model with \( r = 1 \), i.e., the univariate \( DLM \).
The updating equations for such a model, derived by West and Harrison (1997) in Theorem 4.1, are:

**Posterior at** $t - 1$:  

$$(\theta_{t-1} | D_{t-1}) \sim N(m_{t-1}, C_{t-1}).$$

**Prior at** $t$:  

$$(\theta_t | D_{t-1}) \sim N(a_t, R_t),$$

where $a_t = G_t m_t$ and $R_t = G_t C_{t-1} G_t' + W_t$.

**1-step forecast:**  

$$(Y_t | D_{t-1}) \sim N(f_t, Q_t),$$

where $f_t = F_t' a_t$ and $Q_t = F_t' R_t F_t + V_t$.

**Posterior at** $t$:  

$$(\theta_t | D_t) \sim N(m_t, C_t),$$

where $m_t = a_t + A_t e_t$, $C_t = R_t - A_t Q_t A_t'$,

$$A_t = R_t F_t / Q_t^{-1}, \text{ and } e_t = Y_t - f_t.$$

One of the primary interests of the analysis lies in finding the full forecast distributions (which are normal), defining components of which are given by the following results provided by West and Harrison (1997) in Theorem 4.2:

For $0 \leq j < k$, at each time $t$:

(a) State distribution:  

$$(\theta_{t+k} | D_t) \sim N(\alpha_t(k), R_t(k)).$$

(b) Forecast distribution:  

$$(Y_{t+k} | D_t) \sim N(f_t(k), Q_t(k)).$$

(c) State covariances:  

$$C(\theta_{t+k}, \theta_{t+j} | D_t) = C_t(k, j),$$

(d) Observation covariances:  

$$C(Y_{t+k}, Y_{t+j} | D_t) = F_t' C_t(k, j) F_t + j,$$
(e) Other covariances: 
\[ C(\theta_{t+k}, Y_{t+j} | D_t) = C_t(k, j) F_{t+j}, \]
\[ C(Y_{t+k}, \theta_{t+j} | D_t) = F_{t+k}' C_t(k, j), \]
where
\[ f_t(k) = F_{t+k}' a_t(k), \quad Q_t(k) = F_{t+k}' R_t(k) F_{t+k} + V_{t+k}, \]
\[ a_t(k) = G_{t+k} a_t(k - 1), \]
\[ R_t(k) = G_{t+k} R_t(k - 1) G_{t+k}' + W_{t+k}, \]
\[ C_t(k, j) = G_{t+k} C_t(k - 1, j), \quad k = j + 1, \ldots, \]
\[ a_t(0) = m_t, \quad R_t(0) = C_t, \quad C_t(j, j) = R_t(j). \]

Let us now consider the case in which the observational variance \( V_t \) is unknown, but constant over time (i.e., \( V_t = V \) for all \( t \)), and in which all the remaining model variances are scaled by \( V \). Also, let \( \phi = V^{-1} \) be the unknown observation precision. West and Harrison (1997) (def. 4.5) define the DLM as follows:

**Observation equation:** 
\[ Y_t = F_t' \theta_t + \nu_t, \quad \nu_t \sim N(0, V) \]

**System equation:** 
\[ \theta_t = G_t \theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, V W_t^*) \]

**Initial information:** 
\[ (\theta_0 | D_0, \phi) \sim N(m_0, V C_0^*), \]
\[ (\phi | D_0) \sim G(\frac{n_0}{2}, \frac{n_0 \phi_0}{2}), \]

where \( m_0, C_0^*, n_0, S_0, \) and \( W_t^* \) are known, and \( G(a, b) \) denotes the gamma distribution with parameters \( a \) and \( b \).
The following distributional results use the multivariate $T$ distributions\footnote{We say that the $(d \times 1)$ vector $\theta$ has a multivariate $T$ distribution with $\nu$ degrees of freedom, mode $\mu$, and positive definite scale matrix $\Sigma$, if its density is \[ p(\theta) \propto \{1 + \frac{1}{\nu} (\theta - \mu)'\Sigma^{-1}(\theta - \mu)\}^{-\nu/2}, \] writing $\theta \sim T_\nu(\mu, \Sigma)$. Also, $E[\theta] = \mu$ for $\nu > 1$ and $Var[\theta] = \Sigma \frac{\nu}{\nu - 2}$ if $\nu > 2$ (Gelman et al. (1995)). As $h \to \infty$ the distribution converges to the normal $\theta \sim N(\mu, \Sigma)$ (West and Harrison (1997)).} for the state vector at all times (West and Harrison (1997), Theorem 4.3):

(a) Conditional on $V$:

\[ (\theta_{t-1}|D_{t-1}, V) \sim N(m_{t-1}, VC_{t-1}^*), \]
\[ (\theta_t|D_{t-1}, V) \sim N(a_t, VR_t^*), \]
\[ (Y_t|D_{t-1}, V) \sim N(f_t, VQ_t^*), \]
\[ (\theta_t|D_t, V) \sim N(m_t, VC_t^*), \]

where $a_t = G_t m_{t-1}, \quad R_t^* = G_t C_{t-1}^* G_t' + W_t^*,$

\[ f_t = F_t' a_t, \quad Q_t^* = 1 + F_t' R_t^* F_t, \]
\[ e_t = Y_t - f_t, \quad A_t = R_t^* F_t / Q_t^*, \]
\[ m_t = a_t + A_t e_t, \quad C_t^* = R_t^* - A_t A_t' Q_t^*. \]

(b) For the precision $\phi = V^{-1}$:

\[ (\phi|D_{t-1}) \sim G \left( \frac{n_{t-1}}{2}, \frac{n_{t-1} S_{t-1}}{2} \right), \]
\[ (\phi|D_t) \sim G \left( \frac{n_t}{2}, \frac{n_t S_t}{2} \right), \]

where $n_t = n_{t-1} + 1, \quad n_t S_t = n_{t-1} S_{t-1} + e_t^2 / Q_t^*.$
(c) Unconditional on V:

\[(\theta_{t-1}|D_{t-1}) \sim T_{n_{t-1}}(m_{t-1}, C_{t-1}),\]

\[(\theta_t|D_{t-1}) \sim T_{n_{t-1}}(a_t, R_t),\]

\[(Y_t|D_{t-1}) \sim T_{n_{t-1}}(f_t, Q_t),\]

\[(\theta_t|D_t) \sim T_{n_t}(m_t, C_t),\]

where \(R_t = S_{t-1}R_t^*, \ Q_t = S_{t-1}Q_t^*, \ C_t = S_tC_t^*\).

(d) Operational definition of updating equations:

With \(Q_t = F_t' R_t F_t + S_{t-1}\) and \(A_t = R_t F_t / Q_t\),

\[n_t = n_{t-1} + 1, \quad S_t = S_{t-1} + \frac{S_{t-1}}{n_t} \left( \frac{S_{t-1}}{Q_t} - 1 \right),\]

\[m_t = a_t + A_t e_t, \quad C_t = \frac{S_{t-1}}{n_{t-1}} (R_t - A_t A_t' Q_t).\]

The final concept worth mentioning in this section is that of discount factors as an aid to specifying \(W_t\). Let us illustrate this concept based on the univariate constant model. Following the notation by West and Harrison (1997), let \(A\) and \(C\) be the limiting values of \(A_t\) and \(C_t\), respectively. It can be shown that \(W = AC/(1 - A)\), so that \(W\) is a fixed proportion of \(C\). So, "between observations, the addition of the error \(\omega_t\) leads to an additive increase of \(W = 100A/(1 - A)\)% of the initial uncertainty \(C\" (West and Harrison (1997), p. 51). If we denote \(1 - A\) by \(\delta\), then \(R = C/\delta\). Hence, choice of \(\delta\) guides the choice of \(W\). West and Harrison (1997) (p.51) state that "it is [often] convenient and natural to adopt a constant rate of increase of uncertainty ... for all \(t\) rather than just in the limiting case. Thus, for a given
discount factor \( \delta \), typically between 0.8 and 1, choosing \( W_t = C_{t-1}(1-\delta)/\delta \) for each \( t \) implies \( R_t = C_{t-1}/\delta \). This DLM is not a constant model, but quickly converges to the constant DLM \( \{1, 1, V, r V\} \), with \( r = (1-\delta)^2/\delta^v \).

### 1.2.2 Linear Growth Models

Let us first define the second-order polynomial model. West and Harrison (1997) (sec. 7.2.1) define such a model as any DLM for which at any time \( t \), the forecast function \( f_t(k) \) is of the form \( a_{t0} + a_{t1} k \). The forecast function \( f(\cdot) \) is given by

\[
f_t(k) = E[Y_{t+k}|D_t],
\]

and for this particular model can be also written as

\[
f_t(k) = f_t(0) + [f_t(1) - f_t(0)]k,
\]

thus expressing \( a_{t0} \) and \( a_{t1} \) as \( f_t(0) \) and \( f_t(1) - f_t(0) \), respectively.

Writing

\[
\theta_t = \begin{pmatrix} \theta_{t1} \\ \theta_{t2} \end{pmatrix} = \begin{pmatrix} \mu_t \\ \beta_t \end{pmatrix},
\]

a second-order polynomial DLM with a known observational variance \( V_t \) can be described by the following equations (West and Harrison (1997)):

Observation: \( Y_t = \mu_t + \nu_t \),

System: \( \mu_t = \mu_{t-1} + \beta_t + \omega_{t1} \),

\( \beta_t = \beta_{t-1} + \omega_{t2} \),
\[(\theta_{t-1}|D_{t-1}) \sim N(m_{t-1}, C_{t-1}),\]

with

\[
\omega_t = (\omega_{t1}, \omega_{t2})' \sim N(0, W_t), \quad \nu_t \sim N(0, V_t),
\]

\[
m_{t-1} = \begin{pmatrix} m_{t-1} \\ b_{t-1} \end{pmatrix} \quad \text{and} \quad C_{t-1} = \begin{pmatrix} C_{t-1,1} & C_{t-1,3} \\ C_{t-1,3} & C_{t-1,2} \end{pmatrix}.
\]

Thus, \(f_t(k) = m_t + kb_t\). If \(V_t\) is unknown, the normal posterior distributions are simply replaced by Student T distributions.

Now, with \(W_t\) written as

\[
W_t = \begin{pmatrix} W_{t1} & W_{t3} \\ W_{t3} & W_{t2} \end{pmatrix},
\]

the updating equations for the model are (West and Harrison (1997), sec. 7.2.2):

- \((\theta_t|D_{t-1}) \sim N(\alpha_t, R_t)\), where

\[
\alpha_t = \begin{pmatrix} m_{t-1} + b_{t-1} \\ b_{t-1} \end{pmatrix}, \quad R_t = \begin{pmatrix} R_{t1} & R_{t3} \\ R_{t3} & R_{t2} \end{pmatrix},
\]

\[
R_{t1} = C_{t-1,1} + 2C_{t-1,3} + C_{t-1,2} + W_{t1},
\]

\[
R_{t2} = C_{t-1,2} + W_{t2},
\]

\[
R_{t3} = C_{t-1,3} + W_{t3},
\]

- 1-step forecast distribution:

\[(Y_t|D_{t-1}) \sim N(f_t, Q_t),\]

where \(f_t = f_{t-1}(1) = m_{t-1} + b_{t-1}\) and \(Q_t = R_{t1} - V_t\).
• with \( e_t = Y_t - f_t \), and
\[
A_t = \begin{pmatrix} A_{t1} \\ A_{t2} \end{pmatrix} = \begin{pmatrix} R_{t1}/Q_t \\ R_{t2}/Q_t \end{pmatrix},
\]
the posterior distribution of \( \theta_t \) at \( t \) is given by:
\[
(\theta_t | D_t) \sim N \left( \begin{pmatrix} m_t \\ b_t \end{pmatrix}, \begin{pmatrix} C_{t1} & C_{t3} \\ C_{t3} & C_{t2} \end{pmatrix} \right),
\]
where
\[
m_t = m_{t-1} + b_{t-1} + A_{t1}e_t, \quad b_t = b_{t-1} + A_{t2}e_t,
\]
\[
C_{t1} = A_{t1}V_t, \quad C_{t2} = R_{t2} - A_{t2}R_{t3}, \quad C_{t3} = A_{t2}V_t.
\]

Now, we come to the definition of the linear growth model. A linear growth model, as defined by West and Harrison (1997) (def. 7.4), is any second-order polynomial model that can be represented by
\[
\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, V_t, \begin{pmatrix} W_{t1} + W_{t2} & W_{t2} \\ W_{t2} & W_{t2} \end{pmatrix} \right\}.
\]
If we write
\[
\omega_t = \begin{pmatrix} \omega_{t1} \\ \omega_{t2} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} W_{t1} & 0 \\ 0 & W_{t2} \end{pmatrix} \right),
\]
then the form of a linear growth model in terms of the observational and system equations is (West and Harrison (1997)):

Observation: \( Y_t = \mu_t + \nu_t \),

System: \( \mu_t = \mu_{t-1} + \beta_t + \omega_{t1} \),
\( \beta_t = \beta_{t-1} + \omega_{t2} \).
1.2.3 Dynamic Hierarchical Models

In medical studies involving one or more longitudinal responses, the data consist of a number of series of measurements taken on a group of individuals over a certain time period. As Gamerman and Migon (1993) point out, these individual series can be analyzed separately, but since there might be relations between the patients, a single model that can analyze them jointly may be advantageous. This way all the series are part of one multivariate system, and hierarchical models theory (Lindley and Smith (1972)) combined with dynamic linear modeling (Harrison and Stevens (1976)) can be applied to analyze it.

The dynamic hierarchical model proposed by Gamerman and Migon (1993) consists of three parts:

- the observation equation:

\[ Y_t = F_{1t} \theta_{1t} + \nu_{1t}, \quad \nu_{1t} \sim N(0, V_{1t}) \quad (1.13) \]

- the structural equations:

\[ \theta_{1t} = F_{2t} \theta_{2t} + \nu_{2t}, \quad \nu_{2t} \sim N(0, V_{2t}) \quad (1.14) \]

\[ \theta_{2t} = F_{3t} \theta_{3t} + \nu_{3t}, \quad \nu_{3t} \sim N(0, V_{3t}) \]

- the system equation:

\[ \theta_{3t} = G_t \theta_{3,t-1} + w_t, \quad w_t \sim N(0, W_t) \quad (1.15) \]
with prior \( (\theta_{30}|D_0) \sim N(m_{30}, C_{30}) \). \( \nu_{1t}, \nu_{2t}, \nu_{3t}, \) and \( w_t \) are independent, and all the covariance matrices \( V_{1t}, V_{2t}, V_{3t}, W_t \), as well as \( F_{1t}, F_{2t}, \) and \( F_{3t} \) are assumed known. The first and third equation play the same role as in the regular dynamic linear model introduced in previous sections, whereas the structural equations describe the hierarchical structure of the parameters. Although models with more than three levels of hierarchy may be considered, they are not common in statistical applications (Lindley and Smith (1972)). A simpler, two-stage dynamic hierarchical model, is used more frequently, and can be obtained from the model above by setting \( F_{3t} = I \), and \( V_{3t} = 0 \). (Gamerman and Migon (1993))

As noted by Gamerman and Migon (1993) (p. 630), “the key aspect of [the structural] equations ... is the progressive reduction in the dimension of the parameters as the level becomes higher. So, [if] the dimensions of \( \theta_{it} \) are \( r_i, i = 1, 2, 3, \ldots \) [then] \( r_1 > r_2 > r_3 \)”.

Let us consider the following two-stage variation of the presented model, which can be useful in relation to this research (Gamerman and Migon (1993), ex. 1):

\[
\text{Observation: } (Y_{it}|\beta_{it}) \sim N(\beta_{it}, \sigma^2), \quad i = 1, \ldots, n \\
\text{Structure: } (\beta_{it}|\mu_t) \sim N(\mu_t, \tau^2) \\
\text{System: } \mu_t = \mu_{t-1} + w_t, \quad w_t \sim N(0, W_t)
\]

Thus, in the context of clinical trials involving longitudinal data collected on a group of individuals, \( \beta_{it} \)'s can be thought of as true, patient specific trajectories of a longitudinal marker, with common population mean \( \mu_t \).
Gamerman and Migon (1993) (p. 632) note that equations (1.13) (1.15) “do not fully specify the model because they do not make clear the conditioning events”. To take into account the information collected up to time \( t \), Gamerman and Migon (1993) define the \( k \)-stage dynamic hierarchical model as

\[
(Y_t | \theta_{t}) \sim N(F_{t} \theta_{t}, V_{t})
\]

\[
(\theta_{it} | \theta_{i+1,t}, D_{t-1}) \sim N(F_{i+1,t} \theta_{i+1,t}, V_{i+1,t}), \quad i = 1, \ldots, k - 1
\]

\[
(\theta_{kt} | \theta_{k,t-1}, D_{t-1}) \sim N(G_{t} \theta_{k,t-1}, W_{t})
\]

\[
(\theta_{k,0} | D_{0}) \sim N(m_{k,0}, C_{k,0})
\]

where \( F_{it} \) and \( V_{it} \) are assumed known. With the model specified in such a way, the following are true (Gamerman and Migon (1993), Theorem 1):

- prior distribution at \( t \):

\[
(\theta_{it} | D_{t-1}) \sim N(a_{it}, R_{it}), \quad i = 1, \ldots, k
\]

with \( a_{it} = F_{i+1,t} a_{i+1,t}, \quad R_{it} = F_{i+1,t} R_{i+1,t} F_{i+1,t} + V_{i+1,t}, \quad i = 1, \ldots, k - 1, \quad a_{kt} = G_{t} m_{k,t-1}, \quad R_{kt} = G_{t} C_{k,t-1} G_{t}' + W_{t};

- one-step ahead forecast distribution:

\[
(Y_t | D_{t-1}) \sim N(f_t, Q_t)
\]

with \( f_t = F_{t} a_{1t}, \quad Q_t = F_{t} R_{t} F_{t} - V_{1t};
\]
• posterior distribution at $t$:

$$(\theta_{it}|D_t) \sim N(m_{it}, C_{it}), \quad i = 1, \ldots, k$$

with $m_{it} = a_{it} + S_{it} Q^{-1}_x (Y_t - f_t)$, $C_{it} = R_{it} - S_{it} Q^{-1}_x S'_{it}$, $S_{it} = R_{it} E_{0it}$.

$E_{0it} = F_{1,t} \ldots F_{it}$.

These results can be easily extended to multi-step ahead forecast. For details, please refer to Gamerman and Migon (1993), section 3.

A natural extension of the model is to relax the assumption of all the variance matrices being known, which is very rare in applications. Thus, appropriate changes have to be made to allow for variance estimation. Unfortunately, analysis with completely unknown variances rapidly becomes quite complex, and in order to find estimates of the model parameters some numerical techniques, such as MCMC methods, are necessary (Ferreira et al. (1997); Landim and Gamerman (2000); Salvador et al. (2004)).

Gamerman and Migon (1993) note, however, that if the structure of the model allows us to assume that $V_{1t} = \sigma^2 I_n$ (where $n = \text{dim}(Y_t)$), then a conjugate analysis is still possible after all the model variances are scaled by unknown factor $\sigma^2$, with the inverse gamma prior distribution.

More details about dynamic hierarchical models, such as smoothing, relationship between stages of hierarchy, or parameter estimation in a model with fully unknown variance matrices, can be found in Gamerman and Migon (1993); Ferreira et al. (1997); Landim and Gamerman (2000); Salvador et al. (2004).
1.3 A Quick Overview of Simulation Methods

In an analysis that uses the Bayesian approach, all the inference concerning the parameters of interest is based on the posterior distribution of the parameters. However, it is quite common for the posterior not to have a closed form. Thus, the question of how to sample from a distribution that is not available in closed form arises. Markov Chain Monte Carlo (MCMC) techniques provide the answer to this problem. In the following sections some basic MCMC techniques will be presented, with a specific application to the dynamic models setting.

1.3.1 Gibbs Sampling

One of the most popular MCMC schemes is called Gibbs sampling (Geman and Geman (1984); Gelfand and Smith (1990)). It is based on subsequent generations from the full conditional distributions, and particularly useful when direct sampling from the posterior distribution of the parameters is complicated or unavailable, while generations from the full conditional distributions are possible or simply easier.

Adopting the notation by Gamerman (1997), let us assume that the distribution of interest is $\pi(\theta)$, where $\theta = (\theta_1, \ldots, \theta_d)'$, keeping in mind that each $\theta_i$ can be multidimensional. Moreover, let us suppose that the full conditional distributions, denoted by $\pi_i(\theta_i) = \pi(\theta_i|\theta_{-i})$, $i = 1, \ldots, d$, are available and can be sampled from. Gamerman (1997) describes the algorithm of Gibbs sampling in the following way:
Step 1 Set \( j = 1 \) and choose initial value for \( \theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})' \).

Step 2 Draw a new value of \( \theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_d^{(j)})' \) through successive generation of values

\[
\theta_1^{(j)} \sim \pi(\theta_1 | \theta_2^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_d^{(j-1)}) \\
\theta_2^{(j)} \sim \pi(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_d^{(j-1)}) \\
\vdots \\
\theta_d^{(j)} \sim \pi(\theta_d | \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{d-1}^{(j)})
\]

Step 3 Set \( j = j + 1 \) and go back to Step 2.

Once the convergence is reached, \( \theta^{(j)} \) is a draw from the distribution of interest \( \pi(\theta) \).

(Gamerman (1997))

1.3.2 Metropolis-Hastings Algorithm

The Metropolis-Hastings MCMC sampling method (Metropolis et al. (1953); Hastings (1970)) is useful whenever the direct sampling from neither the posterior, \( \pi(\theta) \), nor, unlike in Gibbs sampling, the full conditional distributions, \( \pi_i(\theta_i), i = 1, \ldots, d \), is feasible. Following Gamerman (1997), the sampling is carried out by the following algorithm:

Step 1 Set \( j = 1 \) and choose an initial value \( \theta^{(0)} \).

Step 2 Draw a new value \( \phi \) from the density \( q(\theta^{(j-1)}, \cdot) \), where \( q(\cdot, \cdot) \) is an arbitrary proposal density.
Step 3 Calculate the acceptance probability $\alpha(\theta^{(j-1)}, \phi)$ given by

$$\alpha(\theta, \phi) = \min \left\{ 1, \frac{\pi(\phi) q(\phi, \theta)}{\pi(\theta) q(\theta, \phi)} \right\}$$

Step 4 Generate $u \sim Unif(0, 1)$. If $u \leq \alpha(\theta^{(j-1)}, \phi)$ set $\theta^{(j)} = \phi$. Otherwise, $\theta^{(j)} = \theta^{(j-1)}$.

Step 5 Set $j = j + 1$ and go back to Step 2 until convergence is reached.

Let us note that if $q$ is a symmetric density (such as normal, for example) then $\alpha(\theta, \phi) = \min \left\{ 1, \frac{\pi(\phi)}{\pi(\theta)} \right\}$. Special cases of the method include symmetric chains, random walk chains, independence chains, and others.

1.3.3 Forward Filtering Backward Sampling Algorithm

Retaining the notation of section 1.2.1, the forward filtering backward sampling (FFBS) algorithm, introduced by Frühwirth-Schantzter (1994), is a simulation method, that allows to sample from the conditional distribution $p(\Theta_n | D_n)$, where $\Theta_n = \{ \theta_0, \theta_1, \ldots, \theta_n \}$.

Noting that the above distribution can be written as

$$p(\Theta_n | D_n) = p(\theta_n | D_n)p(\theta_{n-1} | \theta_n, D_{n-1}) \ldots p(\theta_1 | \theta_2, D_1)p(\theta_0 | \theta_1, D_0),$$

the algorithm can be described as follows (West and Harrison (1997), sec. 15.2):

Step 1 Run a standard DLM analysis forward in time (given the data $Y_t$), saving at each time $t = 1, \ldots, n$ the quantities $m_t, C_t, a_t$ and $R_t$.

Step 2 Generate a value of $\theta_n$ from $(\theta_n | D_n) \sim N(m_n, C_n)$. 
Step 3 For \( t = n - 1, n - 2, \ldots, 1, 0 \), draw \( \theta_t \) from \( (\theta_t | \theta_{t+1}, D_t) \sim N(h_t, H_t) \), using
the value of \( \theta_{t+1} \) obtained in Step 2 and with

\[
h_t = m_t + B_t (\theta_{t+1} - a_{t+1}),
\]

\[
H_t = C_t - B_t R_{t+1} B_t',
\]

\[
B_t = C_t G_{t+1} R_{t+1}^{-1}.
\]

Steps 2 and 3 can then be repeated a number of times depending on the desired size of the sample from \( p(\Theta_n | D_n) \) that one wishes to obtain.

1.3.4 Auxiliary Particle Filters

In the dynamic models for survival data context, the time axis is usually divided into time intervals according to the chosen grid \( \tau = \{t_0, t_1, \ldots, t_N\} \), where usually \( t_0 = 0 \), and \( t_N = \infty \). The analysis then proceeds sequentially from one interval to another. Initial prior distribution at \( t_0 \) for the parameters which are allowed to change over time, is usually a well known distribution, such as normal, for example. This prior multiplied by the likelihood of the data for the initial time interval \([t_0, t_1]\), yields the posterior distribution for the parameters at time \( t_1 \), which is often very complex and thus not available in closed form. A pseudo-random sample from this distribution can be fairly easily obtained by means of various MCMC techniques. This posterior, however, is later used as the prior for the parameters in the next time interval \([t_1, t_2]\), causing the prior to be available only in a form of discrete
points, "particles" (Kitagawa (1996)), and associated with them probabilities. Thus, in order to be able to proceed with our analysis and eventually obtain the estimated mean trajectories of the parameters, we need a simulation tool called auxiliary particle filters which allows us to use these "particles" and their probabilities as a discrete approximation of a continuous random variable. The method was developed by Pitt and Shephard (1999) and its details are presented below.

Suppose we have a time series $Y_t$, $t = 1, \ldots, n$, and the parameter vector $\theta_t$. Also, let $p(Y_t|\theta_t)$ be the "measurement" density of $Y_t$, $p(\theta_{t+1}|\theta_t)$ the "evolution" density of $\theta_t$, and $p(\theta_0)$ the initializing density of the parameter vector at time 0. The primary objective is to use simulation to estimate $p(\theta_t|D_t)$ — the density which summarizes all the information about $\theta_t$ at time $t$ given $D_t = \{Y_1, \ldots, Y_t\}$, but which is often difficult to compute. Particle filters approximate the above density by discrete points, "particles", $\theta^1_t, \ldots, \theta^M_t$ (viewed as samples from $p(\theta_t|D_t)$), with their corresponding probabilities $\pi^1_t, \ldots, \pi^M_t$, where $M$ is taken to be large. (Pitt and Shephard (1999))

Let us suppose that $p(\theta_{t+1}|D_t)$ is the "prior" density of $\theta_{t+1}$ at time $t + 1$ (prior to observing $Y_{t+1}$). This density is obtained by propagating the "posterior" density $p(\theta_t|D_t)$ into the future via the "evolution" density $p(\theta_{t+1}|\theta_t)$. This can be written as

$$p(\theta_{t+1}|D_t) = \int p(\theta_{t+1}|\theta_t) \, p(\theta_t|D_t) \, d\theta_t. \quad (1.17)$$

Combining the "prior" (1.17) with the "measurement" density of $Y_{t+1}$ via the Bayes
Theorem yields the “posterior” density of $\theta_{t+1}$ at time $t+1$:

$$p(\theta_{t+1}|D_{t+1}) = \frac{p(Y_{t+1}|\theta_{t+1})p(\theta_{t+1}|D_t)}{p(Y_{t+1}|D_t)},$$

(1.18)

where

$$p(Y_{t+1}|D_t) = \int p(Y_{t+1}|\theta_{t+1})p(\theta_{t+1}|D_t) \, d\theta_{t+1}.$$

The key feature of particle filters is that they allow to approximate the “prior” (1.17) by using the discrete support of the particles to produce the “empirical prior” density

$$\hat{p}(\theta_{t+1}|D_t) = \sum_{j=1}^{M} p(\theta_{t+1}|\theta^j_t) \pi^j_t.$$

(1.19)

Thus, the “empirical posterior” density (which is the approximation to the true “posterior” density (1.18)) can be obtained by combining (1.19) with the “measurement” density of $Y_{t+1}$:

$$\hat{p}(\theta_{t+1}|D_{t+1}) \propto p(Y_{t+1}|\theta_{t+1})\hat{p}(\theta_{t+1}|D_t).$$

(1.20)

This density is then sampled to produce new particles $\theta^1_{t+1}, \ldots, \theta^M_{t+1}$ with probabilities $\pi^1_{t+1}, \ldots, \pi^M_{t+1}$. The whole procedure can then be iterated through the data. (Pitt and Shephard (1999))

Pitt and Shephard (1999) note that sampling from (1.19) can be carried out by first choosing $\theta^j_t$ with probability $\pi^j_t$ and then generating a value from $p(\theta_{t+1}|\theta^j_t)$. Furthermore, if $p(Y_{t+1}|\theta_{t+1})$ can also be evaluated then drawing from the “empirical posterior” density (1.20) may be carried out by one of the following three sampling methods: sampling/importance resampling (SIR) (Rubin (1987); Berzuini et al. (1997); Gordon et al. (1993); Isard and Blake (1996); Kitagawa (1996)), acceptance
sampling (Hürzeler and Künsch (1995)), and MCMC (Gilks et al. (1996)). Pitt and Shephard (1999) point out, however, that due to the mixture structure of the particle filters, “it is difficult to adapt the ... sampling methods without greatly slowing down the running of the filter” (p. 592). The authors argue that “many of these problems are reduced when we perform particle filtering in a higher dimension” (p. 592), introducing an auxiliary variable \( k \), corresponding to an index on the mixture in (1.19).

Thus, the task is to sample from the joint density \( p(\theta_{t+1}, k|D_{t+1}) \) defined as

\[
p(\theta_{t+1}, k|D_{t+1}) \propto p(Y_{t+1} | \theta_{t+1}) p(\theta_{t+1}|\theta_t^k) \pi_t^k, \quad k = 1, \ldots, M, \tag{1.21}
\]

and later discard the index \( k \), producing a sample from the density (1.20).

Sampling from (1.21) can be done using one of the sampling methods mentioned above. In particular, Pitt and Shephard (1999) suggest the following algorithm to sample from (1.21) using the SIR method:

**Step 1** Generate \( R \) (\( R \gg M, R \to \infty \)) proposals \((\theta_{t+1}^j, k^j)\) from the proposal density

\[
g(\theta_{t+1}, k|D_{t+1}) \propto p(Y_{t+1} | \mu_{t+1}^k) p(\theta_{t+1}|\theta_t^k) \pi_t^k, \quad k = 1, \ldots, M,
\]

where \( \mu_{t+1}^k \) is any value (e.g., mean) associated with \( p(\theta_{t+1}|\theta_t^k) \). This can be done as follows:

- simulate \( R \) indices \( k^j, j = 1, \ldots, R \), according to the “first-stage weights”

\[
\lambda_k \propto g(k|^D_{t+1}), \quad k = 1, \ldots, M, \quad \text{where} \quad g(k|D_{t+1}) \propto \pi_t^k p(Y_{t+1} | \mu_{t+1}^k).
\]
for each of the simulated indices \( k^j, j = 1, \ldots, R \), generate a value of \( \theta_{t+1} \) from the “evolution” density \( p(\theta_{t+1} | \theta_t^{k^j}) \).

**Step 2** Construct resampling weights (or, the “second-stage weights”):

\[
w_j = \frac{p(Y_{t+1} | \theta_t^{k^j})}{p(Y_{t+1} | \mu^{k^j}_{t+1})}, \quad \pi_j = \frac{w_j}{\sum_{i=1}^R w_i}, \quad j = 1, \ldots, R.
\]

**Step 3** Convert the obtained non-random \( R \)-sample to a random \( M \)-sample, by resampling the \( \theta_{t+1}^1, \ldots, \theta_{t+1}^R \) using the just computed weights \( \pi_{t+1}^1, \ldots, \pi_{t+1}^R \) (this is why \( R \to \infty \) and \( R \gg M \) is required).

### 1.3.5 Remarks

In the MCMC techniques presented in this section, a value from the distribution of interest \( \pi \) is only obtained when the number of iterations of the chain approaches infinity. Obviously, this is not attainable in practice, and to address the problem, convergence characteristics of the chain have been studied. We may thus pose a question: How many iterations do we need to assure convergence? There is no simple answer to this. There exist, however, various convergence diagnostic procedures that allow us to determine if the generated sample is sufficiently close to a “true” sample from \( \pi \). These include many graphical techniques as well as methods developed by Geweke, Gelman and Rubin, Raftery and Lewis, Heidelberger and Welch, among others.
It is certainly possible to nest the two methods presented above within each other. For example, if the calculation of the full conditional distributions is feasible, but one or more of them is not easy to sample from, the Metropolis-Hastings algorithm can be applied to generate a sample from that distribution.

1.4 Previous Research on Modeling of Longitudinal and Time to Event Data

This section reviews existing approaches for modeling longitudinal measurements, often measured with error or incompletely observed, and survival data, possibly censored. Such models have recently become more popular in clinical trials, due to the possibility of longitudinal biologic markers, such as CD4 count in AIDS clinical trials, to be important predictors of survival. Although the emphasis is put on the models using the Bayesian approach, the methods based on the frequentist paradigm give useful insight into the nature of the problem, and thus, some of them will be presented here as well.

1.4.1 Two-stage Method

Let us first take a look at the approach taken by Tsiatis et al. (1995). This, so called "two-stage" method, applied to CD4 counts and survival in a clinical trial with AIDS patients, is based on the random effects growth-curve and the Cox proportional hazard
models. In the first stage, empirical Bayes estimates for patient specific random effects and their variances are obtained. These estimates are then used as a time-dependent covariate to find the parameters in the Cox model that maximize the partial likelihood. More specifically, the effect of the measurement error is described by

$$Y(t) = Y^*(t) + e(t),$$

where $Y(t)$ is the observed value of the longitudinal covariate (CD4 count) at time $t$, $Y^*(t)$ is its true unobserved value at time $t$, and $e(t)$ is the measurement error, such that $E[e(t)] = 0$, $Var[e(t)] = \sigma^2$, $Cov(e(s), e(t)) = 0$, $s \neq t$, and independent of $Y^*(t)$. The true covariates process is modeled by

$$Y^*_i(t) = \alpha_i + \beta_i t,$$

where

$$\begin{pmatrix}
\alpha_i \\
\beta_i
\end{pmatrix} \sim N
\begin{pmatrix}
\begin{pmatrix}
\mu_{\alpha}^{(t)} \\
\mu_{\beta}^{(t)}
\end{pmatrix},
\begin{pmatrix}
\sigma_{\alpha}^2(t) & \sigma_{\alpha\beta}^{(t)} \\
\sigma_{\alpha\beta}^{(t)} & \sigma_{\beta}^2(t)
\end{pmatrix}
\end{pmatrix}.$$ 

It is unclear why the parameters of the distribution above depend on time but the vector $(\alpha_i, \beta_i)'$ itself does not.

For the survival model it is assumed that censoring and measurement error, and the timing of the visits prior to $t$ are non-informative. Although an analysis in which the last of the assumptions is relaxed was considered, it did not find substantial differences in the hazard rate estimates. From these assumptions and by the law of conditional probability the observed hazard rate $\lambda(t|\bar{Y}(t))$ is written as

$$\lambda(t|\bar{Y}(t)) = \lambda_0(t)E[f(\bar{Y}^*(t), \gamma)|Y(t_1), \ldots, Y(t_j), X \geq t],$$

(1.22)
where $f(\hat{Y}^*(t), \gamma)$ is some function of the history $\hat{Y}^*(t)$ and the parameter $\gamma$, and $X = \min(T, C)$ is the observed right-censored data ($T$ - survival time, $C$ - censoring time). Let us denote the expectation term in the hazard (1.22) by $E[t, \gamma]$. If $E[t, \gamma]$ were known, $\gamma$ can be estimated by maximizing the partial likelihood:

$$
\prod_{i=1}^{n} \left( \frac{E[X_i, \gamma]}{\sum_{j=1}^{n} E[X_i, \gamma] R_j(X_i)} \right)^{\delta_i}
$$

(1.23)

where $\delta_i$ is the censoring indicator, and $R_j(v)$ is the indicator of being at risk at time $v$, $I(X_j \geq v)$. However, since $E[t, \gamma]$ is difficult to obtain, various simplifying assumptions are introduced. First, the hazard is considered as a function of $Y^*(t)$ only, i.e., it only depends on the current value of the longitudinal covariate rather than on the entire history $\hat{Y}^*(t)$. Second, the Cox proportional hazard model is used, that is $f(Y^*(t), \gamma) = \exp(\gamma Y^*(t))$. Finally, normality for both the true longitudinal measurements $Y^*(t)$ and the measurement errors $e(t)$ is assumed.

Wulfsohn and Tsiatis (1997) point out a few drawbacks of the two-stage approach described above. First, it produces parameter estimates that are slightly biased toward the null, as indicated by simulations of Dafni (1993). Second, one may argue that the data are not used as efficiently as possible, since the method does not use any survival information in the longitudinal part of the model. Third, the normality assumption for the random effects in those at risk at each event time may not be reasonable. According to Wulfsohn and Tsiatis (1997), this is because "patients whose covariate trajectories have the steepest negative slopes may be at higher risk of mortality, and thus are removed from the population early on. This may result in
the random effects having a distributional shift toward a non normal distribution as time progresses” (p. 331). Fourth, the first-order approximation to $E[t, \gamma]$, namely

$$E[t, \gamma] = E[f(Y^*(t), \gamma)|\bar{Y}(t), X \geq t] \approx f(E[Y^*(t)|\bar{Y}(t), X \geq t], \gamma),$$

is used in fitting polynomial growth curve models (Tsiatis et al. (1995)). The approximation may not be valid, depending on the scaling of the covariate (Wulfsohn and Tsiatis (1997)).

An alternative to this method is to model the longitudinal covariate process simultaneously relating it to failure. This joint modeling approach, which has recently attracted much attention, has several advantages over the two-stage method presented above and is the subject of the next section.

### 1.4.2 Joint Modeling

In addition to using the data more efficiently, the joint modeling approach reduces the bias due to measurement error of the longitudinal covariate and informative censoring of the survival data, both of which are common problems in longitudinal studies. In other words, it results in more accurate estimates of the strength of the relationship between the covariate and the risk of failure. Moreover, the variance estimates of the survival model parameters correctly reflect the uncertainty in the longitudinal model parameters, and vice versa. (Faucett and Thomas (1996))
Faucett and Thomas (1996) present a joint model that is very similar to the two-stage model introduced by Tsiatis et al. (1995) which was presented in the previous section. The model allows for unequally spaced or missing data, different numbers of observations per individual, and censoring of the survival times. However, Faucett and Thomas (1996), in contrast to Tsiatis et al. (1995), do not allow the parameters describing the distribution of the random effects to depend on time. For the survival part of the model they assume a proportional hazard model with log-linear dependence on the true unobserved longitudinal covariate and with a piecewise constant baseline hazard function defined over some arbitrary time grid $\tau = (t_1, \ldots, t_K)$, where $t_1, \ldots, t_K$ do not have to be related to the covariate measurement times. Gibbs sampling is used to estimate the joint posterior distribution of all unknown parameters given only the observed data.

Wulfsohn and Tsiatis (1997) use the same model as Faucett and Thomas (1996) above, but using the EM likelihood maximization algorithm for parameter estimation. In the E-step, they obtain the expected log-likelihood of the complete data, conditional on the observed data and current parameter estimates, and in the M-step, they maximize the expected log-likelihood to obtain new parameter estimates. Only the estimate for $\gamma$, the relative risk parameter from the Cox model, is obtained using a one-step Newton-Raphson algorithm.
Xu and Zeger (2001) extend the model of Faucett and Thomas (1996), and propose a latent process model for the joint analysis of longitudinal covariate, \( Y \), and time to event data, \( T \), given covariates, such as treatment indicator and/or other risk factors, \( Z \), and then apply the model to estimate the marginal distribution \([T|Z]\) by using relevant information in \( Y \). More specifically, they assume there exists an underlying latent process \( Y^* \) corresponding to \( Y \). It is also assumed that \( T \) and \( Y \) are conditionally independent given \( Y^* \), \( Z \) can affect \( T \) either through \( Y^* \) or directly and \( Z \) only affects \( Y \) through its influence on \( Y^* \). Thus, \( Y \) can be thought of as an imperfect measure of \( Y^* \). The relationship between \( Y^* \) and \( Y \) is described by the following generalized linear model:

\[
g(E[Y_i(t)|Y_i^*(t)]) = Y_i^*(t),
\]

\[
Y_i^*(t) = Z_i(t)\beta + D_i(t)U_i + W_i(t),
\]

where \( g(\cdot) \) is an arbitrary link function, \( Z_i(t) \) is the set of observed covariates, \( \beta \) are regression coefficients, \( D_i(t) \) is a subset of \( Z_i(t) \) for which \( U_i \) are the associated Gaussian random effects with zero mean and unknown covariance matrix \( G \), and \( W_i(t) \) is a stationary Gaussian process with mean zero that produces serial autocorrelation dying away to zero as two observations for an individual become further separated in time. Finally, the failure time model, is specified as

\[
\lambda(t|Y_i^*(t), Z_i(t)) = \lambda_0(t, \gamma_0) \exp[s_1(Y_i^*(t), \gamma_1) + s_2(Z_i(t), \gamma_2)];
\]
where $s_1$ and $s_2$ are parametric models for the direct effect of $Y^*$ and $Z$ on the hazard, respectively, with the simplest case being

$$s_1(Y^*_i(t), \gamma_1) = Y^*_i(t) \gamma_1 \text{ and } s_2(Z_i(t), \gamma_2) = Z_i(t) \gamma_2.$$ 

The parameter estimation is carried out via the combination of Gibbs sampling and Metropolis-Hastings algorithm.

The model developed by Xu and Zeger (2001) is more flexible than the ones described previously. It applies to continuous, count or categorical longitudinal data, and, through the inclusion of $U_i$ and $W_i(t)$, allows for flexible modeling of the autocorrelation in the covariate process. Moreover, it allows for use of all available data even when some measurements are missing.

Wang and Taylor (2001) develop a model that relaxes the straight line assumption on the longitudinal covariate process. The method incorporates an integrated Ornstein-Uhlenbeck (IOU) process into the model, which is the continuous time version of AR(1) model. This way the authors expand the models of Faucett and Thomas (1996) and Wulfsohn and Tsiatis (1997). As before, the joint model consists of two linked sub-models: a model for the longitudinal data and a disease risk model for the survival data. The former includes a patient-specific random intercept $a_i$, the fixed effects of some covariates $Z_i$ (may be time-dependent), a slope $b$ (a population average rate of change of the marker), an IOU process $W_i$ (describing the random fluctuation of the marker around the population average) and an independent measurement error
term $e_i(t)$. Namely,

$$Y_i(t) = Y_i^*(t) + e_i(t)$$

$$Y_i^*(t) = a_i + bt + \beta Z_i(t) + W_i(t),$$

with both $e_i(t)$ and $a_i$ being iid and $e_{ij} \sim N(0, \sigma_e^2)$, $a_i \sim N(\mu_a, \sigma_a^2)$. The inclusion of $W_i$ in the model entails that each individual's longitudinal trajectory is a realization of a stochastic process, and that the slope of that trajectory, which combines $b$ and elements of $W_i(t)$, can vary over time and that it follows an IOU process. The event time data model is formulated by means of Cox PH model:

$$\lambda_i(t) = \lambda_0(t) \exp(\gamma Y_i^*(t) + \omega P_i(t)),$$

where $P_i$ are other covariates (may include some or all of the $Z_i$ covariates), and $\omega$ represents the effect of the covariates. The baseline hazard is assumed to be a step function. Bayesian techniques were used to fit the model, including MCMC sampling as a way of finding parameter estimates.

The advantage of this approach is that it provides a more flexible and reasonable structure for the patient-specific longitudinal trajectories than do the standard random effects models. Also, since covariates are included in both sub models, their effect on the covariate process and survival of the patients can be separated. As a disadvantage we may consider the complexity of the model.

An expanded idea for the longitudinal part of the model was delivered by Lange et al. (1992). They propose a Bayesian hierarchical approach accommodating indi-
individual piecewise-linear growth curves with random unobserved change points, unbalanced and incomplete data and several population level covariates. The method is used to model longitudinal series of CD4 T-cell counts for a group of HIV-infected patients. The error structure is assumed normal with mean zero but with heterogeneous variances, avoiding the implausible, in the authors' opinion, assumption of a common error variance shared by all individuals. To better study the relationship between the longitudinal marker and the progression of HIV infection, they also assume that each participant has a random unknown offset $\tau_i$, representing the time from seroconversion (i.e., development of detectable antibodies to HIV) to the start of the study. Namely, if $t_{ij}$ is the time of the $j^{th}$ measurement of the longitudinal covariate, then $t_{ij}^* = t_{ij} + \tau_i$ is the actual time from seroconversion to the $j^{th}$ measurement time of the $i^{th}$ patient. The $\tau_i$ can be thus considered as random errors but with non-zero mean. In fact, $P(\tau_i > 0) = 1$. A drawback of this approach is that the model introduces a large number of parameters, possibly even bigger than the sample size.

A method that combines ideas of Bayesian hierarchical modeling, nonparametric smoothing of time series, survival analysis, and forecasting into a unified framework was proposed by Berzuini and Larizza (1996). To assure flexible modeling of the longitudinal data, stochastic process theory is used. Namely, let $Y_{kj}$ be the $j^{th}$ response observation and $Z_{kj}$ a covariate vector for the $k^{th}$ individual at time $t_{kj}$. Initially, it is assumed that $(Y_{kj}|\xi_{kj}, \tau_y) \sim N(\xi_{kj}, \tau_y)$ and $(Z_{kj}|\omega_{kj}, \tau_z) \sim N(\omega_{kj}, \tau_z)$. $\omega_{kj}$'s (for
which $Z_{kj}$'s are a proxy measure) are the true factors that influence individual's response. They are later assumed to be constant over time and normally distributed, i.e., $\omega_{kj} = \omega_k \sim N(\mu_\omega, \tau_\omega)$. The remaining specification of the model is as follows:

$$
\xi_{kj} = \beta_{k1} t_{kj} + \beta_2 \omega_k + \epsilon_{kj},
$$

($\beta_2$ is a vector of unknown covariate coefficients),

$$
\beta_{k1} \sim N(\mu_\beta, \tau_\beta),
$$

$$
\epsilon_{k1} \sim N(\mu_\epsilon, \tau_\epsilon),
$$

$$
\epsilon_{k2} \sim N(\mu_\omega, \tau_\omega),
$$

$$
\epsilon_{kj} \sim N((1 + \eta_s)\epsilon_{k,j-1} - \eta_s \epsilon_{k,j-2}, (1 - \eta_s^2)\eta_v), \quad j > 2,
$$

$$
\tau_\omega^{-1}, \tau_\epsilon^{-1}, \tau_z^{-1} \sim \text{Gamma},
$$

$$
\tau_\beta^{-1}, \tau_\omega^{-1}, \eta_v^{-1} \sim \text{Gamma},
$$

$$
\mu_0, \mu_\omega, \mu_\beta, \eta_s \sim \text{Normal}, \quad 0 \leq \eta_s \leq 1.
$$

Berzuini and Larizza (1996) point out that there is a close relationship between the stochastic model above and dynamic models, in which conditionally independent observations are generated by a latent Gaussian Markovian process. Thus, noting that $\epsilon_{kj}$ and $\gamma_{k,j-1} = \epsilon_{kj} - \epsilon_{k,j-1}$ jointly form a two-dimensional Gaussian Markov process, and assuming equally spaced data, leads to the following alternative nonlinear dynamic model representation of the model above:
\[ \epsilon_{kj} = \epsilon_{k,j-1} + \gamma_{k,j-1}. \]

\[ \gamma_{kj} \sim N \left( \eta_s \gamma_{k,j-1}; (1 - \eta_s^2) \frac{\eta_u}{h^2} \right), \]

\[ \gamma_{k1} \sim N(0, \eta_u). \]

\[ \epsilon_{k1} \sim N(\mu_0, \tau_0). \]

where \( h = t_{kj} - t_{k,j-1} \) and all the remaining components of the previous model representation are left unchanged. After introducing the stochastic process approach to time series modeling, the authors go on to model the failure data. They allow for multiple failure events per individual representing consequences of a disease (e.g., infections in patients diagnosed with AIDS). Thus, the survival data is given in the form of counts of failures for each individual in a sequence of time intervals. Then the relationship between the time series data and failures is modeled. This is done by allowing the parameters that underlie each individual’s longitudinal covariate process to act as regressors in a Poisson regression model for the failure counts. More specifically, if \( d_{kj} \) and \( \lambda_{kj} \) are the observed count of failures and the unknown hazard of failure for individual \( k \) in the \( j^{th} \) time interval, respectively, then \( d_{kj} \)’s contribution to the likelihood is of the form \( \lambda_{kj}^{d_{kj}} \exp(-\lambda_{kj}) \). Thus, \( d_{kj} \) can be assumed to follow Poisson distribution with mean \( \lambda_{kj} \). Now, the hazard \( \lambda_{kj} \) is modeled by the proportional hazards model, that can be expressed in the following form:

\[ \log \lambda_{kj} = \lambda_{0j} + \phi_k + \Theta_{kj} + f_r, \]

where \( \lambda_{0j} \) represents the baseline hazard and dependence of hazard on time, \( \phi_k \) are
frailties, allowing for random variability of the hazard across individuals, $\theta_2$ are regression parameters, and $f_r$ represents the dependence of the hazard on current response $\xi_{k,j}$ and on past response history ($\xi_{k,j-1}, \xi_{k,j-2}, \ldots$). MCMC techniques are used for parameter estimation.

Brown and Ibrahim (2003) develop a method that is very similar to the ones presented in Wulfsohn and Tsiatis (1997) and Faucett and Thomas (1996), but relaxes the distributional assumptions on the longitudinal model parameters, by placing a Dirichlet process (DP) prior on the distribution of these parameters. The model incorporates information about missing values, measurement error and links the hazard rate to longitudinal measurements through Cox proportional hazard model. Namely, the longitudinal model is

$$Y_i(t) = Y_i^*(t) + e_i(t), \quad e_i(t) \sim N(0, \sigma^2),$$

$$Y_i^*(t) = \beta_0 + \beta_1 t + \beta_2 t^2,$$

$$\beta_i \sim G, \quad G \sim DP(M, G_0), \quad G_0 = N(b_0, V_0),$$

with $\beta_i = (\beta_{0i}, \beta_{1i}, \beta_{2i})'$, and normal, Wishart and gamma priors placed on $b_0$, $V_0$, and $M$, respectively. If $M \to \infty$, then the model above corresponds to a fully parametric model in which a distributional assumption is placed on $\beta_i$, i.e., $\beta_i \sim G_0$. Gibbs sampling is used to find estimates of $\beta$'s and other parameters of interest.

Brown and Ibrahim (2003) point out that one of the applications of this semi-parametric Bayesian hierarchical model could be to validate a parametric model. No
improvement over the regular parametric model proves the validity of our parametric assumptions. The model, if necessary or desired, could also be easily extended to incorporate other covariates in the trajectory function.

Other approaches for joint modeling of longitudinal and time-to-event data include, but are not limited to, DeGruttola and Tu (1994); Hogan and Laird (1997a,b); Henderson et al. (2000); Tsiatis and Davidian (2001); Song et al. (2002); Brown et al. (2005). Also, a comprehensive overview of some of the existing methods can be found in Troxel (2002) and Tsiatis and Davidian (2004).

1.4.3 Remarks

Joint models for longitudinal and survival data allow for direct assessment of the dependence among various components of the data. As any other statistical methods, however, the models depend on explicit assumptions, especially for the longitudinal part of the data, and thus need to be thoroughly tested and sensitivity analysis must be performed. Also, their computational complexity can be a drawback, but with current advancement in computer technology this becomes a smaller issue. (Troxel (2002))
1.5 Previous Research on Dynamic Models for Survival Data

In most models for survival data, the covariates are assumed to have fixed effects on the risk of failure of an individual. In some applications this may be reasonable and sufficient. In general, however, it is too restrictive, since it is possible that the effects do change with time, as is the case with unemployment studies (Gamerman and West (1987)), for example. Dynamic models, that let the parameters vary in time, allowing the data and prior assumptions to determine the form of variation, provide a solution to such problems. For instance, Gamerman (1991) extends the static Cox PH model by allowing the parameters representing the covariate effects to vary between time intervals, and a system equation provides a description of the stochastic evolution of these parameters over time. The approach uses sequential analysis, based on factorization of the likelihood over the time intervals, and assumes piecewise exponential (PE) distribution for the survival times of the patients. We say that a random variable $T$ has a PE distribution, $T \sim \text{PE}(\lambda, \tau)$, with $\lambda = (\lambda_1, \ldots, \lambda_s)$ and $\tau = \{t_0, \ldots, t_{s-1}\}$, for some integer $s$, and ordered times $t_0 \leq t_1 \leq \ldots \leq t_{s-1}$, if
its hazard is

\[
\lambda(t) = \begin{cases} 
\lambda_1 & \text{if } t \in I_1 = [t_0, t_1), \\
\vdots & \\
\lambda_i & \text{if } t \in I_i = (t_{i-1}, t_i], \\
\vdots & \\
\lambda_s & \text{if } t \in I_s = (t_{s-1}, \infty). 
\end{cases}
\]

The grid \( \tau \) may consist of the ordered death times of the patients or can be selected independently of the data. Assuming \( T \sim PE(\lambda, \tau) \), the conditional survival and density functions of \( T \), given \( T \geq t_i, i = 0, \ldots, s - 1 \), required later in the sequential analysis, are:

\[
S(t|T \geq t_{i-1}) = \exp(-\lambda_i(t - t_{i-1})), \quad t \in I_i, \tag{1.24}
\]

\[
f(t|T \geq t_{i-1}) = \lambda_i \exp(-\lambda_i(t - t_{i-1})), \quad i = 1, \ldots, s. \tag{1.25}
\]

Now, following the notation by Gamerman (1991), let \( T = (T_1, \ldots, T_n) \) be a set of independent survival times observed from \( n \) patients under study, and \( \tau = \{t_1, \ldots, t_N\} \) (with \( t_N \) bigger than the largest observed survival time). The model suggested by Gamerman (1991) is given by the following components:

- observation equation: \( T_j \sim PE(\lambda^{(j)}, \tau), \quad \lambda^{(j)} = (\lambda_1^{(j)}, \ldots, \lambda_N^{(j)}), \quad j = 1, \ldots, n; \)

- guide relation: \( \log \lambda_i^{(j)} = z_i' \beta_i, \quad i = 1, \ldots, N; \)

- evolution equation: \( \beta_i = G_i(b_i)\beta_{i-1} + w_i. \tag{1.26} \)
where $z_j' = (1, z_{j1}, \ldots, z_{jp})'$, with $z_{jk}, k = 1, \ldots, p$, being the value of the $k^{th}$ covariate for the $j^{th}$ individual; $\beta_j' = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{pj})'$ is the set of regression coefficients, with $\beta_{0j} = \log(\lambda_{0j})$ and $\lambda_{0j}$ being the value of the baseline hazard function of the $j^{th}$ individual in the $i^{th}$ time interval; $G_i$ is the system evolution matrix over $I_i$, and usually depends on $b_i = \text{length}(I_i)$ (often $G_i = I_p$, with $I_p$ being a $p$-dimensional identity matrix); and $w_i$ is the cumulative error over $I_i$ (usually $w_i \sim [0, W_i]$).

As mentioned earlier, the likelihood of the data is divided into time intervals. The idea, called temporal factorization of likelihood and proved by Gamerman (1987), "...use[s] a factorized timescale to process data information at each time conditionally on the information from previous times" (Gamerman (1991), p. 65), and is structured as follows. Let $D_i$ be the information available at time $t_i$ of the chosen time grid $\tau$.

As Gamerman (1991) points out, each individual $j = 1, \ldots, n$ contributes to $D_i$ in one of the following three ways:

1. $j$ has died at $T_j = t_j \leq t_i$,

2. $j$ was censored at $T_j = t_j \leq t_i$,

3. $j$ is still alive at $t_i$, so that $T_j > t_i$.

In cases 1 and 2, $T_j$ provides no additional information after $t_i$. With $\Lambda = \{\lambda^{(1)}, \ldots, \lambda^{(n)}\}$, the full data likelihood can be thus written as:
\[
L(T|\Lambda, D_0) = \prod_{i=1}^{N} L_i(T^{(i)}|\Lambda, D_{i-1}),
\]

\[
L_i(T^{(i)}|\Lambda, D_{i-1}) = \prod_{w \in F(I_i)} f(t_{iw}|\Lambda, D_{i-1}) \prod_{q \in R(I_i)} S(t_{iq}|\Lambda, D_{i-1}), \quad i = 1, \ldots, N,
\]
where \( T^{(i)} = \{\min(T_k, t_i), \ k \in F(I_i) \cup R(I_i)\} \), \( t_{iw} = t_w, \ w \in F(I_i), \ t_{iq} = \min\{t_q, t_i\}, \)
\( q \in R(I_i), \ F(I_i) \) is the set of individuals dying in \( I_i \), and \( R(I_i) \) is the set of individuals censored in or surviving through \( I_i \). Now, plugging equations (1.24) and (1.25) into (1.27), the likelihood factors \( \{L_i, \ i = 1, \ldots, N\} \) for the assumed model are

\[
\prod_{j=1}^{r_i} (\lambda_i^{(j)})^{\delta_{ij}} \exp(-\lambda_i^{(j)}(t_{ij} - t_{i-1})),
\]
in which \( r_i \) is the number of individuals that are alive at the beginning of \( I_i \), \( \delta_{ij} \) is the death indicator for the \( j^{th} \) individual, and \( t_{ij} \) is the survival time for the \( j^{th} \) individual in \( I_i \), which is equal to \( t_i \) if the individual survives through \( I_i \). As noted by Gamerman (1991), the Bayesian analysis can be then carried out by assuming full distributions for \( (\boldsymbol{\beta}_i|D_0) \) and \( \omega_i \), and cycling through intervals \( I_i, \ i = 1, \ldots, N \), via the evolution equation (1.26) and likelihood factors (1.28) written in terms of the \( \beta \)'s. The downside of this procedure is its computational complexity, as well as difficulty in choosing the above distributions. Thus, Gamerman (1991) proposes an alternative approach to obtaining the estimates of the model parameters. Let \([m_{t-1}, C_{t-1}]\) be the posterior distribution of \( \beta_{t-1} \) at \( t_{i-1} \), partially specified by its two first moments. Then, the prior distribution for \( \beta \) at \( t_i \) is \([\alpha_i, P_i]\) with \( \alpha_i = G_i(b_i)m_{t-1} \), and \( P_i = G_i(b_i)C_{t-1}G'_i(b_i) + W_i \). This prior is then updated into the posterior distribution of \( \beta \) at time \( t_i \) (denoted by \([m_i, C_i]\)) by processing all the observations that
contribute to the likelihood factor $L_I$. The updating equations for the distributions of the model parameters were derived by Gamerman (1991) using the methods proposed by West et al. (1985) and can be found in Gamerman (1991), section 4.2.

A similar approach for dynamic modeling of discrete time survival data was presented by Fahrmeir (1994). The method is related to the dynamic version of the piecewise exponential model and an extension of point processes studied by Gamerman (1991, 1992). As in Gamerman (1991) the follow-up period is partitioned into intervals $I_t$, $i = 1, \ldots, d$, and each observation is only known to lie in one of these intervals. Following the notation by Fahrmeir (1994), the proposed model is:

- observation model:

$$P(Y_{it} = 1|Y_{i-1}^*, \mathbf{x}_{it}^*, \mathbf{r}_t^*; \gamma_t, \mathbf{\beta}_t) = h(\gamma_t + \mathbf{w}_{it}'\mathbf{\beta}_t),$$

where $h$ is a response function (usually exp), $\mathbf{r}_t^* = (r_{1t}, \ldots, r_{rt})$ is the history of risk indicators up to $t$, where $r_{it} = (r_{it}, i \geq 1)$, with $r_{it}$ being a risk indicator that is equal to 1 if an individual is at risk in $I_t$ and 0 otherwise, $Y_{it}$ is a failure indicator equal to 1 if an individual $i$ fails in $I_t$ and 0 if he survives, with $i \in R_t$, $t = 1, \ldots, t_i$ ($t_i = \min(T_i, C_i)$), $R_t = \{i : r_{it} = 1\}$ is a risk set at $t$, $\mathbf{x}_{it}^* = (\mathbf{x}_1, \ldots, \mathbf{x}_t)$ is the history of possibly time-dependent covariates up to $t$, with $\mathbf{x}_t = (x_{it}, i \in R_t)$, $Y_i^* = (Y_1, \ldots, Y_t)$ is the failure history up to $t$, with $Y_t = (Y_{it}, i \in R_t)$, $\gamma_t$ are unknown baseline hazard parameters, $\mathbf{\beta}_t$ is a possibly time-varying vector of covariate effects, and $\mathbf{w}_{it}$ is an appropriate function of
\( Y^*_{t-1} \) and \( x^*_t \) (in the simplest case \( w_{it} \) equals the vector of covariates).

- transition models:

  - for stochastic variation of baseline hazard parameters \( \{\gamma_t\} \) we may use:
    
    * first or second order random walks:
      
      \( \gamma_t = \gamma_{t-1} + \omega_t \) or \( \gamma_t = 2\gamma_{t-1} - \gamma_{t-2} + \omega_t, \quad \omega_t \sim N(0, \sigma^2_w), \)
    
    * the local linear trend model:
      
      \( \gamma_t = \gamma_{t-1} + \eta_t + \omega_t, \quad \eta_t = \eta_{t-1} + \nu_t, \quad \nu_t \sim N(0, \sigma^2_v), \)
      
      with \( \{\omega_t\} \) and \( \{\nu_t\} \) being mutually independent white noises,

  - for covariate effects:

    \[ \beta_t = \beta_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma), \quad \Sigma \text{ diagonal.} \]

Obviously, further prior structure can be introduced through more sophisticated transition models. If we write \( \alpha_t = (\gamma_t, \beta_t) \), then the transition models above can be written in a form of one general transition model:

\[ \alpha_t = F \alpha_{t-1} + \xi_t, \quad (t \geq 1), \quad (1.29) \]

where \( F \) is a known transition matrix, \( \{\xi_t \sim N(0, Q)\} \) is a white noise sequence, independent of the initial state \( \alpha_0 \sim N(\alpha_0, Q_0) \). Then the models presented above are special cases of the model (1.29). For example, by setting \( F = I \), a joint first-order random walk model for \( \alpha_t \) is obtained.
In order to be able to completely specify the models in terms of likelihoods, Fahrmeir (1994) introduces the following three assumptions:

1. Individual failure indicators $Y_{it}$ ($i \in R_t$) within $I_t$ are conditionally independent, given $Y_{t-1}^*, x_t^*, r_t^*$, and $\alpha_t$.

2. $x_t$ and $r_t$ are independent of $\alpha_{t-1}^* = (\alpha_{t-1}, \ldots, \alpha_0)$, given $Y_{t-1}^*, x_{t-1}^*, r_{t-1}^*$, i.e., "...the covariate and the censoring process contain no information on the parameter sequence" (Fahrmeir (1994), p. 320). As noted by Kalbfeisch and Prentice (1980) (p. 124), this assumption may not be satisfied for internal covariates whose paths are affected by failure.

3. $Y_t$ are conditionally independent of $\alpha_{t-1}^*$, given $Y_{t-1}^*, x_t^*, r_t^*$, and $\alpha_t$, i.e., "conditional information of $\alpha_t^*$ on $Y_t$ is already contained in the current parameter $\alpha_t$ alone" (Fahrmeir (1994), p. 320).

The estimation of the parameters is based on posterior modes and penalized likelihood, and will not be presented here in detail. The method is not a fully Bayesian approach, thus losing some of its useful features, e.g., model comparison based on Bayes factors.
Chapter 2

Preliminary Analysis

The joint dynamic modeling of longitudinal and survival data can be thought of as two separate pieces of analysis that are put together. One of them is the dynamic model for time-to-event data with non-longitudinal covariates, and the second is the dynamic linear model (DLM) for a longitudinal covariate. This chapter includes a couple of real data examples of these two types of models. Let us first take a look at an example of the former one.

2.1 The Gastric Cancer Patients Data Analysis

Data
The data consist of the survival times of 90 patients diagnosed with gastric cancer (Stablein et al. (1981)). The patients are divided equally into two treatment groups: combined treatment of chemotherapy and radiation, and chemotherapy only. There
are 10 censored observations, and two of observed survival times repeat.

2.1.1 The PE Model

Gamerman (1991) adapts the following model for the data. Let us define the time grid \( \tau = (t_0, \ldots, t_{N-1}, t_N) \) as the observed death times, with \( t_0 = 0 \), \( t_{N-1} \) being the last observed failure, and \( t_N = \infty \), so that \( I_i \) is the time interval \( (t_{i-1}, t_i], i = 1, \ldots, N \). Let us also assume that the survival times are \( PE(\lambda, \tau) \) distributed, with \( \lambda = (\lambda_1, \ldots, \lambda_N) \). The hazard function in the \( i \)th interval is modeled as:

\[
\lambda(t) = \exp(\beta_i' z), \quad t \in I_i,
\]

\[
\beta_i = \beta_{i-1} + w_i, \quad w_i \sim N(0, W_i),
\]

where

\[
W_i = \begin{pmatrix}
\sigma_{i0}^2 & 0 \\
0 & \sigma_{ii}^2
\end{pmatrix} = \begin{pmatrix}
0 & 0 \\
0 & 0.04
\end{pmatrix}, \quad \beta_i' = (\beta_{i0}, \beta_{i1}), \quad z' = (1, z), \quad \text{for all } i.
\]

\( z \) is the treatment indicator coded as 1 for the combined treatment, and 0 for chemotherapy only. A non-informative \( N(0, 10^3 I_2) \) prior is assumed for \( \beta_0 \), where \( I_2 \) is a two-dimensional identity matrix.

Analysis

For comparison, it is worth pointing out that Carter et al. (1983) fitted Cox's non-parametric model to the data with two covariates: \( z \) and \( t \cdot z \). In the formulation of
the dynamic extension of the Cox PH model in Gamerman (1991) (sec. 2) given by

\[ \lambda(t) = \exp(z'\beta(t)), \]

(with \( \beta(t) = (\beta_0(t), \beta_1(t)) \)), this is equivalent to assuming \( \beta_1(t) = \alpha_1 + \alpha_2 t \). With \( t \) measured in months, the estimates of \( \alpha_1 \) and \( \alpha_2 \) calculated by Carter et al. (1983) were 1.2771 and -0.0794, respectively. The straight diagonal line in Figure 2.1 represents this solution.

The dynamic analysis of the data was carried out in two ways. First, according to the procedure presented by Gamerman (1991), in which the updating formulas are based on the methods presented by West et al. (1985), and then using the auxiliary particle filter approach. Let us recall that the general idea of the former method is to update the partially specified prior \( (\beta_i|D_{i-1}) \sim [m_{i-1}, P_i] \) (where \( P_i = C_{i-1} + W_i \), and \( m_{i-1}, C_{i-1} \) are respectively the mean and the covariance matrix of \( (\beta_{i-1}|D_{i-1}) \), and \( i = 1, \ldots, N \)) into a posterior \( (\beta_i|D_i) \sim [m_i, C_i] \) by processing \( r_i \) ordered observations that contribute to the likelihood factor \( L_i \) in (1.27). More details of this approach, including full updating equations, can be found in Gamerman (1991) in section 4.2.

The results obtained by Gamerman (1991) were reproduced using this procedure and are as follows. Since \( \sigma_{i0} = 0 \) was used in the analysis, \( \beta_{i0} \) is constant over time (i.e., the distribution of the survival times is exponential when \( z = 0 \)). As can be seen in Figure 2.1, significant variation of the treatment effect parameter \( \beta_{i1} \) over time was detected. The posterior mean of \( \beta_{i1} \) changes from positive values at early times to negative ones as time passes. The mean smoothed trajectory of \( \beta_{i1} \) (also
included in Figure 2.1) follows a similar pattern. It is more useful and more precise than the on-line trajectory, since it is computed given all the information available \( D_N \). (Gamerman (1991), sec. 6.1)

Having all the on-line posterior means, \( m_i \), and covariance matrices, \( C_i \), \( i = 1, \ldots, N \), saved, and writing \( (\theta_i|D_N) \sim [m_i^N, C_i^N] \), with \( m_N^N = m_N \), and \( C_N^N = C_N \), the first two moments of the smoothed posterior distribution were computed according to the following recursive formulas (Gamerman (1991), sec. 4.3):

\[
\begin{align*}
  m_i^N &= m_i + C_i P_{i+1}^{-1} [m_{i+1}^N - m_i], \\
  C_i^N &= C_i - C_i P_{i+1}^{-1} [P_{i+1} - C_{i+1}^N] P_{i+1}^{-1} C_i.
\end{align*}
\]

The complete results of this analysis can be found in Table 2.1.
<table>
<thead>
<tr>
<th>(i)</th>
<th>(m_{1i})</th>
<th>(m_{1i}^N)</th>
<th>(\sigma_{1i}^2)</th>
<th>(\sigma_{1i}^{2N})</th>
<th>(i)</th>
<th>(m_{1i})</th>
<th>(m_{1i}^N)</th>
<th>(\sigma_{1i}^2)</th>
<th>(\sigma_{1i}^{2N})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-9.398</td>
<td>0.182</td>
<td>1.000.04</td>
<td>0.251</td>
<td>41</td>
<td>0.804</td>
<td>0.023</td>
<td>0.562</td>
<td>0.224</td>
</tr>
<tr>
<td>2</td>
<td>-0.483</td>
<td>0.182</td>
<td>1.995</td>
<td>0.211</td>
<td>42</td>
<td>0.197</td>
<td>0.047</td>
<td>0.599</td>
<td>0.213</td>
</tr>
<tr>
<td>3</td>
<td>0.469</td>
<td>0.212</td>
<td>1.507</td>
<td>0.185</td>
<td>43</td>
<td>0.197</td>
<td>0.077</td>
<td>0.422</td>
<td>0.199</td>
</tr>
<tr>
<td>4</td>
<td>0.913</td>
<td>0.270</td>
<td>1.352</td>
<td>0.168</td>
<td>44</td>
<td>0.079</td>
<td>0.083</td>
<td>0.460</td>
<td>0.195</td>
</tr>
<tr>
<td>5</td>
<td>1.250</td>
<td>0.314</td>
<td>1.281</td>
<td>0.158</td>
<td>45</td>
<td>0.147</td>
<td>0.096</td>
<td>0.360</td>
<td>0.186</td>
</tr>
<tr>
<td>6</td>
<td>1.519</td>
<td>0.340</td>
<td>1.242</td>
<td>0.153</td>
<td>46</td>
<td>0.003</td>
<td>0.114</td>
<td>0.398</td>
<td>0.185</td>
</tr>
<tr>
<td>7</td>
<td>0.819</td>
<td>0.366</td>
<td>0.783</td>
<td>0.152</td>
<td>47</td>
<td>0.398</td>
<td>0.147</td>
<td>0.328</td>
<td>0.180</td>
</tr>
<tr>
<td>8</td>
<td>1.055</td>
<td>0.400</td>
<td>0.744</td>
<td>0.144</td>
<td>48</td>
<td>0.438</td>
<td>0.149</td>
<td>0.290</td>
<td>0.182</td>
</tr>
<tr>
<td>9</td>
<td>1.293</td>
<td>0.429</td>
<td>0.721</td>
<td>0.141</td>
<td>49</td>
<td>0.358</td>
<td>0.134</td>
<td>0.328</td>
<td>0.192</td>
</tr>
<tr>
<td>10</td>
<td>1.386</td>
<td>0.427</td>
<td>0.707</td>
<td>0.139</td>
<td>50</td>
<td>0.238</td>
<td>0.125</td>
<td>0.366</td>
<td>0.197</td>
</tr>
<tr>
<td>11</td>
<td>1.555</td>
<td>0.455</td>
<td>0.698</td>
<td>0.140</td>
<td>51</td>
<td>0.021</td>
<td>0.127</td>
<td>0.404</td>
<td>0.197</td>
</tr>
<tr>
<td>12</td>
<td>1.133</td>
<td>0.471</td>
<td>0.571</td>
<td>0.144</td>
<td>52</td>
<td>-0.027</td>
<td>0.151</td>
<td>0.443</td>
<td>0.193</td>
</tr>
<tr>
<td>13</td>
<td>1.346</td>
<td>0.494</td>
<td>0.551</td>
<td>0.141</td>
<td>53</td>
<td>0.307</td>
<td>0.176</td>
<td>0.347</td>
<td>0.185</td>
</tr>
<tr>
<td>14</td>
<td>1.423</td>
<td>0.488</td>
<td>0.538</td>
<td>0.141</td>
<td>54</td>
<td>0.211</td>
<td>0.169</td>
<td>0.385</td>
<td>0.184</td>
</tr>
<tr>
<td>15</td>
<td>1.102</td>
<td>0.489</td>
<td>0.495</td>
<td>0.144</td>
<td>55</td>
<td>0.478</td>
<td>0.170</td>
<td>0.317</td>
<td>0.179</td>
</tr>
<tr>
<td>16</td>
<td>1.144</td>
<td>0.499</td>
<td>0.472</td>
<td>0.139</td>
<td>56</td>
<td>0.279</td>
<td>0.133</td>
<td>0.356</td>
<td>0.181</td>
</tr>
<tr>
<td>17</td>
<td>1.180</td>
<td>0.529</td>
<td>0.457</td>
<td>0.137</td>
<td>57</td>
<td>0.545</td>
<td>0.119</td>
<td>0.301</td>
<td>0.177</td>
</tr>
<tr>
<td>18</td>
<td>1.376</td>
<td>0.587</td>
<td>0.448</td>
<td>0.138</td>
<td>58</td>
<td>0.492</td>
<td>0.061</td>
<td>0.340</td>
<td>0.181</td>
</tr>
<tr>
<td>19</td>
<td>1.041</td>
<td>0.624</td>
<td>0.438</td>
<td>0.140</td>
<td>59</td>
<td>0.713</td>
<td>0.005</td>
<td>0.292</td>
<td>0.180</td>
</tr>
<tr>
<td>20</td>
<td>1.276</td>
<td>0.690</td>
<td>0.418</td>
<td>0.136</td>
<td>60</td>
<td>0.944</td>
<td>-0.104</td>
<td>0.264</td>
<td>0.187</td>
</tr>
<tr>
<td>21</td>
<td>1.481</td>
<td>0.732</td>
<td>0.405</td>
<td>0.134</td>
<td>61</td>
<td>0.280</td>
<td>-0.285</td>
<td>0.303</td>
<td>0.202</td>
</tr>
<tr>
<td>22</td>
<td>1.593</td>
<td>0.748</td>
<td>0.397</td>
<td>0.133</td>
<td>62</td>
<td>0.240</td>
<td>-0.360</td>
<td>0.342</td>
<td>0.213</td>
</tr>
<tr>
<td>23</td>
<td>1.776</td>
<td>0.752</td>
<td>0.391</td>
<td>0.134</td>
<td>63</td>
<td>-0.146</td>
<td>-0.434</td>
<td>0.381</td>
<td>0.220</td>
</tr>
<tr>
<td>24</td>
<td>1.950</td>
<td>0.721</td>
<td>0.388</td>
<td>0.136</td>
<td>64</td>
<td>-0.302</td>
<td>-0.462</td>
<td>0.420</td>
<td>0.223</td>
</tr>
<tr>
<td>25</td>
<td>1.964</td>
<td>0.648</td>
<td>0.386</td>
<td>0.141</td>
<td>65</td>
<td>-0.367</td>
<td>-0.476</td>
<td>0.459</td>
<td>0.222</td>
</tr>
<tr>
<td>26</td>
<td>1.650</td>
<td>0.567</td>
<td>0.392</td>
<td>0.148</td>
<td>66</td>
<td>-0.024</td>
<td>-0.486</td>
<td>0.349</td>
<td>0.217</td>
</tr>
<tr>
<td>27</td>
<td>1.616</td>
<td>0.515</td>
<td>0.376</td>
<td>0.149</td>
<td>67</td>
<td>-0.062</td>
<td>-0.543</td>
<td>0.388</td>
<td>0.223</td>
</tr>
<tr>
<td>28</td>
<td>1.825</td>
<td>0.470</td>
<td>0.367</td>
<td>0.154</td>
<td>68</td>
<td>0.021</td>
<td>-0.598</td>
<td>0.315</td>
<td>0.225</td>
</tr>
<tr>
<td>29</td>
<td>1.456</td>
<td>0.373</td>
<td>0.383</td>
<td>0.163</td>
<td>69</td>
<td>-0.321</td>
<td>-0.684</td>
<td>0.354</td>
<td>0.240</td>
</tr>
<tr>
<td>30</td>
<td>1.646</td>
<td>0.321</td>
<td>0.362</td>
<td>0.167</td>
<td>70</td>
<td>-0.386</td>
<td>-0.726</td>
<td>0.393</td>
<td>0.252</td>
</tr>
<tr>
<td>31</td>
<td>1.496</td>
<td>0.223</td>
<td>0.384</td>
<td>0.176</td>
<td>71</td>
<td>-0.442</td>
<td>-0.763</td>
<td>0.432</td>
<td>0.260</td>
</tr>
<tr>
<td>32</td>
<td>0.732</td>
<td>0.147</td>
<td>0.399</td>
<td>0.182</td>
<td>72</td>
<td>-0.637</td>
<td>-0.797</td>
<td>0.335</td>
<td>0.265</td>
</tr>
<tr>
<td>33</td>
<td>0.968</td>
<td>0.147</td>
<td>0.353</td>
<td>0.182</td>
<td>73</td>
<td>-0.705</td>
<td>-0.811</td>
<td>0.374</td>
<td>0.287</td>
</tr>
<tr>
<td>34</td>
<td>1.149</td>
<td>0.112</td>
<td>0.327</td>
<td>0.189</td>
<td>74</td>
<td>-0.766</td>
<td>-0.820</td>
<td>0.414</td>
<td>0.306</td>
</tr>
<tr>
<td>35</td>
<td>0.821</td>
<td>0.039</td>
<td>0.358</td>
<td>0.205</td>
<td>75</td>
<td>-0.539</td>
<td>-0.824</td>
<td>0.325</td>
<td>0.324</td>
</tr>
<tr>
<td>36</td>
<td>0.638</td>
<td>0.011</td>
<td>0.390</td>
<td>0.218</td>
<td>76</td>
<td>-0.626</td>
<td>-0.857</td>
<td>0.365</td>
<td>0.364</td>
</tr>
<tr>
<td>37</td>
<td>0.541</td>
<td>0.000</td>
<td>0.424</td>
<td>0.227</td>
<td>77</td>
<td>-0.781</td>
<td>-0.882</td>
<td>0.404</td>
<td>0.404</td>
</tr>
<tr>
<td>38</td>
<td>0.447</td>
<td>-0.009</td>
<td>0.458</td>
<td>0.232</td>
<td>78</td>
<td>-0.881</td>
<td>-0.891</td>
<td>0.444</td>
<td>0.444</td>
</tr>
<tr>
<td>39</td>
<td>0.172</td>
<td>-0.015</td>
<td>0.493</td>
<td>0.233</td>
<td>79</td>
<td>-0.892</td>
<td>-0.892</td>
<td>0.484</td>
<td>0.484</td>
</tr>
<tr>
<td>40</td>
<td>0.020</td>
<td>0.002</td>
<td>0.525</td>
<td>0.230</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: On-line \((m_{1i}, \sigma_{1i}^2)\) and smoothed \((m_{1i}^N, \sigma_{1i}^{2N})\) posterior means and variances of \(\beta_{1i}, i = 1, \ldots, N\) (Gamerman (1991) method).
Figure 2.1: Estimated posterior mean trajectories for the treatment difference effect: Gamerman (1991) method vs. auxiliary particle filter method.

One potential problem with Gamerman's procedure arises if we choose to take a different time grid than the one of ordered observed failure times. Since the distributions $(\beta_{i-1}|D_{i-1})$ are only partially specified in terms of their mean and covariance matrix and may be improper, then if there are no observations in $I_i$, the posterior distribution $(\beta_i|D_i)$ may be improper as well.

To avoid this, the data can be analyzed in a different way. This is by going through intervals $I_i$, via the evolution equation (1.26) and likelihood factors (1.28) expressed in terms of $\beta_i$'s. Full distributions for $(\beta_i|D_0)$ and $w_i$ are assumed, and are $N(0, 10^3 I_2)$ and $N(0, W_i)$, respectively, where $W_i$ is as defined earlier in the section. For the first interval, Metropolis-Hastings algorithm can be used to generate a sample from $(\beta_1|D_1)$. The algorithm, however, requires a closed form prior distribution for
the parameter, and hence cannot be used beyond the initial time interval. This is because when processing the $i^{th}$ interval, the prior distribution for the parameter in this interval, $(\beta_i|D_{i-1})$, is, through the evolution equation, directly related to the posterior distribution of the parameter in the previous interval, $(\beta_{i-1}|D_{i-1})$, which is not available in closed form. Thus, a method that could use a discrete posterior sample from $(\beta_{i-1}|D_{i-1})$ to approximate the prior $(\beta_i|D_{i-1})$ is necessary here. This is exactly what auxiliary particle filters, described in section 1.3.4, do. The results of this method applied to the gastric cancer patients data, assuming $PE(\lambda, \tau)$ distribution for the survival times, can be found in Table 2.2 and are also plotted in Figures 2.1 and 2.2. From Figure 2.1 we can see that both methods performed similarly. If we take the results obtained according to Gamerman (1991) updating equation procedure as the point of reference, then we can see that the auxiliary particle filter performed equally well, with subtle differences in the estimates at the early stages. It should be noted that the posterior means for $\beta_0$ agreed and were -6.53 in both cases.

As can be seen in Figure 2.2, the posterior variances for both methods are quite different, being fairly close at the early stages, and as time goes by, differing more significantly. The values obtained from auxiliary particle filter method are significantly smaller than their corresponding values from Gamerman (1991) method. This, perhaps, is due to the nature of the auxiliary particle filter methods, which can result in underestimated variances of the parameters. The pattern of the on-line variance trajectories for the treatment difference effect are as expected: larger values at the be-
<table>
<thead>
<tr>
<th>i</th>
<th>$m_{i1}$</th>
<th>$m_{i1}^N$</th>
<th>$\sigma_{i1}^2$</th>
<th>$\sigma_{i1}^2$</th>
<th>i</th>
<th>$m_{i1}$</th>
<th>$m_{i1}^N$</th>
<th>$\sigma_{i1}^2$</th>
<th>$\sigma_{i1}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-18.324</td>
<td>0.201</td>
<td>174.685</td>
<td>0.195</td>
<td>41</td>
<td>-0.268</td>
<td>-0.048</td>
<td>0.299</td>
<td>0.147</td>
</tr>
<tr>
<td>2</td>
<td>-0.916</td>
<td>0.205</td>
<td>2.677</td>
<td>0.155</td>
<td>42</td>
<td>-0.372</td>
<td>-0.008</td>
<td>0.312</td>
<td>0.146</td>
</tr>
<tr>
<td>3</td>
<td>0.694</td>
<td>0.227</td>
<td>0.954</td>
<td>0.122</td>
<td>43</td>
<td>-0.153</td>
<td>0.043</td>
<td>0.304</td>
<td>0.141</td>
</tr>
<tr>
<td>4</td>
<td>1.231</td>
<td>0.291</td>
<td>0.845</td>
<td>0.108</td>
<td>44</td>
<td>0.001</td>
<td>0.072</td>
<td>0.289</td>
<td>0.135</td>
</tr>
<tr>
<td>5</td>
<td>1.629</td>
<td>0.360</td>
<td>0.852</td>
<td>0.101</td>
<td>45</td>
<td>-0.062</td>
<td>0.085</td>
<td>0.252</td>
<td>0.129</td>
</tr>
<tr>
<td>6</td>
<td>1.900</td>
<td>0.386</td>
<td>0.809</td>
<td>0.097</td>
<td>46</td>
<td>-0.138</td>
<td>0.113</td>
<td>0.239</td>
<td>0.128</td>
</tr>
<tr>
<td>7</td>
<td>1.470</td>
<td>0.395</td>
<td>0.590</td>
<td>0.097</td>
<td>47</td>
<td>0.003</td>
<td>0.163</td>
<td>0.245</td>
<td>0.130</td>
</tr>
<tr>
<td>8</td>
<td>1.223</td>
<td>0.417</td>
<td>0.452</td>
<td>0.095</td>
<td>48</td>
<td>0.288</td>
<td>0.195</td>
<td>0.231</td>
<td>0.131</td>
</tr>
<tr>
<td>9</td>
<td>1.487</td>
<td>0.438</td>
<td>0.480</td>
<td>0.093</td>
<td>49</td>
<td>0.307</td>
<td>0.185</td>
<td>0.236</td>
<td>0.135</td>
</tr>
<tr>
<td>10</td>
<td>1.578</td>
<td>0.428</td>
<td>0.477</td>
<td>0.090</td>
<td>50</td>
<td>0.182</td>
<td>0.170</td>
<td>0.254</td>
<td>0.139</td>
</tr>
<tr>
<td>11</td>
<td>1.686</td>
<td>0.432</td>
<td>0.492</td>
<td>0.091</td>
<td>51</td>
<td>-0.053</td>
<td>0.171</td>
<td>0.256</td>
<td>0.142</td>
</tr>
<tr>
<td>12</td>
<td>1.475</td>
<td>0.456</td>
<td>0.410</td>
<td>0.095</td>
<td>52</td>
<td>-0.185</td>
<td>0.209</td>
<td>0.279</td>
<td>0.144</td>
</tr>
<tr>
<td>13</td>
<td>1.403</td>
<td>0.476</td>
<td>0.352</td>
<td>0.096</td>
<td>53</td>
<td>0.074</td>
<td>0.268</td>
<td>0.294</td>
<td>0.143</td>
</tr>
<tr>
<td>14</td>
<td>1.523</td>
<td>0.478</td>
<td>0.338</td>
<td>0.093</td>
<td>54</td>
<td>0.258</td>
<td>0.297</td>
<td>0.293</td>
<td>0.139</td>
</tr>
<tr>
<td>15</td>
<td>1.345</td>
<td>0.464</td>
<td>0.300</td>
<td>0.092</td>
<td>55</td>
<td>0.430</td>
<td>0.303</td>
<td>0.284</td>
<td>0.134</td>
</tr>
<tr>
<td>16</td>
<td>1.163</td>
<td>0.470</td>
<td>0.266</td>
<td>0.090</td>
<td>56</td>
<td>0.419</td>
<td>0.287</td>
<td>0.258</td>
<td>0.129</td>
</tr>
<tr>
<td>17</td>
<td>1.125</td>
<td>0.507</td>
<td>0.244</td>
<td>0.090</td>
<td>57</td>
<td>0.492</td>
<td>0.267</td>
<td>0.250</td>
<td>0.126</td>
</tr>
<tr>
<td>18</td>
<td>1.251</td>
<td>0.590</td>
<td>0.269</td>
<td>0.095</td>
<td>58</td>
<td>0.658</td>
<td>0.232</td>
<td>0.263</td>
<td>0.124</td>
</tr>
<tr>
<td>19</td>
<td>1.165</td>
<td>0.670</td>
<td>0.249</td>
<td>0.100</td>
<td>59</td>
<td>0.820</td>
<td>0.167</td>
<td>0.257</td>
<td>0.118</td>
</tr>
<tr>
<td>20</td>
<td>1.178</td>
<td>0.767</td>
<td>0.247</td>
<td>0.103</td>
<td>60</td>
<td>1.170</td>
<td>0.066</td>
<td>0.233</td>
<td>0.112</td>
</tr>
<tr>
<td>21</td>
<td>1.529</td>
<td>0.861</td>
<td>0.266</td>
<td>0.104</td>
<td>61</td>
<td>0.570</td>
<td>-0.124</td>
<td>0.170</td>
<td>0.107</td>
</tr>
<tr>
<td>22</td>
<td>1.735</td>
<td>0.898</td>
<td>0.259</td>
<td>0.101</td>
<td>62</td>
<td>0.209</td>
<td>-0.288</td>
<td>0.194</td>
<td>0.115</td>
</tr>
<tr>
<td>23</td>
<td>1.930</td>
<td>0.906</td>
<td>0.256</td>
<td>0.098</td>
<td>63</td>
<td>-0.073</td>
<td>-0.390</td>
<td>0.203</td>
<td>0.119</td>
</tr>
<tr>
<td>24</td>
<td>2.208</td>
<td>0.882</td>
<td>0.260</td>
<td>0.096</td>
<td>64</td>
<td>-0.344</td>
<td>-0.452</td>
<td>0.209</td>
<td>0.123</td>
</tr>
<tr>
<td>25</td>
<td>2.206</td>
<td>0.797</td>
<td>0.239</td>
<td>0.091</td>
<td>65</td>
<td>-0.432</td>
<td>-0.472</td>
<td>0.227</td>
<td>0.128</td>
</tr>
<tr>
<td>26</td>
<td>1.933</td>
<td>0.684</td>
<td>0.223</td>
<td>0.090</td>
<td>66</td>
<td>-0.226</td>
<td>-0.479</td>
<td>0.243</td>
<td>0.130</td>
</tr>
<tr>
<td>27</td>
<td>1.677</td>
<td>0.598</td>
<td>0.211</td>
<td>0.090</td>
<td>67</td>
<td>-0.041</td>
<td>-0.521</td>
<td>0.268</td>
<td>0.130</td>
</tr>
<tr>
<td>28</td>
<td>1.786</td>
<td>0.551</td>
<td>0.208</td>
<td>0.093</td>
<td>68</td>
<td>-0.023</td>
<td>-0.592</td>
<td>0.239</td>
<td>0.126</td>
</tr>
<tr>
<td>29</td>
<td>1.660</td>
<td>0.458</td>
<td>0.209</td>
<td>0.096</td>
<td>69</td>
<td>-0.267</td>
<td>-0.687</td>
<td>0.201</td>
<td>0.125</td>
</tr>
<tr>
<td>30</td>
<td>1.596</td>
<td>0.369</td>
<td>0.214</td>
<td>0.099</td>
<td>70</td>
<td>-0.477</td>
<td>-0.771</td>
<td>0.218</td>
<td>0.132</td>
</tr>
<tr>
<td>31</td>
<td>1.564</td>
<td>0.272</td>
<td>0.219</td>
<td>0.102</td>
<td>71</td>
<td>-0.523</td>
<td>-0.825</td>
<td>0.248</td>
<td>0.138</td>
</tr>
<tr>
<td>32</td>
<td>1.017</td>
<td>0.154</td>
<td>0.206</td>
<td>0.103</td>
<td>72</td>
<td>-0.714</td>
<td>-0.874</td>
<td>0.204</td>
<td>0.199</td>
</tr>
<tr>
<td>33</td>
<td>0.892</td>
<td>0.118</td>
<td>0.216</td>
<td>0.107</td>
<td>73</td>
<td>-0.831</td>
<td>-0.906</td>
<td>0.201</td>
<td>0.151</td>
</tr>
<tr>
<td>34</td>
<td>1.153</td>
<td>0.098</td>
<td>0.215</td>
<td>0.111</td>
<td>74</td>
<td>-0.883</td>
<td>-0.919</td>
<td>0.236</td>
<td>0.170</td>
</tr>
<tr>
<td>35</td>
<td>1.021</td>
<td>0.017</td>
<td>0.203</td>
<td>0.115</td>
<td>75</td>
<td>-0.751</td>
<td>-0.926</td>
<td>0.238</td>
<td>0.186</td>
</tr>
<tr>
<td>36</td>
<td>0.747</td>
<td>-0.048</td>
<td>0.222</td>
<td>0.126</td>
<td>76</td>
<td>-0.684</td>
<td>-0.955</td>
<td>0.239</td>
<td>0.206</td>
</tr>
<tr>
<td>37</td>
<td>0.597</td>
<td>-0.074</td>
<td>0.267</td>
<td>0.136</td>
<td>77</td>
<td>-0.840</td>
<td>-1.000</td>
<td>0.257</td>
<td>0.235</td>
</tr>
<tr>
<td>38</td>
<td>0.457</td>
<td>-0.084</td>
<td>0.317</td>
<td>0.142</td>
<td>78</td>
<td>-0.983</td>
<td>-1.025</td>
<td>0.277</td>
<td>0.268</td>
</tr>
<tr>
<td>39</td>
<td>0.187</td>
<td>-0.095</td>
<td>0.322</td>
<td>0.144</td>
<td>79</td>
<td>-1.031</td>
<td>-1.031</td>
<td>0.305</td>
<td>0.305</td>
</tr>
<tr>
<td>40</td>
<td>-0.107</td>
<td>-0.082</td>
<td>0.313</td>
<td>0.146</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: On-line ($m_{i1}, \sigma_{i1}^2$) and smoothed ($m_{i1}^N, \sigma_{i1}^2$) posterior means and variances of $\beta_{i1}$, $i = 1, \ldots, N$ (auxiliary particle filter method).
Figure 2.2: Estimated posterior variance trajectories for the treatment difference effect: Gamerman (1991) method vs. auxiliary particle filter method.

ginning, resulting from the large variance of the non-informative prior, settling down after a few time intervals have been processed, when the value of the flat prior rapidly diminishes as data is received, and then, as time passes, having a downward trend, which represents an increase in information about the parameter due to additional observations taken into account.
2.1.2 The Lognormal Model

Let us now assume that the survival times are lognormally distributed with parameters $\mu$ and $\sigma^2$ (i.e. $T \sim LN(\mu, \sigma^2)$). This means that $\log(T) \sim N(\mu, \sigma^2)$. For comparison, we will use the same time grid as in the PE model. Let us model the mean of $\log(T)$ as $\mu_j(t) = \beta^*_j z_j$, for $t \in I_i$, $i = 1, \ldots, N$, and $j = 1, \ldots, n$. Then,

$$\log(T^*_j) = \beta^*_j z_j + v^*_j, \quad v^*_i \sim N(0, \sigma^2),$$

$$\beta_i = \beta_{i-1} + w_i, \quad w_i \sim N(0, W_i),$$

where

$$W_i = \begin{pmatrix} \sigma^2_{10} & 0 \\ 0 & \sigma^2_{11} \end{pmatrix} = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.2 \end{pmatrix}, \quad \text{for all } i = 1, \ldots, N,$$

with $\beta^*_i = (\beta_{10}, \beta_{11})$, and $z^*_j = (1, z_j)$. Similarly to the PE model, assume $(\beta_0|D_0) \sim N(0, 10^3 I_2)$.

Analysis

For this model, in order to find estimates of the $\beta^*_i$ parameters, we will use the auxiliary particle filter method only. However, in order to carry out the procedure, we will need an estimate of the unknown parameter $\sigma^2$. For this purpose, let us notice that the model above looks very much like a regular dynamic linear model with unknown observational variance. Thus, a $G(\frac{n_0}{2}, \frac{\nu_0 S_0}{2})$ distribution is used for the initial information about the precision $\phi = \sigma^{-2}$, $(\phi|D_0)$, with $n_0 = S_0 = 1$. As the data
comes in successively, the parameters of this distribution are updated according to the updating equations of section 1.2.1 for the dynamic linear model with unknown observational variance, and then, the posterior estimate of $\sigma^2$ at time $t_i$, which is equal to $1/S_i$, is used in the auxiliary particle filter estimation of the parameters of the posterior distribution ($\beta_i|D_i$). The quantities necessary to obtain the estimates of $n_i$ and $S_i$, $i = 1, \ldots, N$, are computed based on the generated sample from the posterior ($\beta_{i-1}|D_{i-1}$). After all the data was processed the posterior estimate of $\phi (\sigma^2)$ at $t_{N-1}$ turned out to be 2.679 (0.373) with standard deviation 0.426 (0.063). This number was later used to find the estimates of the mean and covariance matrix of the smoothing distributions ($\beta_i|D_N$), for $i = 1, \ldots, N$.

Monte Carlo Backward Smoothing

Künsch (2003) (sec. 3.4) presents a recursive algorithm to generate samples from the conditional distribution ($\beta_{1:N}|D_N$), where $\beta_{1:N} = (\beta_1, \ldots, \beta_N)$. At time $t_N$, we use the particle filter sample of size $M$ and draw an index $j$ according to the appropriate weights from the particle filter, and set $\beta_{N,j}^N = \beta_{N,j}$. We then proceed backward in time, generating a value of $\beta_{i,j}^N$, $i = N-1, \ldots, 1$, from the density proportional to:

$$g(\beta_{i+1,j}^N|\beta_{i,j}^N)L(Y_{i:N} | \beta_{i,j}^N)\frac{1}{M}\sum_{n=1}^{M} g(\beta_{i,j}^N | \beta_{i-1,n}),$$

(2.1)

where $g(\beta_{i+1,j}^N|\beta_{i,j}^N) = N(\beta_{i+1,j}^N, W_{i+1})$, evaluated at $\beta_{i+1,j}^N$; $g(\beta_{i,j}^N | \beta_{i-1,n}) = N(\beta_{i-1,n}, W_i)$, and $L(Y_{i:N} | \beta_{i,j}^N) = \prod_{k=1}^{M} f(y_{ik} | \beta_{i,j}^N)^{\delta_{ik}} S(y_{ik} | \beta_{i,j}^N)^{1-\delta_{ik}}$ is the likelihood of the data over
the risk set at time $t_i$ with $f(\cdot\mid\cdot)$ and $S(\cdot\mid\cdot)$ being the density and survival function for the survival times, respectively. We can then repeat the above procedure to obtain a sample from $(\mu_{1,N}\mid D_N)$ of the desired size. Various MCMC sampling techniques may be used to generate the sample. Here, the Metropolis-Hastings algorithm is applied. One disadvantage of this method is that if $M$ is large, say $10^4$ or larger, it gets very expensive computationally. Therefore, we modify the procedure slightly. Instead of taking all $M$ elements in $\frac{1}{M} \sum_{k=1}^{M} g(\mu_{i,j}\mid \mu_{i-1,k})$ mixture, for each $i = N-1, \ldots, 1$ we randomly draw $K$ of them, $K << M$, according to the weights $\pi_{i-1} = (\pi_{i-1,1}, \ldots, \pi_{i-1,M})$ obtained in the particle filter stage. In other words we draw a sample from a density proportional to

$$g(\mu_{i+1,j}\mid \mu_{i,j}) L(Y_i\mid \mu_{i,j}) \sum_{k=1}^{K} g(\mu_{i,j}\mid \mu_{i-1,k}) \pi_{i-1,k},$$

(2.2)

The results based on samples of sizes 5000 (with $K = 100$) and 1000 (with $K = 500, 1000$), with $M = 10000$, drawn from (2.2), were compared with the results from a relatively small sample of a 100 drawn from (2.1). The simulations showed that employing the modification does not alter the results significantly, while the speed of the program is greatly increased.

The posterior on-line and smoothed estimates for $\beta_{0i}$ and $\beta_{1i}$, $i = 1, \ldots, N$, across all time intervals are shown in Table 2.3. Also, to be able to compare the lognormal model with the PE model, the predictive survival curves are computed for both of the models and plotted in Figure 2.3. The smoothed estimates of the parameters were used to generate these curves, allowing us to predict the survival time for a
new individual given the treatment group he/she was assigned to, and after the data available on all the individuals have been observed. Following Gamerman (1991) (sec. 5), the predictive survival is computed by factorizing the survival function, where the factors are conditioned on all the previously collected data $D_N$. Namely, with $X$ being the survival time of a patient, the survival function for that patient is

$$S(x|D_N) = S(x|X > t_{i-1}, D_N) \prod_{j=1}^{i-1} S(t_j|X > t_{j-1}, D_N), \quad x \in I_i, \quad i = 1, \ldots, N.$$ 

The biggest differences in the performance of these two models are visible for the chemotherapy alone treatment group ($z=0$). We can see that the PE model greatly underestimates the survival in early stages, while the lognormal model follows the Kaplan-Meier estimate more closely. This in fact is the case for both treatment
groups along the entire time scale, but for the chemotherapy and radiation treatment group (z=1) the differences are not as apparent. This, however, may be due to the fact that the PE model assumed the evolution variance 0 for $\beta_{10}$, while the lognormal model allowed $\beta_{10}$ to change over time. Further investigation would be necessary in order to see the implications of this assumption. The only aspect regarding the lognormal model estimates that we could be concerned about is the overestimation of the survival of the patients for times from 100 to about 300 days, which is visible on the plot in a form of the little hump for this particular time interval. It is not obvious to us why this takes place.
<table>
<thead>
<tr>
<th>i</th>
<th>$m_{i0}$</th>
<th>$m_{i1}$</th>
<th>$m_{i2}$</th>
<th>$m_{i3}$</th>
<th>i</th>
<th>$m_{i0}$</th>
<th>$m_{i1}$</th>
<th>$m_{i2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.600</td>
<td>3.776</td>
<td>14.748</td>
<td>0.603</td>
<td>41</td>
<td>5.142</td>
<td>5.636</td>
<td>1.665</td>
</tr>
<tr>
<td>2</td>
<td>4.217</td>
<td>3.904</td>
<td>0.724</td>
<td>0.288</td>
<td>42</td>
<td>5.154</td>
<td>5.802</td>
<td>1.947</td>
</tr>
<tr>
<td>3</td>
<td>5.816</td>
<td>5.211</td>
<td>-0.339</td>
<td>0.032</td>
<td>43</td>
<td>4.964</td>
<td>5.968</td>
<td>1.096</td>
</tr>
<tr>
<td>4</td>
<td>5.435</td>
<td>5.728</td>
<td>-0.746</td>
<td>-0.944</td>
<td>44</td>
<td>5.228</td>
<td>6.205</td>
<td>0.953</td>
</tr>
<tr>
<td>5</td>
<td>6.111</td>
<td>5.761</td>
<td>-1.215</td>
<td>-0.929</td>
<td>45</td>
<td>6.563</td>
<td>6.442</td>
<td>0.203</td>
</tr>
<tr>
<td>6</td>
<td>7.131</td>
<td>5.906</td>
<td>-1.672</td>
<td>-0.735</td>
<td>46</td>
<td>6.588</td>
<td>6.410</td>
<td>0.076</td>
</tr>
<tr>
<td>7</td>
<td>5.541</td>
<td>5.864</td>
<td>-0.725</td>
<td>-0.674</td>
<td>47</td>
<td>6.135</td>
<td>6.208</td>
<td>-0.192</td>
</tr>
<tr>
<td>8</td>
<td>6.379</td>
<td>6.036</td>
<td>-0.782</td>
<td>-0.835</td>
<td>48</td>
<td>6.771</td>
<td>6.187</td>
<td>-0.784</td>
</tr>
<tr>
<td>9</td>
<td>5.993</td>
<td>6.139</td>
<td>-1.119</td>
<td>-0.987</td>
<td>49</td>
<td>6.298</td>
<td>6.089</td>
<td>-0.593</td>
</tr>
<tr>
<td>10</td>
<td>7.216</td>
<td>6.237</td>
<td>-1.338</td>
<td>-0.734</td>
<td>50</td>
<td>5.952</td>
<td>6.010</td>
<td>-0.058</td>
</tr>
<tr>
<td>11</td>
<td>7.417</td>
<td>6.182</td>
<td>-1.781</td>
<td>-0.739</td>
<td>51</td>
<td>6.327</td>
<td>6.015</td>
<td>0.616</td>
</tr>
<tr>
<td>12</td>
<td>5.666</td>
<td>6.027</td>
<td>-0.885</td>
<td>-0.799</td>
<td>52</td>
<td>5.879</td>
<td>5.828</td>
<td>0.882</td>
</tr>
<tr>
<td>13</td>
<td>5.739</td>
<td>6.137</td>
<td>-0.947</td>
<td>-0.976</td>
<td>53</td>
<td>5.309</td>
<td>5.705</td>
<td>0.345</td>
</tr>
<tr>
<td>14</td>
<td>7.315</td>
<td>6.400</td>
<td>-1.515</td>
<td>-0.823</td>
<td>54</td>
<td>5.261</td>
<td>5.779</td>
<td>0.305</td>
</tr>
<tr>
<td>15</td>
<td>5.961</td>
<td>6.414</td>
<td>-0.811</td>
<td>-0.738</td>
<td>55</td>
<td>5.310</td>
<td>5.898</td>
<td>0.064</td>
</tr>
<tr>
<td>16</td>
<td>7.061</td>
<td>6.629</td>
<td>-0.917</td>
<td>-0.816</td>
<td>56</td>
<td>5.933</td>
<td>6.013</td>
<td>0.196</td>
</tr>
<tr>
<td>17</td>
<td>7.721</td>
<td>6.797</td>
<td>-1.366</td>
<td>-0.851</td>
<td>57</td>
<td>5.864</td>
<td>6.016</td>
<td>-0.057</td>
</tr>
<tr>
<td>18</td>
<td>7.140</td>
<td>6.587</td>
<td>-1.611</td>
<td>-0.994</td>
<td>58</td>
<td>5.369</td>
<td>6.011</td>
<td>-0.080</td>
</tr>
<tr>
<td>19</td>
<td>6.558</td>
<td>6.413</td>
<td>-0.873</td>
<td>-0.898</td>
<td>59</td>
<td>5.281</td>
<td>6.139</td>
<td>-0.287</td>
</tr>
<tr>
<td>20</td>
<td>5.933</td>
<td>6.371</td>
<td>-0.779</td>
<td>-1.289</td>
<td>60</td>
<td>5.448</td>
<td>6.369</td>
<td>-0.881</td>
</tr>
<tr>
<td>21</td>
<td>5.739</td>
<td>6.392</td>
<td>-1.195</td>
<td>-1.401</td>
<td>61</td>
<td>7.153</td>
<td>6.678</td>
<td>-0.142</td>
</tr>
<tr>
<td>22</td>
<td>6.865</td>
<td>6.498</td>
<td>-1.744</td>
<td>-1.375</td>
<td>62</td>
<td>6.552</td>
<td>6.699</td>
<td>0.409</td>
</tr>
<tr>
<td>23</td>
<td>6.934</td>
<td>6.463</td>
<td>-2.177</td>
<td>-1.485</td>
<td>63</td>
<td>6.864</td>
<td>6.581</td>
<td>0.898</td>
</tr>
<tr>
<td>24</td>
<td>6.868</td>
<td>6.429</td>
<td>-2.582</td>
<td>-1.402</td>
<td>64</td>
<td>6.715</td>
<td>6.473</td>
<td>1.174</td>
</tr>
<tr>
<td>26</td>
<td>6.533</td>
<td>6.493</td>
<td>-1.468</td>
<td>-0.816</td>
<td>66</td>
<td>5.631</td>
<td>6.188</td>
<td>0.669</td>
</tr>
<tr>
<td>27</td>
<td>7.186</td>
<td>6.515</td>
<td>-1.255</td>
<td>-0.715</td>
<td>67</td>
<td>5.246</td>
<td>6.316</td>
<td>0.547</td>
</tr>
<tr>
<td>28</td>
<td>6.873</td>
<td>6.383</td>
<td>-1.645</td>
<td>-0.727</td>
<td>68</td>
<td>6.132</td>
<td>6.583</td>
<td>0.377</td>
</tr>
<tr>
<td>29</td>
<td>6.553</td>
<td>6.295</td>
<td>-0.984</td>
<td>-0.390</td>
<td>69</td>
<td>7.097</td>
<td>6.835</td>
<td>0.380</td>
</tr>
<tr>
<td>30</td>
<td>6.357</td>
<td>6.241</td>
<td>-0.919</td>
<td>-0.309</td>
<td>70</td>
<td>6.678</td>
<td>6.763</td>
<td>0.749</td>
</tr>
<tr>
<td>31</td>
<td>6.106</td>
<td>6.227</td>
<td>-0.661</td>
<td>-0.018</td>
<td>71</td>
<td>6.106</td>
<td>6.725</td>
<td>0.947</td>
</tr>
<tr>
<td>32</td>
<td>6.517</td>
<td>6.335</td>
<td>0.450</td>
<td>0.245</td>
<td>72</td>
<td>7.320</td>
<td>6.881</td>
<td>0.466</td>
</tr>
<tr>
<td>33</td>
<td>6.108</td>
<td>6.253</td>
<td>0.146</td>
<td>0.089</td>
<td>73</td>
<td>7.145</td>
<td>6.820</td>
<td>0.293</td>
</tr>
<tr>
<td>34</td>
<td>6.252</td>
<td>6.154</td>
<td>-0.472</td>
<td>0.121</td>
<td>74</td>
<td>6.604</td>
<td>6.567</td>
<td>0.559</td>
</tr>
<tr>
<td>35</td>
<td>6.763</td>
<td>6.241</td>
<td>-0.171</td>
<td>0.528</td>
<td>75</td>
<td>6.680</td>
<td>6.538</td>
<td>0.277</td>
</tr>
<tr>
<td>36</td>
<td>6.449</td>
<td>6.063</td>
<td>0.378</td>
<td>0.829</td>
<td>76</td>
<td>6.711</td>
<td>6.570</td>
<td>0.268</td>
</tr>
<tr>
<td>37</td>
<td>5.797</td>
<td>5.782</td>
<td>0.600</td>
<td>0.930</td>
<td>77</td>
<td>6.747</td>
<td>6.586</td>
<td>0.824</td>
</tr>
<tr>
<td>38</td>
<td>5.223</td>
<td>5.706</td>
<td>0.711</td>
<td>1.021</td>
<td>78</td>
<td>6.612</td>
<td>6.601</td>
<td>1.259</td>
</tr>
<tr>
<td>40</td>
<td>5.453</td>
<td>5.675</td>
<td>1.459</td>
<td>1.042</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: On-line ($m_{i0}$, $m_{i1}$) and smoothed ($m_{i0}^{N}$, $m_{i1}^{N}$) posterior means and of $\beta_{i0}$ and $\beta_{i1}$, $i = 1, \ldots, N$ (auxiliary particle filter method).
2.2 The Sales Data Analysis

To illustrate the dynamic linear model for longitudinal data, we reproduce the results of example 3.4.2 in West and Harrison (1997) using the data given in Table 3.3 of the text.

Data

The data consists of 42 measurements (taken over the years 1975–1985) of two time-dependent variables: a response variable $Y_t$, representing quarterly sales of a certain company (given in standardized units), and a time-dependent covariate $F_t$, representing the corresponding total market sales. $Y_t$ ranges from 44 to 80.2, with mean 60.9 and standard deviation 9.65, versus $F_t$ ranging from 105.5 to 183, with mean 138.51 and standard deviation 21.63.

Model

The main goal of the analysis is to investigate the pattern of change of the relationship between $Y_t$ and $F_t$ over time. West and Harrison (1997) point out that if we plotted each of the series against time, a clear annual seasonal pattern would be visible. Plotting both of the series against each other, however, removes this pattern (Figure 2.4), suggesting a straight line relationship between the two variables with
Figure 2.4: Company sales ($Y_t$) vs. total market sales ($F_t$).

intercept 0. Hence, the following linear dynamic model with unknown, but constant observational variance $V$, can be fitted to the data (West and Harrison (1997), sec. 3.3.1):

**Observation equation:** $Y_t = F_t \theta_t + \nu_t, \quad \nu_t \sim N(0, V)$

**System equation:** $\theta_t = \theta_{t-1} + \omega_t, \quad \omega_t \sim T_{n_t-1}(0, W_t)$

**Initial information:** $(\theta_0|D_0) \sim T_1(m_0, C_0),$

$(\phi|D_0) \sim G(\frac{n_t-1}{2}, \frac{n_t-1}{2})$, where $\phi = \frac{1}{\tau}$ is an observational precision, and $\theta_t$ is the expected proportion of the company’s sales in total market sales in quarter $t$. Based on the knowledge of the market prior to 1975, $m_0$ and $C_0$ are set to 0.45 and 0.0025, respectively. Also, a relatively non–informative prior is chosen for $\phi$ with $n_0 = S_0 = 1$. 
Figure 2.5: One-step ahead forecast errors: $\delta = 0.6$ (left), $\delta = 1$ (right).

**Analysis**

Let $\delta$ be the discount factor defined in section 1.2.1. Following West and Harrison (1997), two types of analysis are considered: one, with $\delta = 0.6$, that allows $\theta_t$ to vary significantly over time, and second, with $\delta = 1$, corresponding to the static regression model with $\theta_t = \theta_0$ for all $t = 1, \ldots, 42$. The results of both analyses for the last 20 quarters can be found in Table 2.4, with $m^5_t$ denoting smoothed estimates of $\theta_t$ values.

Our results agreed with the original findings of West and Harrison (1997) (sec. 3.4.2) which can be summarized as follows. From Figure 2.5, where the one-step ahead forecast errors $e_t$ are plotted against time, we can see that the model with
\( \delta = 0.6 \) underforecasts the response variable \( Y_t \) at later times, as reflected by the positive values of \( \epsilon_t \) at those times. The same pattern is visible for the corresponding plot of \( \epsilon_t \)'s for \( \delta = 1 \). Here, however, not only most of the errors are positive, but their values tend to grow with time, demonstrating deepening inadequacy of the model as time progresses. This poor performance of the model results from the fact that the model adapts to the data much more slowly than the competing, highly adaptive model with \( \delta = 0.6 \), which produces much smaller forecast errors. Figure 2.6, which shows the on-line estimates \( m_t \) with 90\% HPD intervals from both models, visually compares the rates at which the models adapt to the data as we move forward in time. West and Harrison (1997) note that although neither of the two models can anticipate the increase in the values of \( \theta_t \) with time, the model with \( \delta = 0.6 \), due to the small value of the discount factor, adapts quite well to new data, producing much more accurate one-step ahead forecasts than the model with \( \delta = 1 \). Thus, having learned how poorly the static model fits the data, we can now see how greatly the plot in Figure 2.4 may be misleading, indicating such a model. The authors also note that to see the pattern of change of \( \theta_t \) over time and the necessity of a dynamic model, it is enough to plot the empirical estimates of \( \theta_t \) values (given by \( Y_t/F_t \)) versus time. This plot, together with on-line and smoothed estimates \( m_t \) and \( m_t^S \), respectively, is displayed in Figure 2.7, with the \( Y_t/F_t \) ratios appearing as dots. It shows that the model with \( \delta = 0.6 \) captures the changes in the parameter trajectory quite well. To smoothed estimates were found via FFBS algorithm (sec. 1.3.3).
Figure 2.6: On-line estimated trajectories of $\theta_t$ with 90% HPD.

Figure 2.7: Ratio $Y_t/F_t$ vs. time, with on-line and smoothed estimates.
<table>
<thead>
<tr>
<th>( t )</th>
<th>( F_t )</th>
<th>( Y_t )</th>
<th>( f_t )</th>
<th>( Q_t^{1/2} )</th>
<th>( S_t^{1/2} )</th>
<th>( m_t )</th>
<th>( m_t^2 )</th>
<th>( C_t^{1/2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>112.2</td>
<td>49.8</td>
<td>48.69</td>
<td>1.004</td>
<td>0.854</td>
<td>0.437</td>
<td>0.440</td>
<td>0.0041</td>
</tr>
<tr>
<td>24</td>
<td>149.5</td>
<td>66.6</td>
<td>65.29</td>
<td>1.158</td>
<td>0.859</td>
<td>0.441</td>
<td>0.443</td>
<td>0.0039</td>
</tr>
<tr>
<td>25</td>
<td>156.9</td>
<td>70.0</td>
<td>68.15</td>
<td>1.164</td>
<td>0.851</td>
<td>0.443</td>
<td>0.444</td>
<td>0.0037</td>
</tr>
<tr>
<td>26</td>
<td>123.4</td>
<td>54.8</td>
<td>54.69</td>
<td>1.032</td>
<td>0.835</td>
<td>0.443</td>
<td>0.444</td>
<td>0.0038</td>
</tr>
<tr>
<td>27</td>
<td>108.8</td>
<td>48.7</td>
<td>48.25</td>
<td>0.993</td>
<td>0.823</td>
<td>0.445</td>
<td>0.445</td>
<td>0.0041</td>
</tr>
<tr>
<td>28</td>
<td>153.3</td>
<td>67.7</td>
<td>68.17</td>
<td>1.155</td>
<td>0.811</td>
<td>0.443</td>
<td>0.444</td>
<td>0.0037</td>
</tr>
<tr>
<td>29</td>
<td>158.3</td>
<td>70.4</td>
<td>70.16</td>
<td>1.111</td>
<td>0.798</td>
<td>0.444</td>
<td>0.445</td>
<td>0.0034</td>
</tr>
<tr>
<td>30</td>
<td>119.5</td>
<td>52.7</td>
<td>53.05</td>
<td>0.959</td>
<td>0.787</td>
<td>0.443</td>
<td>0.446</td>
<td>0.0037</td>
</tr>
<tr>
<td>31</td>
<td>107.7</td>
<td>49.1</td>
<td>47.71</td>
<td>0.937</td>
<td>0.802</td>
<td>0.447</td>
<td>0.448</td>
<td>0.0040</td>
</tr>
<tr>
<td>32</td>
<td>145.0</td>
<td>64.8</td>
<td>64.79</td>
<td>1.101</td>
<td>0.790</td>
<td>0.447</td>
<td>0.449</td>
<td>0.0037</td>
</tr>
<tr>
<td>33</td>
<td>155.3</td>
<td>70.0</td>
<td>69.39</td>
<td>1.088</td>
<td>0.781</td>
<td>0.449</td>
<td>0.450</td>
<td>0.0035</td>
</tr>
<tr>
<td>34</td>
<td>123.1</td>
<td>55.3</td>
<td>55.23</td>
<td>0.956</td>
<td>0.770</td>
<td>0.449</td>
<td>0.452</td>
<td>0.0036</td>
</tr>
<tr>
<td>35</td>
<td>109.2</td>
<td>50.1</td>
<td>49.02</td>
<td>0.923</td>
<td>0.774</td>
<td>0.452</td>
<td>0.453</td>
<td>0.0039</td>
</tr>
<tr>
<td>36</td>
<td>144.8</td>
<td>65.6</td>
<td>65.43</td>
<td>1.064</td>
<td>0.764</td>
<td>0.452</td>
<td>0.454</td>
<td>0.0036</td>
</tr>
<tr>
<td>37</td>
<td>160.6</td>
<td>72.7</td>
<td>72.66</td>
<td>1.071</td>
<td>0.754</td>
<td>0.453</td>
<td>0.456</td>
<td>0.0033</td>
</tr>
<tr>
<td>38</td>
<td>119.1</td>
<td>55.2</td>
<td>53.90</td>
<td>0.908</td>
<td>0.764</td>
<td>0.456</td>
<td>0.458</td>
<td>0.0036</td>
</tr>
<tr>
<td>39</td>
<td>109.5</td>
<td>51.5</td>
<td>49.93</td>
<td>0.916</td>
<td>0.783</td>
<td>0.460</td>
<td>0.459</td>
<td>0.0039</td>
</tr>
<tr>
<td>40</td>
<td>144.8</td>
<td>66.2</td>
<td>66.65</td>
<td>1.075</td>
<td>0.775</td>
<td>0.459</td>
<td>0.458</td>
<td>0.0037</td>
</tr>
<tr>
<td>41</td>
<td>165.8</td>
<td>75.5</td>
<td>76.08</td>
<td>1.103</td>
<td>0.768</td>
<td>0.457</td>
<td>0.458</td>
<td>0.0033</td>
</tr>
<tr>
<td>42</td>
<td>127.4</td>
<td>58.5</td>
<td>58.23</td>
<td>0.940</td>
<td>0.760</td>
<td>0.458</td>
<td>0.458</td>
<td>0.0034</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( t )</th>
<th>( \delta = 1 )</th>
<th>( F_t )</th>
<th>( Y_t )</th>
<th>( f_t )</th>
<th>( Q_t^{1/2} )</th>
<th>( S_t^{1/2} )</th>
<th>( m_t )</th>
<th>( m_t^2 )</th>
<th>( C_t^{1/2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>112.2</td>
<td>49.8</td>
<td>48.31</td>
<td>0.977</td>
<td>0.990</td>
<td>0.431</td>
<td>–</td>
<td>0.0014</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>149.5</td>
<td>66.6</td>
<td>64.42</td>
<td>1.014</td>
<td>1.060</td>
<td>0.432</td>
<td>–</td>
<td>0.0015</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>156.9</td>
<td>70.0</td>
<td>67.71</td>
<td>1.086</td>
<td>1.128</td>
<td>0.432</td>
<td>–</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>123.4</td>
<td>54.8</td>
<td>53.34</td>
<td>1.144</td>
<td>1.141</td>
<td>0.433</td>
<td>–</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>108.8</td>
<td>48.7</td>
<td>47.07</td>
<td>1.153</td>
<td>1.161</td>
<td>0.433</td>
<td>–</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>153.3</td>
<td>67.7</td>
<td>66.37</td>
<td>1.186</td>
<td>1.166</td>
<td>0.433</td>
<td>–</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>158.3</td>
<td>70.4</td>
<td>68.59</td>
<td>1.192</td>
<td>1.191</td>
<td>0.434</td>
<td>–</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>119.5</td>
<td>52.7</td>
<td>51.84</td>
<td>1.205</td>
<td>1.182</td>
<td>0.434</td>
<td>–</td>
<td>0.0015</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>107.7</td>
<td>49.1</td>
<td>46.74</td>
<td>1.193</td>
<td>1.235</td>
<td>0.434</td>
<td>–</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>145.0</td>
<td>64.8</td>
<td>62.98</td>
<td>1.255</td>
<td>1.255</td>
<td>0.435</td>
<td>–</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>155.3</td>
<td>70.0</td>
<td>67.52</td>
<td>1.278</td>
<td>1.305</td>
<td>0.435</td>
<td>–</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>123.1</td>
<td>55.3</td>
<td>53.59</td>
<td>1.320</td>
<td>1.317</td>
<td>0.436</td>
<td>–</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>109.2</td>
<td>50.1</td>
<td>47.58</td>
<td>1.329</td>
<td>1.364</td>
<td>0.436</td>
<td>–</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>144.8</td>
<td>65.6</td>
<td>63.14</td>
<td>1.385</td>
<td>1.403</td>
<td>0.437</td>
<td>–</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>160.6</td>
<td>72.7</td>
<td>70.11</td>
<td>1.429</td>
<td>1.445</td>
<td>0.437</td>
<td>–</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>119.1</td>
<td>55.2</td>
<td>52.06</td>
<td>1.459</td>
<td>1.510</td>
<td>0.438</td>
<td>–</td>
<td>0.0017</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>109.5</td>
<td>51.5</td>
<td>47.92</td>
<td>1.522</td>
<td>1.593</td>
<td>0.438</td>
<td>–</td>
<td>0.0018</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>144.8</td>
<td>66.2</td>
<td>63.45</td>
<td>1.615</td>
<td>1.630</td>
<td>0.439</td>
<td>–</td>
<td>0.0018</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>165.8</td>
<td>75.5</td>
<td>72.73</td>
<td>1.659</td>
<td>1.664</td>
<td>0.439</td>
<td>–</td>
<td>0.0019</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>127.4</td>
<td>58.5</td>
<td>55.96</td>
<td>1.681</td>
<td>1.689</td>
<td>0.440</td>
<td>–</td>
<td>0.0019</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.4: Sales data example results.
As a further comparison of the two competing models, let us take a closer look at the results contained in Table 2.4. The same pattern visible on the plots is quite apparent here as well, with the values of $m_t$ gradually approaching 0.46, for the dynamic model, and slowly converging to 0.44, for the static model. The uncertainty associated with these values is significantly greater in the more adaptive model due to a very small value of the discount factor. Finally, let us note that the model with $\delta = 1$ greatly overestimates the observational variance $V$, represented by the estimates $S_t$. This is because the model wrongly interprets the variability associated with the evolution changes in $\theta_t$ as the measurement error of the data. This in turn causes the one-step forecast variances $Q_t$ to be much smaller in the model with $\delta = 0.6$, compared to the static model with $\delta = 1$. (West and Harrison (1997), sec. 3.4.2)
Chapter 3

A Dynamic Model for Survival Data with Longitudinal Covariates

3.1 Introduction

Two fully Bayesian approaches for modeling survival data with longitudinal covariates are developed. The first one, which from now on will be referred to as the 2-stage model, is conceptually similar to the ideas presented in Tsiatis et al. (1995). In the first step, estimates of the patient-specific trajectories that underlie the observed repeated measurement profiles are obtained via dynamic hierarchical models. These estimates are then used as a time-dependent covariate in the Cox PH model, to find the conditional posterior distribution (given the just estimated longitudinal covariate profiles of the patients) of the corresponding survival model parameters. This postc-
rior distribution is obtained by combining the Cox's Partial Likelihood function with the prior distributions for the parameters. The second approach, referred to as the joint model, uses the same model for the longitudinal covariate as the 2-stage approach described above, but models the survival data in a way similar to Garmann (1991). The estimates of the model parameters are obtained via sampling from their joint posterior distribution, which combines the complete joint likelihood of the longitudinal and survival data with prior knowledge of the parameters. The two models are evaluated and compared to each other and to one of currently available methods in a simulation study. The details of both models are presented in the sections below.

3.2 General Setup, Notation and Assumptions

Let us consider a group of $n$ patients followed over an interval $[0, t_N]$, and time discretized over a grid $\tau = \{t_0, \ldots, t_N\}$, with $t_0 = 0$. Denote by $X_i$ the survival time for the $i^{th}$ individual. As in most cases with survival data, the data is subject to right censoring. Therefore, we observe $T_i = \min(X_i, C_i)$ (where $T_i$ and $C_i$ correspond to the patient's true, possibly unobserved, failure time and potential censoring time, respectively), and the failure indicator $\delta_i$, equal to 1 if failure was observed (i.e. $X_i \leq C_i$) and 0 otherwise. Each patient also provides $n_i$ measurements of a longitudinal covariate taken at times $\{t_{ij} : t_{ij} \leq T_i\}$, where $j = 1, \ldots, n_i$. These measurements are $Y_i = \{Y_{ij} : t_{ij} \leq T_i\}$. This means that there are no measurements available after the failure occurs. The grid $\tau$ is evenly spaced and chosen is such a way that it includes
all the measurement times $t_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_i$, as well as all the event times of the patients. Although the model requires complete knowledge of the true longitudinal covariate history for all patients, it allows for missing data. Through its hierarchical structure, which allows for borrowing strength across patients, the missing observations can be easily imputed and updated at each iteration of the posterior distribution sampling algorithm. This also means that the baseline longitudinal measurements at time $t_0$ may, but do not have to be available. Here we assume that $Y_0 = \{Y_{i0}, i = 1, \ldots, n\}$ is not observed.

The following assumptions are used throughout:

- The timing of the measurements for each individual is non-informative. That is, the measurements are scheduled independently of patient's observed history.

- The censoring of the event times is non-informative, i.e., the censoring and failure times are independent.

- The longitudinal profiles across individuals are conditionally independent, given the parameters.

- If there are ties in the data, the failure is assumed to occur before the censored observation, and both are included in the risk set at a given time.

- The observational and evolution variances in the dynamic model are assumed constant over time. The model, however, can be easily extended to a more general case of time varying variances.
bullet The longitudinal measurements do not have to be taken at the same times for all individuals nor do they have to be taken at even time increments for each patient. The time grid \( \tau \) used in the analysis, however, needs to be evenly spaced and, as mentioned earlier, needs to be fine enough to include all the measurement times of the longitudinal covariate as well as all the event/censoring times of the patients. This means, that for \( i = 1, \ldots, n \), \( T_i = t_j \) for some \( j = 1, \ldots, N \). This will usually lead to the necessity of imputing the "missing" observations at times when measurements for a given patient were simply not taken. If it is reasonable to do so, however, and especially if there are very few longitudinal covariate measurements taken at the observed event (or censoring) times, we can assume discrete, or grouped, survival data. That is, all we know about the event times of the patients is that they occurred in a certain time interval and we set these times to the beginning of that interval.

bullet For simplicity, it is assumed that if patient \( i \) is lost to follow up, the time at which this patient's last covariate measurement was taken is used as the patient's censoring time. In other words, we have no information about the patient's failure after \( t_{in_i} \). Because of the chosen grid, this also implies that \( T_i = t_{in_i} = t_j \), for some \( j = 1, \ldots, N \).

bullet The hazard of the \( i^{th} \) individual at time \( t_j \) depends only on the underlying true value of the longitudinal covariate of that individual at time \( t_j \), not his/her entire history up to time \( t_j \).
Other more implicit assumptions will become apparent as we introduce the models.

3.3 The 2-stage Model

Let us now take a look at the 2 stage approach for modeling survival data with longitudinal covariates.

3.3.1 Longitudinal Data Model

First, let us define the model for the longitudinal part of the data. With \( \theta_i = \{\theta_{ij}, j = 1, \ldots, n_i\} \) (where \( \theta_{ij} = \theta_i(t_{ij}) \)) representing the true, unobserved longitudinal covariate profile for the \( i^{th} \) patient, we adapt the following hierarchical dynamic linear growth curve model for the longitudinal data:

- The observation equation:

\[
Y_{ij} = \theta_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2_\epsilon),
\]  

(3.1)

- The structural equation:

\[
\theta_{ij} = \theta_{i,j-1} + \gamma_{j-1} + \omega_{ij}, \quad \omega_{ij} \sim N(0, \sigma^2_\omega),
\]  

(3.2)

- The evolution equation:

\[
\gamma_j = \gamma_{j-1} + \nu_j, \quad \nu_j \sim N(0, \sigma^2_\nu), \quad j = 1, \ldots, n_i - 1,
\]  

(3.3)
• Priors:

\[ \theta_{i0} \sim N(m_\theta, c_\theta^2), \quad \gamma_0 \sim N(m_\gamma, c_\gamma^2), \]

\[ \sigma_\epsilon^{-2} \sim G(n_\epsilon, S_\epsilon), \quad \sigma_\omega^{-2} \sim G(n_\omega, S_\omega), \quad \sigma_\nu^{-2} \sim G(n_\nu, S_\nu), \]

• Hyperpriors:

\[ m_\theta \sim N(\mu_\theta, \xi_\theta^2), \quad m_\gamma \sim N(\mu_\gamma, \xi_\gamma^2), \quad c_\theta^{-2} \sim G(n_c, S_c). \]

where all the error terms \( \epsilon_{ij}, \omega_{ij}, \text{ and } \nu_{ij} \) are assumed to be mutually and internally independent, and the variances \( \sigma_\epsilon^2, \sigma_\omega^2, \sigma_\nu^2, \text{ and } c_\theta^2 \) are assumed unknown.

The observation equation describes the effect of the measurement error, while equations (3.2) and (3.3) provide a description of the stochastic evolution of the underlying longitudinal covariate process over time. Namely, the value of the covariate at time \( t_j \) for the \( i^{th} \) patient depends on its previous value (at time \( t_{j-1} \)) plus an increment \( \gamma_{j-1} \), that itself is allowed to evolve over time according to the evolution equation (hrcr, simple random walk). Such model allows not only for a local linear behavior in the longitudinal process, but will also track other shapes, such as quadratic or higher level polynomials. In other words, the model does not impose any particular shape on the patient’s longitudinal profile, but lets the shape be determined by the data. Also, through its hierarchical structure it allows for borrowing strength across the patients and estimating both the patient specific and overall longitudinal covariate trajectories.
Since the increments $\gamma_j, j = 1, \ldots, N$, are not indexed by $i$, the model implicitly assumes that any changes that occur in the longitudinal process between two times $t_j$ and $t_k$ are the same for all the patients. This may seem as a restrictive assumption, and we will discuss its implications in greater detail later in the simulation study results section. There is something that is gained, however, by implementing the model in this form. Contrary to the simple random walk approach, for example, if a certain trend is observed in the longitudinal profile history up to time $t$, the model will allow for this trend to be continued past time $t$. This is especially useful if any missing values need to be imputed or a projection into the future needs to be made. Also, together with parameters $\sigma^2_v, \sigma^2_{\theta}, \sigma^2_{\rho}, m_\theta, c^2_{\theta}$, which are common for all the individuals under study, $\gamma_j$'s provide a link between all the patients allowing for both estimation of the population longitudinal trajectory and the prediction of the marker's profile for a new patient.

Including the hyperpriors for the prior moments of the distributions of $m_\theta$ and $m_\gamma$ allows for prediction and makes the model less sensitive to the prior assumptions about the initial levels $\theta_{i0}$, and the initial increment $\gamma_0$. Also, putting a hyperprior on $c^2_{\gamma}$ allows the data to decide how to distribute the weight between $\mu_0$ and the data while estimating $m_\theta$. The more precise the data is at time $t_1$ (or $t_0$, if the measurements are taken at baseline), the more weight the data will be given.
3.3.1.1 The Likelihood and the Joint Prior

Let $\phi_1 = \{\Theta, \sigma^2_\epsilon\}$, $\phi_2 = \{\Gamma, \Theta_0, \sigma^{-2}_\omega\}$, and $\phi_3 = \{m_\theta, m_\gamma, c^{-2}_\theta, c^{-2}_\nu\}$ where $\Theta = \{\theta_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_i\}$, $\Theta_0 = \{\theta_{i0}, i = 1, \ldots, n\}$, and $\Gamma = \{\gamma_j; j = 0, 1, \ldots, N - 1\}$ (for simplicity, the parameters we do not wish to estimate from the data were omitted above). If $D = \{Y, T\}$, where $Y$ is the observed longitudinal data for $n$ patients, and $T$ are the survival times of all the individuals under study, then the observed longitudinal data likelihood is:

$$L(\phi_1 | D) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} N(Y_{ij} | \theta_{ij}, \sigma^2_\epsilon). \quad (3.6)$$

Now, we have the following hierarchical "prior" structure:

$$\pi(\phi_1 | \phi_2) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} N(\theta_{ij} | \theta_{i,j-1} - \gamma_{j-1}, \sigma^{-2}_\omega) \ G(\sigma^{-2}_\epsilon | n_x, S_x), \quad (3.7)$$

$$\pi(\phi_2 | \phi_3) = \prod_{i=1}^{n} \left( N(\theta_{i0} | m_\theta, c^{-2}_\theta) \prod_{j=1}^{n_i} N(\gamma_j | \gamma_{j-1}, \sigma^2_\nu) \right) N(\gamma_0 | m_\gamma, c^{-2}_\gamma) \ G(\sigma^{-2}_\omega | n_\omega, S_\omega), \quad (3.8)$$

$$\pi(\phi_3) = N(m_\theta | \mu_\theta, \xi^2_\theta) \ N(m_\gamma | \mu_\gamma, \xi^2_\gamma) \ G(\sigma^{-2}_\nu | n_\nu, S_\nu) \ G(c^{-2}_\omega | n_c, S_c). \quad (3.9)$$

Now, combining (3.7), (3.8) and (3.9), we get the joint prior for $\Phi = (\phi_1, \phi_2, \phi_3)$:

$$\pi(\Phi) = \pi(\phi_1, \phi_2, \phi_3) \quad = \pi(\phi_1 | \phi_2, \phi_3) \ \pi(\phi_2 | \phi_3) \ \pi(\phi_3) \quad (3.10)$$

$$= \pi(\phi_1 | \phi_2) \ \pi(\phi_2 | \phi_3) \ \pi(\phi_3).$$

The last equality holds because $\phi_3$ affects $\phi_1$ only through $\phi_2$, and therefore $\phi_1$ and $\phi_3$ are conditionally independent given $\phi_2$. Let us also note that $\pi(\phi_3)$ itself depends
on some hyperparameters.

3.3.1.2 The Posterior Distribution of $\Phi$

The joint posterior distribution of the parameters $\Phi$ is proportional to the product of the prior (3.10) and the likelihood (3.6). That is:

$$
\pi(\Phi|D) \propto L(\phi_1|D) \pi(\Phi).
$$

(3.11)

Let us note that only $\phi_1$ is present in $L(\phi_1|D)$. This is because the sampling distribution of the data depends only on $\phi_1$, and the parameters $\phi_3$ affect the data only through $\phi_2$, which in turn affects the data only through $\phi_1$.

3.3.2 Survival Data Model

3.3.2.1 The Model and Likelihood

The traditional Cox PH model with a time-dependent covariate is used to model the survival data. The hazard at time $t_j$ of the $i^{th}$ patient is expressed as

$$
\lambda(t_j|\theta_i) = \lambda_0(t_j) \exp(\beta \theta_{ij}),
$$

(3.12)

where $\theta_i = \{\theta_{ij}, j = 1, \ldots, n_i\}$. Once the estimates of the patient specific longitudinal profiles $\theta_i, i = 1, \ldots, n_i$, are obtained in the first step of the analysis, they are used to find the conditional posterior distribution of the survival parameters (given $\Theta$), that combines the prior information about these parameters with Cox’s partial likelihood
given by:

$$PL(\beta|D, \phi_1) = \prod_{i=1}^{d} \frac{\exp(\theta_{iT_i(0)}\beta)}{\sum_{k \in \mathcal{R}_i} \exp(\theta_{kT_i(0)}\beta)},$$

(3.13)

where $d$ is the number of distinct event times among $n$ patients (i.e., $n - d$ is the number of right-censored observations), $T_{(i)}$ is the $i^{th}$ ordered survival time, $i = 1, \ldots, d$, $\mathcal{R}_i$ is the risk set at time $T_{(j)}$, and $\theta_{kT_{(i)}}$ is the true unobserved longitudinal covariate value of the $k^{th}$ patient in $\mathcal{R}_i$ at time $T_{(i)}$. Assuming this model for survival data, the baseline hazard can be left unspecified since it does not enter into the partial likelihood function.

Let us emphasize here that rather than using the “raw”, observed longitudinal covariate values as the time-dependent covariate in the Cox PH model, we use the measurement–error–adjusted, unobserved, true longitudinal profile values estimated at all time grid points up to each patient’s survival/censoring time. This is due to the fact that the longitudinal data is often measured with substantial amount of measurement error and including the observed longitudinal history in the model may lead to inaccurate estimation of regression parameters that describe the relationship between hazard and true value of the covariate, as indicated by Prentice (1982).

### 3.3.2.2 The Conditional Posterior Distribution of $\beta$ given $\phi_1$

Since our model is fully Bayesian, we first need to specify a prior for the parameter $\beta$. We assume:

$$\pi(\beta) = N(\mu_\beta, \sigma^2_\beta).$$

(3.14)
Now, following Ibrahim et al. (2001) (problem 7.2), we combine this prior with partial likelihood (3.13) to obtain the posterior distribution for $\beta$ conditional on the parameters $\phi_1$:

$$
\pi(\beta|D, \phi_1) \propto PL(\beta|D, \phi_1) \pi(\beta). \tag{3.15}
$$

### 3.3.3 The Joint Posterior Distribution

Having all the pieces ready, we can write out the joint posterior distribution of all the model parameters. By the law of conditional probability we get:

$$
\pi(\beta, \Phi|D) = \pi(\beta|D, \phi_1) \pi(\Phi|D). \tag{3.16}
$$

where the two components on the right side of the equation are given by (3.11) and (3.15), respectively.

Again, similarly to the case of the likelihood for the longitudinal data, let us note that there is no $\phi_2$ nor $\phi_3$ on the conditioning side of $\pi(\beta|D, \phi_1)$. This is because partial likelihood (3.13), and what follows, the conditional posterior distribution $\pi(\beta|D, \phi_1)$, depends on $\phi_3$ only through $\phi_2$, and on $\phi_2$ only through $\phi_1$.

### 3.3.3.1 Sampling from the Joint Posterior Distribution

Ibrahim et al. (2001) (problem 7.2) suggest the following 2 step procedure to sample from the posterior distribution $\pi(\beta, \Phi|D)$:
Step 1 Obtain $M$ samples $\{\Phi^m, m = 1, \ldots, M\}$, from

$$
\pi(\Phi|D) \propto L(\phi_1|D) \pi(\Phi),
$$

where $\Phi^m = (\phi_1^m, \phi_2^m, \phi_3^m)$.

Step 2 For each $m = 1, \ldots, M$, obtain $K$ samples $\{\beta_k, k = 1, \ldots, K\}$, from

$$
\pi(\beta|D, \phi_i^m) \propto PL(\beta|D, \phi_i^m) \pi(\beta)
$$

using the sample $\phi_i^m$ generated in Step 1 (to obtain $\phi_i^m$ from $\Phi^m$ simply discard $\phi_2^m$ and $\phi_3^m$ in the generated sample $\Phi^m$).

The sampling in Step 1 can be easily carried out via Gibbs sampling algorithm (section 1.3.1) and is fairly straightforward since each of the full conditional distributions of the parameters required for Gibbs sampling is either normal or gamma. This is because both normal and gamma are conjugate for normal, and the likelihood component of this distribution involves only normal distributions, while the prior distributions for the parameters are either normal or gamma.

Sampling from $\pi(\beta|D, \phi_i^m)$ in Step 2 is a little bit more complicated, due to the fact that the distribution is not available in closed form. A rejection sampling method, such as Metropolis-Hastings algorithm (section 1.3.2) needs to be used.

More details regarding the form of the full conditionals used in Step 1, are included in section A.1.

Having performed this 2-step procedure we have $MK$ samples of $\beta$ and $M$ samples of $\Phi$ from the joint posterior distribution of all the model parameters $\pi(\beta, \Phi|D)$. 
Based on this sample we can draw statistical conclusions about the parameters, and in turn answer questions of interest about the phenomena and the population under study.

3.3.4 Remarks

The advantage of this model is that it is fairly easy to implement, and through the use of the Cox PH model and partial likelihood, we do not need to worry about estimating the baseline hazard, which often is not of primary interest. One could argue, however, that the model does not make the most efficient use of the data since the estimates of the longitudinal model parameters \( \Phi \) are not adjusted for the censoring of the survival times of the patients. Also, while the variance estimates of the survival model parameters will reflect the uncertainty in the estimates of the longitudinal model parameters, the opposite is not true. Moreover, because of the use of partial likelihood, the adopted time grid does not have to include the censoring times of the patients. This may be advantageous in situations when there are very few patients with their longitudinal profile measurements taken at the censoring times and choosing a very fine grid may be either unreasonable or simply not possible.

The following section presents the second approach, the joint model, which addresses some of the issues mentioned above.
3.4 The Joint Model

Contrary to the 2 stage approach, the joint model combines the survival and longitudinal likelihoods into the joint complete likelihood of the data, which in turn combined with the prior knowledge of the investigator, yields the joint posterior distribution of all the model parameters. Let us now take a look at the details of this joint modeling approach.

3.4.1 Longitudinal Data Model

The longitudinal part of the model is exactly the same as for the 2 stage method, and it will not be repeated here. For details, please, refer to section 3.3.1.

3.4.2 Survival Data Model

The survival model adopted here follows the general ideas presented by Gamerman (1991). The piecewise exponential distribution is assumed for the survival times of the patients, i.e. the hazard function is a step function over a prespecified time grid. Let $\lambda^{(i)}(t|\theta_i)$ be the hazard for patient $i$ at time $t$. Note, that $\lambda^{(i)}(t|\theta_i)$ depends on the observed longitudinal profile $Y_i$ of the $i^{th}$ patient only through $\theta_i$, and therefore $Y_i$ is not present on the conditioning side of $\lambda^{(i)}(t|\theta_i)$. To further simplify the notation, let us from now write the hazard as $\lambda^{(i)}(t)$, remembering its dependence on $\theta_i$. $\lambda^{(i)}(t)$ may depend on other covariates as well, if one desires to include them in the model.

Using the same notation as thus far, the model can be summarized as follows:
• The observation equation:

\[ T_i \sim PE(\lambda^{(i)}(t), \tau), \quad \text{where} \quad \lambda^{(i)} = (\lambda^{(i)}_1, \lambda^{(i)}_2, \ldots, \lambda^{(i)}_N), \]  

\[ \lambda^{(i)}(t) = \lambda^{(i)}_j, \quad \text{for} \quad t \in I_j = (t_{j-1}, t_j], \quad j = 1, \ldots, N; \]  

\[ \lambda^{(i)}_j = \exp(\alpha_j + \beta \theta_{ij}), \quad j = 1, \ldots, N, \]  

• The guide relation:

\[ \lambda^{(i)}_j = \exp(\alpha_j + \beta \theta_{ij}), \quad j = 1, \ldots, N, \]  

• The evolution equation:

\[ \alpha_j = \alpha_{j-1} + \eta_j, \quad \eta_j \sim N(0, \sigma^2_\eta), \quad j = 1, \ldots, N, \]  

• Priors:

\[ \alpha_0 \sim N(m_\alpha, \sigma^2_\alpha), \quad \sigma^{-2}_\eta \sim G(n_\eta, S_\eta) \quad \beta \sim N(m_\beta, \sigma^2_\beta), \]  

The parameters \( \alpha_j \) represent the \( \log(\lambda_{0j}) \), i.e. the logarithm of baseline hazard in the \( j^{th} \) interval, and are common for all the patients. They are allowed to vary between time intervals of the defined grid \( \tau \). \( \lambda_{0j} \) can be interpreted as the prior mean of the baseline hazard \( \lambda_{01} \) in \( I_1 = (t_0, t_1) \).

### 3.4.3 The Joint Likelihood

With the model above and \( \Lambda = \{\lambda^{(i)}, i = 1, \ldots, n\}, \psi = \{A, \beta, \sigma^2_\eta\} \), where \( A = \{\alpha_j, j = 0, 1, \ldots, N\} \), the survival data likelihood is:

\[ L(\Lambda | D) = \prod_{i=1}^{n} \left( \frac{1}{T_i} \lambda^{(i)}(T_i) \right)^{\delta_{iT_i}} \exp \left( -\int_0^{T_i} \lambda^{(i)}(t) \, dt \right), \]
where \( \delta_{ij} = \delta_i \) if \( j = n_i \) and 0 otherwise.

Because of the PE distribution for the survival times, the integral in the likelihood above becomes a sum, and we have the following:

\[
L(\Lambda|D) = \prod_{i=1}^{n} \left( \lambda_{T_i}^{(i)} \right)^{\delta_{T_i}} \exp \left( -\sum_{j=1}^{T_i} \lambda_j^{(i)} \Delta t \right),
\]

where \( \Delta t = t_j - t_{j-1} \). \( \Delta t \) does not depend on \( j \) because of the evenly spaced time grid. Expressing (3.21) in terms of \( \psi \), we obtain:

\[
L(\psi|D) = \prod_{i=1}^{n} \left( e^{\alpha_{T_i} + \beta_{T_i} \tau_i} \right)^{\delta_{T_i}} \exp \left( -\sum_{j=1}^{T_i} e^{\alpha_j + \beta_j \theta_{ij}} \Delta t \right)
= \prod_{j=1}^{N} \prod_{i \in R_j} \left( e^{\alpha_j + \beta_j \theta_{ij}} \right)^{\delta_{ij}} \exp \left( -e^{\alpha_j + \beta_j \theta_{ij}} \Delta t \right).
\]

(3.21)

Now, combining the above survival data likelihood with the likelihood for the longitudinal data (3.6) we obtain the complete joint data likelihood (recall that \( T_i = t_{in_i} \)):

\[
L(\phi_1, \psi|D) = \prod_{i=1}^{n} \left( e^{\alpha_{T_i} + \beta_{T_i} \tau_i} \right)^{\delta_{T_i}} \prod_{j=1}^{T_i} N(Y_{ij}|\theta_{ij}, \sigma_y^2) \exp \left( -e^{\alpha_j + \beta_j \theta_{ij}} \Delta t \right).
\]

(3.22)

### 3.4.4 The Joint Posterior Distribution

#### 3.4.4.1 The Prior for \( \psi \)

Given the assumed model, the prior for \( \psi \) is

\[
\pi(\psi) = \prod_{j=1}^{N} N(\alpha_j|\alpha_{j-1}, \sigma_\alpha^2) \ N(\alpha_0|m_\alpha, c_\alpha^2) \ N(\beta|m_\beta, c_\beta) \ G(\sigma_y^{-2}|n_y, S_y).
\]

(3.23)
3.4.4.2 The Joint Prior for $(\psi, \Phi)$

Let us recall from section 3.3.1.1 that

$$
\pi(\Phi) = \pi(\phi_1|\phi_2) \pi(\phi_2|\phi_3) \pi(\phi_3).
$$

(3.24)

where the components of this equation are given by (3.7), (3.8) and (3.9). Assuming that the priors for $\psi$ and $\Phi$ are independent, the joint prior distribution of $(\psi, \Phi)$ is

$$
\pi(\psi, \Phi) = \pi(\psi)\pi(\Phi).
$$

(3.25)

3.4.4.3 The Joint Posterior for $(\psi, \Phi)$

Putting together the joint likelihood (3.22) with the prior (3.25) we obtain the joint posterior distribution of all the model parameters:

$$
\pi(\psi, \Phi|D) \propto L(\phi_1, \psi|D)\pi(\psi, \Phi)
$$

(3.26)

3.4.4.4 Sampling from the Joint Posterior

Similarly to the 2-stage model, we can use Gibbs sampling to sample from (3.26). The full conditional distributions of most of the model parameters are either normal or gamma, and therefore are easy to sample from. A few parameters, however, namely $\Theta, A$ and $\beta$, do not have their full conditionals available in closed form and have to be sampled from using a rejection sampling technique, such as Metropolis-Hastings algorithm. All the full conditional distributions for the parameters are included in section A.2.
3.4.5 Remarks

As it was pointed out earlier, the joint model uses the data more efficiently than the 2-stage model. The parameters of the longitudinal model are now adjusted for the censoring of the survival data, and the variance estimates of these parameters reflect the uncertainty in the survival model parameters. The model, through the use of the joint likelihood of the longitudinal and survival data, requires us to obtain an estimate of the longitudinal covariate value at the event/censoring time of each patient. Thus, the chosen grid has to include not only the observed failure times, as it was the case in the 2-stage model, but the censoring times as well.
Chapter 4

Simulation Study

To measure the quality of the newly developed methods, a simulation study was carried out. The results of this study are presented in sections below.

4.1 The Data

100 data sets were generated. Each of them consisted of $n = 50$ simulated patients observed over the course of $N = 24$ time units. These can be thought of as hours, days, weeks, months, etc., but for the purposes of this analysis, let us from now on refer to them as days. For each patient two types of data were generated: a survival (or censoring) time $T_i$, and a set of longitudinal marker measurements $Y_i = \{Y_{ij} : j = 1, \ldots, n_i\}$ taken at times $t_i = 1, \ldots, n_i$, where $n_i = T_i \leq 24, i = 1, \ldots, 50$. Although this may be rare in real-world clinical trials, all the data, besides the measurements at baseline time $t_0 = 0$, are available. Given this type of data generated, the time
grid used in the analysis was \( \tau = (t_0, t_1, \ldots, t_{24}) = (0, 1, \ldots, 24) \). On average, 20% of the simulated patients were allowed to be censored.

### 4.1.1 Generating the Data

Each data set was generated according to the following algorithm:

1. Draw 50 quintuples \((\rho_0, \rho_1, \rho_2, \rho_3, \rho_4)_i, i = 1, \ldots, 50\), from the \( N(M, \Sigma) \) distribution, with:

\[
M = \begin{pmatrix} 100.2, -4, 3, -4 \end{pmatrix},
\]

\[
\Sigma = I_5 \begin{pmatrix} 10, 0.2, 0.2, 0.2, 0.2 \end{pmatrix},
\]

where \( I_5 \) is a 5-dimensional identity matrix.

2. Calculate the true longitudinal covariate value for patient \( i \) at time \( t_j, j = 1, \ldots, 24 \), according to the following piecewise linear function of time:

\[
\theta_i(t) = \rho_{0i} + \rho_{1i} t + \rho_{2i} (t - 5)_+ + \rho_{3i} (t - 15)_+ + \rho_{4i} (t - 18)_+.
\]

where \((t - x)_+ = t - x\) if \( t \geq x \) and 0 otherwise.

3. Generate \( Y_{ij} \) from \( N(\theta_{ij}, \sigma^2) \), with \( \sigma^2 = 20 \).

4. Assuming \( \beta = -0.05 \) and \( \alpha(t) = \log(1.2)t \) for \( t \in (t_{j-1}, t_j] \), generate the survival time for the \( i^{th} \) patient as follows:
A Set \( j = 1 \).

B Calculate cumulative hazard \( \Lambda_i(t_j) = \int_0^{t_j} e^{\alpha(t) + \beta \theta_i(t)} dt \).

C Generate \( u_1 \sim Unif(0, 1) \). If \( u_1 \leq 1 - e^{-\Lambda_i(t_j)} \) set \( T_i = t_j \); otherwise increase \( j \) by one and go back to B.

5. Generate \( u_2 \sim Unif(0, 1) \). If \( u_2 \leq 0.2 \) the observation is censored and \( \delta_i = 0 \). Otherwise \( \delta_i = 1 \). If the observation turns out censored, the generated failure time is used as the censoring time for the patient.

Figure 4.1: Sample generated longitudinal trajectories for 50 patients over 24 days
Thus, the longitudinal data is generated in such a way that each patient's trajectory has a shift in direction at day 5, 15 and 18, and evolves over time according to this patient's specific set of slopes ($\rho_{1i}, \rho_{2i}, \rho_{3i}, \rho_{4i}$). A sample generated longitudinal data set is shown in Figure 4.1.

Across all generated data sets, the median survival was 16 days and 19.4% of observations turned out censored. About 10% of the patients died before day 9, and about 10% of the individuals survived beyond day 22. A Kaplan-Meier estimate of survival for the same data set as shown in Figure 4.1 is displayed in Figure 4.2.

![Figure 4.2](image_url)  
Figure 4.2: Kaplan-Meier estimate of the sample generated survival data
4.2 Computation

4.2.1 Handling the Ties in the Data

Since there were ties in the generated survival data, Breslow approximation to the partial likelihood was used. In our setting, the Breslow likelihood is given by:

\[ BL(\beta|D, \phi_1) = \prod_{i=1}^{d} \frac{\exp\left(\beta \sum_{k=1}^{D_i} \theta_{kT(i)}\right)}{\sum_{m \in \mathcal{R}_i} \exp\left(\beta \theta_{mT(i)}\right)} D_i. \] (4.1)

where \( d \) is the number of distinct even times, \( T(i) \) is the \( i \)th ordered event time, \( \mathcal{R}_i \) is the risk set at time \( T(i) \), and \( D_i \) is the number of individuals failing at time \( T(i) \).

4.2.2 The Priors

Throughout the simulations, non-informative priors were used. Namely,

- \( G(0.001, 0.001) \) for parameters \( \sigma^2, \sigma^2, \sigma^2, \sigma^2, \) and \( \sigma^2 \).
- \( N(0, 10^6) \) for parameters \( m_0, m_1, \beta \).
- \( N(0, 10^4) \) for \( \alpha_0 \).
- \( c_0^2 \) in the prior for \( \gamma_0 \) was set to \( 10^4 \), implicitly giving more weight to the data than to \( \mu_1 \) in estimating \( m_1 \), through \( \gamma_0 \) (see the full conditional distributions for \( \gamma_0 \) and \( m_1 \) in section A.1 for the dependence structure of these parameters).
4.2.3 Simulation

For the 2-stage model we obtained a sample of 30,000 from the posterior distribution $\pi(\Phi|D)$ and discarded the first 10,000 as a burn-in period. Of the remaining 20,000 samples, we used 500 (uniformly distributed across the 20,000 iterations) to sample from $\pi(\beta|D, \phi_1)$. For each of these 500 sample points we generated 1,000 $\beta$'s, obtaining a sample of 500,000 and keeping the last 100,000 to assure convergence. For all 100 data sets, the simulation took about 6 hours on Pentium M 1.60GHz processor, with 504 MB of RAM. For the joint model, for each data set 50,000 sample points from the joint posterior $\pi(\psi, \Phi|D)$ were generated, where the first 20,000 were discarded as the burn-in period. This simulation took about 5 hours for all the data sets.

For the parameters, posterior sample of which was drawn using a rejection sampling method, the acceptance rate was consistently 30-40% in both models.

4.3 Results

4.3.1 Posterior Estimates of $\sigma_\epsilon^2$, $\sigma_\omega^2$, $\sigma_\nu^2$, $\sigma_\eta^2$, $m_\theta$, $m_\gamma$, $c_\theta^2$, and $\beta$

The table below shows some of the characteristics of the posterior distributions of all the model parameters based on all 100 data sets. Posterior means, standard errors and 95% credible sets are included. Recall, that the true values of $\sigma_\epsilon^2$, $m_\theta$ and $\beta$ are 20, 100 and -0.05, respectively.
<table>
<thead>
<tr>
<th></th>
<th>2-stage ($\sigma_i^2 = 20$)</th>
<th>Joint ($\sigma_i^2 = 20$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>sd</td>
</tr>
<tr>
<td>$\sigma_i^2$</td>
<td>17.93</td>
<td>1.25</td>
</tr>
<tr>
<td>$\sigma_i^2$</td>
<td>3.19</td>
<td>0.62</td>
</tr>
<tr>
<td>$\sigma_i^2$</td>
<td>2.74</td>
<td>1.26</td>
</tr>
<tr>
<td>$\sigma_i^2$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$m_\theta$</td>
<td>99.81</td>
<td>1.33</td>
</tr>
<tr>
<td>$m_\gamma$</td>
<td>2.05</td>
<td>0.99</td>
</tr>
<tr>
<td>$c_\delta$</td>
<td>5.58</td>
<td>3.57</td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.04</td>
<td>0.015</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>0.24</td>
<td>2.27</td>
</tr>
</tbody>
</table>

Table 4.1: Simulation study results (2-stage and joint method, $\sigma_i^2 = 20$).

From Table 4.1 we can see that both models perform very well in analyzing the data, and the results from both of them are almost identical. The estimates of the parameters $m_\theta$ and $\beta$ are very close to their true values, being well within one standard error of the truth. The joint model's posterior mean for $\beta$ is equal to the true value of the parameter used to generate the data, although the 95% credible set is slightly wider than for the 2-stage model. The observational variance $\sigma_i^2$ seems to be underestimated and this is consistent for both methods. We believe that this is due to the fact that, as it was mentioned earlier, the changes in the levels of the longitudinal trajectories between two different time points are assumed by the model to be the same for all the patients. As we can see in Figure 4.5, however, the patient specific trajectories are estimated quite accurately and for different patients they evolve differently, which seems to contradict the assumption. Thus, we believe that
to make up for the differences between the longitudinal profiles across individuals, the evolution variance $\sigma^2_\omega$ is overestimated, causing the observational variance $\sigma^2_\tau$ to be underestimated. In other words, some of the variability in the model accounts for the evolution changes in the individual trajectories and gets thrown into $\sigma^2_\omega$ instead of accounting for the measurement error of the observed data reflected by $\sigma^2_\tau$. Although, this may seem as an undesirable property, we pay a small price for the ability to estimate both patient specific and overall trajectories, as well as being able to make predictions for new patients.

### 4.3.2 Posterior Estimates of $\Theta$ and $\Gamma$: Overall and Patient Specific Posterior Mean Trajectories

Let us now take a look at the estimated overall and patient specific trajectories. Figure 4.3 displays the mean posterior longitudinal trajectories across all the data sets from both models with the truth plotted in black and 95% pointwise credible set at all time points of the grid. Both models estimate the true overall trajectory of the longitudinal profile quite well, with a little more deviation from the truth toward the end of the considered time grid. This is due to the fact that by day 18-20, many patients have already experienced the event or were censored and no longer provided measurements of the time-dependent covariate. Due to the smaller amount of data, the estimates at these times are not as precise as the ones at the earlier times when more data were available. Also, let us recall that the value of the survival
Figure 4.3: Overall posterior mean trajectories with 95% credible set: 2-stage and joint method.

Figure 4.4: Posterior mean slopes of the piecewise linear population longitudinal profile over time with 95% credible set: 2-stage and joint method.
regression coefficient $\beta$ used to generate the data was negative. This means that the patients whose longitudinal profiles are at a higher level are more likely to survive for a longer period of time. This is exactly what we observed in the 100 simulated data sets, where the patients who were still alive late in the follow-up period had their longitudinal trajectories higher than the true population average. Thus, the estimates of the overall longitudinal trajectory at times toward the end of the time grid are based on the data that comes from this small group of patients (there were only a few such patients in each data set), whose longitudinal profiles are not very "representative" of the entire population at those times. This causes the population longitudinal trajectory to be overestimated at later times, and in turn, what we will see later, the hazard to be underestimated. Also, let us note that the credible set limits grow wider as we move along the time scale. This is nothing unexpected, since due to the nature of our generated data, the between-patient variability in the longitudinal covariate values increases as we move forward in time.

Figure 4.4 provides a slightly different look at the estimated overall longitudinal trajectory. It plots the posterior estimate (mean) of the slope of the piecewise linear trajectory at each time $t_j$. The plot confirms what we have seen in Figure 4.3. Namely, the truth is estimated quite closely by both models, with more deviations from it toward the end of the follow-up period. This plot shows also smooth transitions of the estimated slopes at the times of their change in value (days 5, 15 and 18). In other words, the neighboring slopes tend to pull toward each other. If we look at the
full conditional distributions for $\gamma_j$, $j = 0, \ldots, N - 1$ in section A.1, we can see that this is nothing surprising, since each $\gamma_j$ explicitly depends on its closest "neighbors", $\gamma_{j-1}$ and $\gamma_{j+1}$ (in case of $\gamma_0$ and $\gamma_{N-1}$ there is only one neighboring value). We can regulate the amount of this kind of smoothing by setting a more informative prior for the evolution variance $\sigma_\gamma^2$, or/and $\sigma_V^2$, according to our prior knowledge or initial beliefs about the nature of the longitudinal covariate evolution over time. This is because both of these variances take part in controlling how much of the variability in the longitudinal profiles will be reflected in the posterior estimates of these profiles. The smaller these variances are, the smoother the estimated trajectories will be, with the extreme case of $\sigma_\gamma^2 = \sigma_V^2 = 0$, in which the estimated overall trajectory will be a straight line with a constant slope.

Patient specific posterior mean trajectories for a couple of simulated patients are shown in Figure 4.5. One patient who experienced the event of interest and one censored individual are presented. For the latter one, the "missing" part of the patient’s longitudinal covariate trajectory was imputed and included in the plot. As expected, due to the hierarchical structure of the model, which allowed for borrowing strength across patients, the imputed trajectory of the patient follows an overall population pattern at the individual’s specific level. This is a very useful feature of the model.
Figure 4.5: Patient specific posterior mean trajectories of the longitudinal profile over time: 2-stage and joint method.

Figure 4.6: Posterior predictive mean trajectories with 95% credible set: 2-stage and joint method.
4.3.3 Posterior Predictive Mean Trajectories

Should we want to predict the covariate trajectory for a new patient, we can sample from the posterior predictive distribution for the new data (please refer to section 1.1.3 for the definition of the posterior predictive distribution). A sample from such a distribution can be generated by using the already available draws from the posterior distribution of the model parameters, and then using these values to draw from the sampling distribution of the data. Based on this generated sample the posterior predictive mean trajectory with 95% credible set were computed and are shown in Figure 4.6.

As previously, both models performed equally well, following a similar pattern as the overall posterior mean trajectory. The 95% credible set limits, however, are wider in the case of the posterior predictive mean trajectory due to added uncertainty that comes from generating the new data from its sampling distribution.

4.3.4 Baseline Hazard and Hazard

Figure 4.7 depicts the posterior means of $\alpha_j$'s at each time point. The estimate seems to slightly underestimate the truth, but it is well within the 95% credible set limits. Since during the simulation of the posterior sample, high correlation between $\alpha_j$'s and $\beta$ was detected, for each $j = 1, \ldots, N$, the hazard

$$\lambda_j^{(m)} = \exp(\alpha_j^{(m)} + \beta^{(m)}g_j^{(P)})$$
Figure 4.7: Posterior means of $\alpha_j$'s over time with 95% credible set: joint model.

Figure 4.8: Estimated vs. "true" hazard: joint model.
was computed and compared to the “truth”:

\[ \lambda_j = \exp(\log(1.2)t_j - 0.05\theta_j^{(P)}) \],

where \( \alpha_j^{(m)} \) and \( \beta^{(m)} \) are the \( m^{th} \) generated posterior samples, \( m = 1, \ldots, M = 30,000 \), of \( \alpha_j \) and \( \beta \), respectively, and \( \theta_j^{(P)} \) is the posterior mean of \( \theta_j \). The result is shown in Figure 4.8. We see a pattern similar to what was observed in the plot of the logarithm of baseline hazard vs. time, with the posterior mean slightly underestimating the truth. This is the effect of the dependence of \( \lambda_j \) on \( \alpha_j \), i.e., if \( \alpha_j \) is underestimated, \( \lambda_j \) will be underestimated as well.

### 4.4 Sensitivity Analysis

#### 4.4.1 Varying the Amount of Noise in the Data

To investigate the effect of the measurement error on the results obtained from our models, 100 new data sets were generated assuming \( \sigma^2 = 5 \). Since both models seem to perform very similarly, only the joint model was fitted to the new data. As previously, 50,000 samples from the joint posterior distribution of the parameters were generated, and only the last 30,000 retained. Based on this sample, the posterior estimates of the parameters were obtained and are reported in Table 4.2. For comparison, the results for \( \sigma^2 = 20 \) were also included in the table. As it was in the case of \( \sigma^2 = 20 \), we observe that our model underestimates the observational variance \( \sigma^2 \). This confirms our earlier suspicion that some of the variability in the data that
comes from the measurement error is "borrowed" by the evolution variance $\sigma_\gamma^2$ to reveal true differences between patients' individual longitudinal trajectories. Recall, that $\gamma_j$ parameters are the same across all the patients causing all the individuals to experience the same changes in their profiles between two time points, which is contrary to what we observed on the plots of the patient specific trajectories of these profiles. With respect to the estimates of other parameters, their posterior means are within the same range of values as for the model with increased variance of the measurement error. The only exceptions are $c_\delta^2$ and $\sigma_\gamma^2$. $c_\delta^2$ for the data with $\sigma_\epsilon^2 = 5$ is much larger than for the corresponding estimate obtained from the data with $\sigma_\epsilon^2 = 20$ and is much closer to the true value of 10 used to generate the data. Also, $\sigma_\gamma^2$ is this time a little smaller than previously. This most likely is due to the fact that the data is more precise and there is less to "borrow" from $\sigma_\epsilon^2$. As to the estimates of the survival parameters, the posterior means of $\beta$ and $\alpha_0$ are virtually the same for both
sets of data, but the standard errors of these parameters for the data generated with \( \sigma^2 = 20 \) are significantly bigger, and, what follows, their corresponding 95% credible sets are wider.

A similar characteristic is visible on all the plots (Figures 4.9, 4.10, 4.11, and 4.12) that further investigate the influence of the measurement error on the results. Namely, the posterior means of all depicted parameter trajectories are very much the same for both sets of the data generated with different values of the measurement error variance, but the 95% credible set limits for the data generated with \( \sigma^2 = 20 \) are slightly wider than for \( \sigma^2 = 5 \). This is nothing unexpected since larger measurement error adds extra uncertainty to the obtained estimates of the model parameters. This is obviously not a downside of our method, since any good model fitted to these two types of data would be expected to “catch” this extra variability, and our approach clearly performed very well reflecting this feature of the data. In Figure 4.10 we can also see that the changes between the estimated value of the slope of the piecewise linear overall longitudinal trajectory at the transition times of 5, 15 and 18 days, occur a little more smoothly when \( \sigma^2 = 20 \). This is because the extra noise in the data caused these changes to be detected by the model a little less accurately than in the case of \( \sigma^2 = 5 \), and pull toward each other even more at the transition points.
Figure 4.9: Overall posterior mean longitudinal trajectories ($\sigma_z^2 = 5$ vs. $\sigma_z^2 = 20$).

Figure 4.10: Posterior means of the slopes of the overall piecewise linear longitudinal trajectory ($\sigma_z^2 = 5$ vs. $\sigma_z^2 = 20$).
Figure 4.11: Posterior predictive mean trajectories ($\sigma^2 = 5$ vs. $\sigma^2 = 20$).

Figure 4.12: Posterior means of $\log($baseline hazard$)$, i.e. $\alpha_j$, $j = 1, \ldots, N$ ($\sigma^2 = 5$ vs. $\sigma^2 = 20$).
4.4.2 Sensitivity to the Prior Moments

The nature of Bayesian methods requires from the investigator to specify the prior distributions for the parameters. This subjective input may have a significant influence on the obtained results. If no prior knowledge about the parameters is available, non-informative priors can be used, which was done in the analyses presented in this chapter. However, the concept of a non-informative prior is somewhat vague, and one prior that is non-informative in the mind of one investigator, may not be portrayed as non-informative enough by another. To address this issue we perform a number of simulations varying the prior moments of some of the parameters. We focus only on the parameters, true values of which we know, that is, $\sigma^2$, $m_\theta$, $\gamma_0$, $\beta$ and $\alpha_0$. Due to the significant computational time required to run the simulation on all 100 data sets we chose the first 10 and averaged the results. In this analysis we chose the data generated with $\sigma_i^2 = 20$. When the prior distribution of an investigated parameter was varied, the rest of the prior distributions were set to the ones specified in section 4.2.2.

4.4.2.1 Varying the Priors for $\beta$, $\alpha_0$, $\gamma_0$, $m_\theta$ and $\sigma_i^2$

The values of the posterior estimates for most of the parameters did not differ across different priors used for some parameters, and therefore were not reported in Table 4.3.
<table>
<thead>
<tr>
<th>$m_\beta = 0, \xi_\beta^2 = 10^6$</th>
<th>$m_\beta = 10, \xi_\beta^2 = 10^4$</th>
<th>$m_\beta = 100, \xi_\beta^2 = 10^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>sd</td>
<td>mean</td>
</tr>
<tr>
<td>$\sigma_\epsilon^2$</td>
<td>17.77</td>
<td>1.14</td>
</tr>
<tr>
<td>$m_\theta$</td>
<td>100.21</td>
<td>1.40</td>
</tr>
<tr>
<td>$m_\gamma$</td>
<td>2.17</td>
<td>0.77</td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.043</td>
<td>0.02</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>-0.41</td>
<td>2.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$m_\alpha = 0, \xi_\alpha^2 = 10^4$</th>
<th>$m_\alpha = 10, \xi_\alpha^2 = 10^4$</th>
<th>$m_\alpha = 100, \xi_\alpha^2 = 10^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>sd</td>
<td>mean</td>
</tr>
<tr>
<td>$\sigma_\epsilon^2$</td>
<td>17.77</td>
<td>1.14</td>
</tr>
<tr>
<td>$m_\theta$</td>
<td>100.21</td>
<td>1.40</td>
</tr>
<tr>
<td>$m_\gamma$</td>
<td>2.17</td>
<td>0.77</td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.043</td>
<td>0.02</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>-0.41</td>
<td>2.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mu_\epsilon = 0, \xi_\epsilon^2 = 10^4$, $\xi_\delta^2 = 10^5$</th>
<th>$\mu_\epsilon = 0, \xi_\epsilon^2 = 10^5$, $\xi_\delta^2 = 10^6$</th>
<th>$\mu_\epsilon = 1000, \xi_\epsilon^2 = 10^5$, $\xi_\delta^2 = 10^6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>sd</td>
<td>mean</td>
</tr>
<tr>
<td>$\sigma_\epsilon^2$</td>
<td>17.77</td>
<td>1.14</td>
</tr>
<tr>
<td>$m_\theta$</td>
<td>100.21</td>
<td>1.40</td>
</tr>
<tr>
<td>$m_\gamma$</td>
<td>2.17</td>
<td>0.77</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>1.96</td>
<td>0.83</td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.043</td>
<td>0.02</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>-0.41</td>
<td>2.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mu_\alpha = 0, \xi_\epsilon^2 = 10^6$</th>
<th>$\mu_\alpha = -100, \xi_\epsilon^2 = 10^4$</th>
<th>$\mu_\alpha = 500, \xi_\epsilon^2 = 10^6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>sd</td>
<td>mean</td>
</tr>
<tr>
<td>$\sigma_\epsilon^2$</td>
<td>17.77</td>
<td>1.14</td>
</tr>
<tr>
<td>$\sigma_\delta^2$</td>
<td>3.12</td>
<td>0.65</td>
</tr>
<tr>
<td>$\sigma_\gamma^2$</td>
<td>2.34</td>
<td>0.88</td>
</tr>
<tr>
<td>$m_\theta$</td>
<td>100.21</td>
<td>1.40</td>
</tr>
<tr>
<td>$m_\gamma$</td>
<td>2.17</td>
<td>0.77</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>1.96</td>
<td>0.83</td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.043</td>
<td>0.02</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>-0.41</td>
<td>2.10</td>
</tr>
</tbody>
</table>

Table 4.3: Sensitivity analysis results (based on 10 data sets, $\xi_\epsilon^2 = 20$).
Varying the prior moments of $\beta$: Three different values for the prior mean of $\beta$ were used: 0, 10 and 100. Lack of knowledge about $\beta$ was assumed in all cases, as reflected by a large prior variance of $10^6$ (for $m_\beta = 0$), and $10^4$ (for $m_\beta = 10$ or $m_\beta = 100$). The results in Table 4.3 show that the posterior estimates of the parameters across different values of prior moments of $N(m_\beta, c_\beta^2)$ were virtually the same, proving that as long as a large prior variance was used, the prior mean did not influence the results.

Varying the prior moments of $\alpha_0$: Three values for the prior mean of $\alpha_0$ were used: 0, 10, and 100, with constant prior variance of $10^4$. Similarly to the case of $\beta$ the results were the same for different priors used, with the estimates of $\beta$ and $\alpha_0$ being slightly pulled up when the prior mean of 100 was used. Let us note that the value of 100 is far from the true value of $\alpha_0 = 0.18$, and thus we have a reason to suspect that further increase in the prior mean of $\alpha_0$ would cause this trend to continue. Increasing the prior variance to even a larger value at the same time, however, would decrease the effect of this prior on the resulting estimates, and this is recommended if we have no prior knowledge about the possible range of values of $\alpha_0$.

Varying the prior variance of $\gamma_0$ and prior moments of $m_\gamma$: With each triplet representing $(c_\gamma^2, \mu_\gamma, \xi_\gamma^2)$, three different settings were used: $(10^4, 0, 10^6)$, $(10^6, 0, 10^6)$, and $(10^6, 100, 10^6)$. Comparing the results from the first two settings, we can see that in the second case, when the prior variance for $\gamma_0$ was
equal to the prior variance for \( m_\gamma \) (i.e. \( \gamma_0 \) was given the same weight as \( \mu_\gamma \) in estimating \( m_\gamma \)) the posterior estimate of \( m_\gamma \) was further from its true value (2), but the uncertainty of this estimate increased significantly. The hierarchical structure of the model, however, dealt with the problem very well, since by setting \( c_\gamma^2 \) to a very large value of \( 10^6 \), we "told" the model, that we are not very confident in this estimate of \( m_\gamma \), and the data should play a leading part in determining the value of \( \gamma_0 \). This happened to be the case, and the posterior mean of \( \gamma_0 \) is quite close to its true value. This was one of the reasons why this additional level of hierarchy on \( m_\gamma \) was introduced, to make the method less sensitive to prior assumptions. If there was no prior distribution for \( m_\gamma \), the value of this parameter, as the prior mean for \( \gamma_0 \), would affect the results much more by pulling the value of the estimate of \( \gamma_0 \) up (or down, depending on the value of \( m_\gamma \)), especially if this prior value would be far from the truth. Moving on to the third setting, we observe a similar phenomena. The prior mean of \( m_\gamma \) is very large (100) compared to the truth. Since the variances \( c_\gamma^2 \) and \( \xi_\gamma^2 \) are equal, again the amount of weight given to \( \gamma_0 \) and \( \mu_\gamma \) in estimating \( m_\gamma \) is the same, and thus the prior is a compromise between the values of these two parameters. We are interested in the estimated value of \( \gamma_0 \), however, and by setting \( c_\gamma^2 \) to \( 10^6 \) we once again do not express much confidence in the estimate of \( m_\gamma \), giving it very little weight in estimating \( \gamma_0 \). \( \gamma_0 \) is thus implicitly determined by the data, through the estimates of the parameters in \( \Theta \).
Varying the prior moments of $m_\theta$: With each pair representing $(\mu_\theta, \xi_\theta^2)$, three different settings were used: $(0, 10^6)$, $(-100, 10^4)$, and $(500, 10^6)$. Changing the prior mean and variance of $m_\theta$ did not significantly change the values of the posterior estimates of the model parameters, although some of the used values were very far away from the truth.

Varying the prior moments of $\sigma^2$: With each pair representing $(n, S)$, three different settings were used: $(0.001, 0.001)$, $(10, 0.1)$, and $(25, 0.5)$. Thus, in the first setting the prior mean and variance of $\sigma^2$ are 1 and 1000, respectively, in the second one 100 and 1000, and in the third 50 and 100. In both cases, the increase of the prior mean for the parameter caused the model to further underestimate its true value. This was most apparent in the third setting when we not only increased the mean from 0 to 50, but also decreased the variance from 1000 to 100. Contrary to expectations this has not caused the estimate to be pulled up toward the prior mean, but underestimated the truth even more with the posterior mean value of 16.01 compared to the value of 17.77 when the prior mean 1 was used.

To summarize, the method proves to be quite robust to changes in the prior assumptions about the parameters. As long as these priors were non-informative, which was reflected by their large variances, the model adapted very well to the data. Instead of prior beliefs (which in some cases were very far from the truth), the method allowed the data to determine the values of the estimates of the parameters.
4.5 Comparison to an Existing Method

4.5.1 The Method

Ibrahim et al. (2001) in chapter 7 suggests a Bayesian version of the model presented by Tsiatis et al. (1995) (for the details regarding the approach taken by Tsiatis et al. (1995), please refer to section 1.4.1). Adjusting the approach to our notation, the following model for the longitudinal covariate is suggested. Given $T_i \geq t$, $i = 1, \ldots, n$:

$$
\theta_i(t) = b_{0i} + b_{1i}t, \tag{4.2}
$$

where

$$
\begin{pmatrix}
  b_{0i} \\
  b_{1i}
\end{pmatrix} \sim N\left(\begin{pmatrix}
  \mu_0 \\
  \mu_1
\end{pmatrix}, \begin{pmatrix}
  \sigma_0^2 & \sigma_{01} \\
  \sigma_{01} & \sigma_1^2
\end{pmatrix}\right) \equiv N(\mu, \Sigma). \tag{4.3}
$$

Also, assuming that $T_i \geq t$ and $Y_i = (Y_i(t_1), Y_i(t_2), \ldots, Y_i(t_{n_i}))$,

$$
Y_i(t_j) = \theta_i(t_j) + \epsilon_{ij}, \tag{4.4}
$$

where $\epsilon_{ij}$'s are iid and $\epsilon_{ij} \sim N(0, \sigma_i^2)$, and $t_1 < t_2 < \ldots < t_{n_i} \leq t$ are the observation times for $Y_i$. The model thus assumes that over the course of time the true covariate history for each patient follows a linear pattern, with patient-specific slope and intercept, and that what we observe are these true values plus a normally distributed measurement error.
For the survival part, the hazard at time $t$ for the $i^{th}$ patient, given the patient's observed longitudinal covariate history is expressed as:

$$\lambda(t|Y_i) = \lambda_0(t)E[f(\tilde{\theta}_i(t), \beta)|Y_i, T_i \geq t],$$

where the function $f(\cdot)$ of the true covariate history $\tilde{\theta}_i(t)$ of the $i^{th}$ patient up to time $t$ is taken to be $\exp(\beta\theta_i(t))$. That is, only the current value of the entire true covariate history $\tilde{\theta}_i(t)$ is assumed to be predictive of survival at time $t$. Let us denote the expectation in the above equation by $E[t, \beta]$. In order to estimate $\beta$ using partial likelihood (1.23) (with $E[t, \gamma]$ replaced by $E[t, \beta]$) we need to find $E[t, \beta]$ at each ordered event time $T(i), i = 1, \ldots, d$.

To find this expectation, let us first notice that the joint distribution $(\theta_i(t), Y_i|T_i \geq t)$ is multivariate normal with mean

$$\mu_i = (\mu_0 + \mu_1 t, \mu_0 + \mu_1 t_1, \ldots, \mu_0 + \mu_1 t_n)'$$

and the covariance matrix

$$\Sigma_i = \begin{pmatrix} \Sigma_{11i} & \Sigma_{12i} \\ \Sigma_{21i} & \Sigma_{22i} \end{pmatrix},$$

where

$$\Sigma_{11i} = \sigma_0^2 + 2t\sigma_{01} + t^2\sigma_1^2,$$

$$\Sigma_{12i} = \left( \sigma_0^2 + (t + t_1)\sigma_{01} + tt_1\sigma_1^2, \sigma_0^2 + (t + t_2)\sigma_{01} + tt_2\sigma_1^2, \ldots, \sigma_0^2 + (t + t_n)\sigma_{01} + tt_n\sigma_1^2 \right),$$

and $\Sigma_{22i} = \left( (\Sigma_{22i})_{k_1 k_2} \right)$ with

$$(\Sigma_{22i})_{k_1 k_2} = \sigma_0^2 + \sigma_{01}(t_{k_1} + t_{k_2}) + \sigma_1^2 t_{k_1} t_{k_2}$$
for $k_1 \neq k_2$, $1 \leq k_1, k_2 \leq n_i$, and

$$(\Sigma_{22i})_{kk} = \sigma_0^2 + 2\sigma_{01} t_k + \sigma_1^2 t_k^2 + \sigma_i^2$$

for $1 \leq k \leq n_i$.

Thus, $(\theta_i(t)|Y_i, T_i \geq t)$ is also normal with mean

$$\mu(t|Y_i) = \mu_0 + \mu_1 t + \Sigma_{12i} \Sigma_{22i}^{-1} (Y_i - \mu_i),$$

and variance

$$\sigma_2(t|Y_i) = \Sigma_{11i} - \Sigma_{12i} \Sigma_{22i}^{-1} \Sigma_{21i}.$$

Now, recognizing

$$E[t, \beta] = E[e^{\beta \theta_i(t)}|Y_i, T_i \geq t]$$

as the moment generating function of $(\theta_i(t)|Y_i, T_i \geq t)$ it follows that for the $i^{th}$ patient

$$E[t, \beta] = \exp \left( \beta \mu(t|Y_i) + \frac{\beta^2}{2} \sigma^2(t|Y_i) \right).$$

Having $E[t, \beta]$ found, we can proceed to the estimation of the model parameters. The procedure used to sample from the joint posterior distribution is the same 2 step procedure that was used in section 3.3.3.1 to sample from the joint posterior distribution $\pi(\beta, \Phi|D)$ of all the model parameters, but with $\Phi$ now being $(b, \mu, \Sigma^{-1}, \sigma_i^{-2})$, where $b = (b_1, b_2, \ldots, b_i)$, and $b_i = (b_{0i}, b_{1i})$. Please, refer to section 3.3.3.1 for details regarding this procedure. Similarly, Gibbs sampling can be used to sample from the posterior distribution of newly defined $\Phi$, and Metropolis-Hastings algorithm to
sample from the conditional posterior distribution of $\beta$ given $\Phi$. The ties in the data are again handled via Breslow approximation to the partial likelihood. The form of the likelihood, posterior distribution of $\Phi$ and full conditional distributions of the parameters in $\Phi$ are included in section A.3.

4.5.2 The Priors

The following priors were used in the analysis:

- $N(0, I_2 10^6)$ for $\mu$,

- $W(3, I_2 10^{-3})$ for $\Sigma^{-1}$, where $W(\rho, R)$ is a Wishart distribution with $\rho$ degrees of freedom and a precision matrix $R$,

- $G(0.001, 0.001)$ for $\sigma_i^{-2}$,

- $N(0, 10^6)$ for $\beta$.

Each $b_i$ is a combination of the observed data for the $i^{th}$ patient and the parameters $\mu$ and $\Sigma$. Thus $N(\mu, \Sigma)$ can be considered as the prior distribution for each $b_i$.

4.5.3 Simulation

The posterior distribution sampling algorithm was run for 15,000 iterations, on the same 100 data sets as the 2-stage and joint methods (with $\sigma_i^2 = 20$). The first 5,000 iterations were discarded as the burn-in period. Out of the remaining 10,000 iterations, every 20th (in total 500 samples) was used to generate 1,000 values of $\beta$. 
Of the total 500,000 generated \( \beta \)'s, only the last 100,000 were kept. Based on 10,000 generated values of \( \Phi \) and 100,000 values of \( \beta \) the statistical conclusions were drawn.

The simulation took about 5 hours on Pentium M 1.60GHz processor with 504MB RAM.

### 4.5.4 Results

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>sd</th>
<th>95%L</th>
<th>95%U</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_2^2 )</td>
<td>33.35</td>
<td>1.93</td>
<td>29.30</td>
<td>36.81</td>
</tr>
<tr>
<td>( \mu_0 )</td>
<td>109.87</td>
<td>0.71</td>
<td>108.6</td>
<td>111.26</td>
</tr>
<tr>
<td>( \mu_1 )</td>
<td>-1.01</td>
<td>0.10</td>
<td>-1.18</td>
<td>-0.82</td>
</tr>
<tr>
<td>( \sigma_0^2 )</td>
<td>2.97</td>
<td>4.58</td>
<td>0.015</td>
<td>15.49</td>
</tr>
<tr>
<td>( \sigma_{01} )</td>
<td>0.18</td>
<td>0.51</td>
<td>-1.12</td>
<td>0.93</td>
</tr>
<tr>
<td>( \sigma_1^2 )</td>
<td>0.27</td>
<td>0.076</td>
<td>0.14</td>
<td>0.42</td>
</tr>
<tr>
<td>( \beta )</td>
<td>-0.046</td>
<td>0.018</td>
<td>-0.08</td>
<td>-0.012</td>
</tr>
</tbody>
</table>

Table 4.4: Simulation study results (Ibrahim et al. (2001) method).

The results in Table 4.4 show that the method highly overestimated the observational variance \( \sigma_2^2 \). This is because the model imposes a linear pattern on the longitudinal trajectories of the patients, not allowing for any changes in the shape of these profiles to be reflected in their estimates. Thus, a lot of what in reality are true deviations from a straight line, is taken for a measurement error. This is confirmed by the overall posterior mean trajectory showed in Figure 4.13, where we can see how the estimated path "cuts" through all the true shifts in the overall true profile, assuming that these shifts resulted from the imprecision of the observed data. Let us also note that \( \Sigma \), which represents the between patient variability of the covariate trajectories,
Figure 4.13: Overall posterior mean trajectories with 95% credible set: 2-stage and joint vs. Ibrahim et al. (2001)

is fairly small compared to $\sigma^2$. This means that a lot more weight is given to the overall population mean than to the data in estimating $b_i$'s. The small value of $\Sigma$ does not allow for much variability in $b_i$'s, and the true differences between the individual longitudinal profiles of the patients are again mistaken for the measurement error in the data. These findings are consistent with the plots of the patient specific trajectories for patient 9 and 34 displayed in Figure 4.14. Especially for patient 34 we see how inaccurate the model was in estimating this patient's true covariate profile, pulling the individual's trajectory toward its population mean. Also, let us note that the 95% credible set bands are much wider for the posterior predictive mean trajectory for a new patient (Figure 4.15) than for the overall posterior mean trajectory estimated from the available data. This again is due to the largely overestimated
Figure 4.14: Patient specific posterior mean trajectories with 95% credible set: 2-stage and joint vs. Ibrahim et al. (2001)

Figure 4.15: Posterior predictive mean trajectories with 95% credible set: 2-stage and joint vs. Ibrahim et al. (2001)
observational variance which adds a lot of uncertainty when drawing new values from a sampling distribution of the data given the posterior estimates of the parameters. Moreover, since the model does not have the ability to adapt to the changes in the shape of the covariate profiles as we move along the time scale, the baseline value at time $t_0 = 0$ is also overestimated.

With respect to the estimated value of the survival regression coefficient $\beta$ the model performed equally well compared to the 2-stage approach and the joint method. Perhaps the differences in this aspect of the model would become more apparent if the true generated longitudinal data was even further away from a linear pattern assumed by this model.

4.6 Remarks

The comparison of the 2-stage and joint models to the method by Ibrahim et al. (2001) reveals a clear advantage of dynamic modeling approach in the context of joint modeling of survival data with longitudinal covariates. The dynamic models, as confirmed by the results of the simulation study presented in previous sections, allow for changes in the evolution of the longitudinal profiles over time and allow these changes to be determined by the data, not an assumed model, which was the case in the Ibrahim et al. (2001) method. Moreover, through the hierarchical structure of these models, they allow for a proper separation of the within and between patient variability in the longitudinal data, thanks to which one is able to accurately esti-
mate both patient specific and population longitudinal trajectories, as well as make predictions for new individuals.
Chapter 5

Conclusions and Future Considerations

5.1 Advantages

Let us summarize the advantages, limitations and possible future extensions of the proposed approach for joint analysis of longitudinal and survival data.

First of all, the model is very flexible in a sense that it does not force the longitudinal covariate trajectories to follow any particular shape, but allows the evolution of these trajectories to be shaped by the information contained in the data. In many studies involving longitudinal data, assuming a linear, quadratic or any other shape for the trajectory of the time-dependent covariate, may be too restrictive, and may cloud the real nature of the process that occurs in patients, due to the studied dis-
ease, or applied treatment. This was confirmed by the results of the simulation study carried out in chapter 4.

Second, the model allows to capture both patient-specific changes in the longitudinal profiles as well as the changes that occur at the level of the entire population. This allows to answer questions of interest regarding the within and between patient variability of the measured response.

Third, because of the hierarchical structure of the model, the method allows for valid prediction of the path of the longitudinal response for new patients.

Fourth, since the approach is fully Bayesian it allows for incorporation of prior knowledge of the parameters that can come from the experience of the investigator or previous studies. If such knowledge is not available, the investigator can avoid making a strong subjective input into the analysis by setting very flat priors for the model parameters. The model was proved to be quite resistant to the changes in these prior moments, as showed by the sensitivity analysis in section 4.4.

Also, because of the nature of Bayesian techniques, the analyst is always allowed to look at the data and, if desired, make decisions while collecting it. New data can always be added, and the posterior distributions obtained in the analysis with the data available thus far can simply serve as our prior knowledge in the analysis of the incoming data.

Although both presented methods performed equally well on the simulated data, a slight edge is given to the joint model, due to the fact that, as we mentioned before,
the data is used more efficiently and the variance estimates of the longitudinal model parameters are adjusted for the uncertainty in the survival model parameters, and vice versa. Moreover, censoring is now reflected in the longitudinal part of the model as well.

5.2 Future Considerations

The following future considerations for this project may be considered:

- Throughout, we have assumed an evenly spaced time grid that includes both the measurement times of the longitudinal covariate of all the patients, and their recorded survival times. The reason for this assumption was that we wanted to allow for even propagation of error over time. Even if the longitudinal measurements for different patients were taken at different times, assuming an evenly spaced grid for the analysis will still allow for the uncertainty of the parameters to be properly “built up” between the successive times of the data collection for each patient. One of the future considerations of this work, however, could be to consider an unevenly spaced grid and investigate gains and losses of such approach.

- Another extension of the model is to allow the variances in both the longitudinal and survival models to vary in time. This could allow for uneven propagation of error should unevenly spaced time grid wished to be used.
• Gamerman (1991) fitted a dynamic model to the survival data with fixed covariates that allowed the regression coefficient $\beta$ to change between time intervals of the assumed time grid. In our approach, however, we have assumed that the parameter $\beta$ is constant over time, thus a dynamic extension of the survival model where both log(baseline hazard) parameters $\alpha_j$ and $\beta$ are allowed to evolve as time passes could be considered.

• Often the goal of longitudinal studies is not only to study the patterns of change in the longitudinal marker profiles of the patients, but also to investigate how a particular treatment affects these changes and the survival of the patients. Thus, it would be worthwhile to develop an extension of the model presented in this project that would include treatment effect parameters and/or other covariates of interest.

• Further investigation of the robustness of the newly developed methods is required. For example, the issues of the amount of missing data, the size of the sample, and more heterogeneity between the individual trajectories of the patients, need to be addressed.

• Censoring and timing of the visits were assumed non-informative in the model. The patients whose longitudinal marker measurements reflect their poor condition, however, may be withdrawn from the study early, or may be missing their visits more often. Thus informative dropout and missing data mechanisms could
be considered as an extension of the method.

- It would also be worth investigating the relationship between survival of the patients and different functionals of the longitudinal profiles. Here, we have used only the current value as the most informative aspect of the longitudinal trajectory at each time $t$.

- Finally, as a sort of “by-product” of this research, usefulness of auxiliary particle filter methods in dynamic models for survival data was investigated. In the preliminary analysis of chapter 2, however, the variances in the fitted models were assumed known, and thus, an extension to fully unknown variances could be considered.
Appendix A

Sampling from the Joint Posterior Distribution

In this section the full forms of the posterior distributions from 2-stage, joint and Ibrahim et al. (2001) methods are presented, and the details regarding full conditional distributions used in Gibbs sampling are provided.

A.1 2-stage Model

Throughout, the following parameterization for the $G(\alpha, \beta)$ distribution is used:

If $X \sim G(\alpha, \beta)$, where $\alpha, \beta > 0$, then

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$
A.1.1 The Posterior Distribution of $\Phi$

The posterior distribution (3.11) for $\Phi$ up to the proportionality constant is:

$$
\pi(\Phi|D) \propto \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi c_\theta}} \exp \left( -\frac{1}{2c_\theta^2} \left( \theta_{i0} - m_\theta \right)^2 \right) \right\} \cdot \prod_{j=1}^{n_s} \left\{ \frac{1}{\sqrt{2\pi \sigma_t}} \exp \left( -\frac{1}{2\sigma_t^2} \left( y_{ij} - \theta_{ij} \right)^2 \right) \right\}
\cdot \frac{1}{\sqrt{2\pi \sigma_\omega}} \exp \left( -\frac{1}{2\sigma_\omega^2} \left( \theta_{ij} - \theta_{i,j-1} - \gamma_{j-1} \right)^2 \right)
\cdot \frac{1}{\sqrt{2\pi \sigma_\nu}} \exp \left( -\frac{1}{2\sigma_\nu^2} \left( \gamma_{j} - \gamma_{j-1} \right)^2 \right) \right\}
\cdot \frac{1}{\sqrt{2\pi \xi_\gamma}} \exp \left( -\frac{1}{2\xi_\gamma^2} \left( m_\gamma - \mu_\gamma \right)^2 \right)
\cdot \frac{1}{\Gamma(n_t)} \left( \sigma_\epsilon^{-2} \right)^{n_t-1} e^{-S_t \sigma_\epsilon^{-2}} \cdot \frac{1}{\Gamma(n_\omega)} \left( \sigma_\omega^{-2} \right)^{n_\omega-1} e^{-S_\omega \sigma_\omega^{-2}}
\cdot \frac{1}{\Gamma(n_\nu)} \left( \sigma_\nu^{-2} \right)^{n_\nu-1} e^{-S_\nu \sigma_\nu^{-2}} \cdot \frac{1}{\Gamma(n_\xi)} \left( \sigma_\xi^{-2} \right)^{n_\xi-1} e^{-S_\xi \sigma_\xi^{-2}}
$$

(A.1)

The full conditional distribution required to sample from A.1 via Gibbs sampling are constructed by “pulling out” only those elements of (A.1) that depend on the parameter of interest.

A.1.2 Full Conditionals for $\Phi$

Denoting by $n_{\pi_j}$ the number of patients at risk at $t_j$, $j = 1, \ldots, N$, the full conditional distributions of all the parameters in $\Phi$ given all the other parameters and the data are:
\( \theta_{i0}, i = 1, \ldots, n \) (assuming \( Y_{i0} \) is not observed):

\[
\pi(\theta_{i0}|\Phi_{(-\theta_{i0})}, D) \propto \frac{1}{\sqrt{2\pi}\sigma_{\epsilon}} \exp\left( -\frac{1}{2\sigma_{\epsilon}^2} \left( \theta_{i0} - m_\theta \right)^2 \right)
\]

\[
\frac{1}{\sqrt{2\pi}\sigma_\omega} \exp\left( -\frac{1}{2\sigma_\omega^2} \left( \theta_{i1} - \theta_{i0} - \gamma_0 \right)^2 \right)
\]

\[= N(M_{\theta_{i0}}, \Sigma_{\theta_{i0}}), \]

where

\[
M_{\theta_{i0}} = \Sigma_{\theta_{i0}} \left( \frac{\theta_{i1} - \gamma_0}{\sigma_\omega^2} + \frac{m_\theta}{\sigma_\theta^2} \right), \\
\Sigma_{\theta_{i0}} = \left( \sigma_\omega^{-2} + \sigma_\theta^{-2} \right)^{-1}.
\]

\( \theta_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_{i-1} \):

\[
\pi(\theta_{ij}|\Phi_{(-\theta_{ij})}, D) \propto \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp\left( -\frac{1}{2\sigma_\epsilon^2} \left( Y_{ij} - \theta_{ij} \right)^2 \right)
\]

\[
\frac{1}{\sqrt{2\pi}\sigma_\omega} \exp\left( -\frac{1}{2\sigma_\omega^2} \left( \theta_{ij} - \theta_{i,j-1} - \gamma_{j-1} \right)^2 \right)
\]

\[
\frac{1}{\sqrt{2\pi}\sigma_\omega} \exp\left( -\frac{1}{2\sigma_\omega^2} \left( \theta_{i,j+1} - \theta_{ij} - \gamma_j \right)^2 \right)
\]

\[= N(M_{\theta_{ij}}, \Sigma_{\theta_{ij}}), \]

where

\[
M_{\theta_{ij}} = \Sigma_{\theta_{ij}} \left( \frac{Y_{ij}}{\sigma_\epsilon^2} + \frac{\theta_{i,j-1} + \gamma_{j-1} + \theta_{i,j+1} - \gamma_j}{\sigma_\omega^2} \right), \\
\Sigma_{\theta_{ij}} = \left( \sigma_\epsilon^{-2} + 2\sigma_\omega^{-2} \right)^{-1}.
\]

\( \theta_{i,n_i}, i = 1, \ldots, n \):

\[
\pi(\theta_{i,n_i}|\Phi_{(-\theta_{i,n_i})}, D) \propto \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp\left( -\frac{1}{2\sigma_\epsilon^2} \left( Y_{i,n_i} - \theta_{i,n_i} \right)^2 \right)
\]

\[
\frac{1}{\sqrt{2\pi}\sigma_\omega} \exp\left( -\frac{1}{2\sigma_\omega^2} \left( \theta_{i,n_i} - \theta_{i,n_i-1} - \gamma_{n_i-1} \right)^2 \right)
\]

\[= N(M_{\theta_{i,n_i}}, \Sigma_{\theta_{i,n_i}}), \]
where

\[ M_{\theta_{i,n_i}} = \sum_{\theta_{i,n_i}} \left( \frac{Y_{i,n_i}}{\sigma_\varepsilon^2} + \frac{\theta_{i,n_i-1} + \gamma_{n_i-1}}{\sigma_\omega^2} \right), \]

\[ \Sigma_{\theta_{i,n_i}} = \left( \sigma_\varepsilon^{-2} + \sigma_\omega^{-2} \right)^{-1}. \]

• \( \gamma_0 \):

\[ \pi(\gamma_0|\Phi_{(-\gamma_0)}, D) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_\nu} \exp \left( -\frac{1}{2\sigma_\nu^2} \left( \gamma_i - \gamma_0 \right)^2 \right) \]
\[ \cdot \frac{1}{\sqrt{2\pi}\sigma_\nu} \exp \left( -\frac{1}{2\sigma_\nu^2} \left( \gamma_1 - \gamma_0 \right)^2 \right) \]
\[ \cdot \frac{1}{\sqrt{2\pi}\sigma_\gamma} \exp \left( -\frac{1}{2\sigma_\gamma^2} \left( \gamma_0 - m_\gamma \right)^2 \right) \]

\[ = N(M_{\gamma_0}, \Sigma_{\gamma_0}), \]

where

\[ M_{\gamma_0} = \sum_{\gamma_0} \left( \frac{\sum_{i=1}^{n} (\theta_{i1} - \theta_{i0})}{\sigma_\omega^2} + \frac{\gamma_1}{\sigma_\nu^2} + \frac{m_\gamma}{\sigma_\gamma^2} \right), \]

\[ \Sigma_{\gamma_0} = \left( N\sigma_\omega^{-2} + \sigma_\nu^{-2} + \sigma_\gamma^{-2} \right)^{-1}. \]

• \( \gamma_j, j = 1, \ldots, N - 2 \):

\[ \pi(\gamma_j|\Phi_{(-\gamma_j)}, D) \propto \prod_{i \in \mathcal{R}_j} \frac{1}{\sqrt{2\pi}\sigma_\omega} \exp \left( -\frac{1}{2\sigma_\omega^2} \left( \theta_{i,j+1} - \theta_{i,j} - \gamma_j \right)^2 \right) \]
\[ \cdot \frac{1}{\sqrt{2\pi}\sigma_\nu} \exp \left( -\frac{1}{2\sigma_\nu^2} \left( \gamma_j - \gamma_{j-1} \right)^2 \right) \]
\[ \cdot \frac{1}{\sqrt{2\pi}\sigma_\nu} \exp \left( -\frac{1}{2\sigma_\nu^2} \left( \gamma_{j+1} - \gamma_j \right)^2 \right) \]

\[ = N(M_{\gamma_j}, \Sigma_{\gamma_j}), \]

where

\[ M_{\gamma_j} = \sum_{\gamma_j} \left( \frac{\sum_{i \in \mathcal{R}_j} (\theta_{i,j+1} - \theta_{i,j})}{\sigma_\omega^2} + \frac{\gamma_{j-1} + \gamma_{j-1}}{\sigma_\nu^2} \right), \]
\[ \Sigma_{\gamma_i} = \left( n_{\mathcal{R}_i} \sigma_\omega^{-2} + 2 \sigma_\nu^{-2} \right)^{-1}. \]

- \( \gamma_{N-1} \):

\[
\pi(\gamma_{N-1} | \Phi(-\gamma_{N-1}), D) \propto \prod_{i \in \mathcal{R}_{N-1}} \frac{1}{\sqrt{2\pi\sigma_\omega}} \exp\left( -\frac{1}{2\sigma_\omega^2} \left( \theta_{iN} - \theta_{i,N-1} - \gamma_{N-1} \right)^2 \right) \\
\cdot \frac{1}{\sqrt{2\pi\sigma_\nu}} \exp\left( -\frac{1}{2\sigma_\nu^2} \left( \gamma_{N-1} - \gamma_{N-2} \right)^2 \right) \\
= N(M_{\gamma_{N-1}}, \Sigma_{\gamma_{N-1}}),
\]

where

\[
M_{\gamma_{N-1}} = \Sigma_{\gamma_{N-1}} \left( \frac{\sum_{i \in \mathcal{R}_{N-1}} (\theta_{iN} - \theta_{i,N-1})}{\sigma_\omega^2} + \frac{\gamma_{N-2}}{\sigma_\theta^2} \right),
\]

\[
\Sigma_{\gamma_{N-1}} = \left( n_{\mathcal{R}_{N-1}} \sigma_\omega^{-2} + \sigma_\nu^{-2} \right)^{-1}.
\]

- \( m_\theta \):

\[
\pi(m_\theta | \Phi(-m_\theta), D) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\theta}} \exp\left( -\frac{1}{2\sigma_\theta^2} \left( \theta_{i\theta} - m_\theta \right)^2 \right) \\
\cdot \frac{1}{\sqrt{2\pi\xi_\theta}} \exp\left( -\frac{1}{2\xi_\theta^2} \left( m_\theta - \mu_\theta \right)^2 \right) \\
= N(M_{m_\theta}, \Sigma_{m_\theta}),
\]

where

\[
M_{m_\theta} = \Sigma_{m_\theta} \left( \frac{\sum_{i=1}^{n} \theta_{i\theta}}{\sigma_\theta^2} + \frac{\mu_\theta}{\xi_\theta} \right),
\]

\[
\Sigma_{m_\theta} = \left( n_{\mathcal{R}_{\theta}}^{-2} + \xi_\theta^{-2} \right)^{-1}.
\]
\( c_{\theta}^{-2} \):

\[
\pi(c_{\theta} \mid \Phi_{(-c_{\theta})}, D) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi c_{\theta}}} \exp \left( -\frac{1}{2c_{\theta}^2} \left( \theta_{i0} - m_{\theta} \right)^2 \right) \\
\cdot \frac{S_{c_{\theta}}^{n_{c}}}{\Gamma(n_{c})} \left( c_{\theta}^{-2} \right)^{n_{c}-1} e^{-S_{c_{\theta}} c_{\theta}^{-2}}
\]

\[= G(n_{c}^*, S_{c}^*), \]

where

\[n_{c}^* = n_{c} + \frac{n}{2}, \]

\[S_{c}^* = S_{c} + \frac{1}{2} \sum_{i=1}^{n} (\theta_{i0} - m_{\theta})^2. \]

\( m_{\gamma} \):

\[
\pi(m_{\gamma} \mid \Phi_{(-m_{\gamma})}, D) \propto \frac{1}{\sqrt{2\pi c_{\gamma}}} \exp \left( -\frac{1}{2c_{\gamma}^2} \left( \gamma_{0} - m_{\gamma} \right)^2 \right) \\
\cdot \frac{1}{\sqrt{2\pi \xi_{\gamma}}} \exp \left( -\frac{1}{2\xi_{\gamma}^2} \left( m_{\gamma} - \mu_{\gamma} \right)^2 \right)
\]

\[= N(M_{m_{\gamma}}, \Sigma_{m_{\gamma}}), \]

where

\[M_{m_{\gamma}} = \Sigma_{m_{\gamma}} \left( \frac{\gamma_{0}}{c_{\gamma}^2} + \frac{\mu_{\gamma}}{\xi_{\gamma}^2} \right), \]

\[\Sigma_{m_{\gamma}} = \left( c_{\gamma}^{-2} + \xi_{\gamma}^{-2} \right)^{-1}. \]

\( \sigma_{\epsilon}^{-2} \):

\[
\pi(\sigma_{\epsilon} \mid \Phi_{(-\sigma_{\epsilon})}, D) \propto \prod_{i=1}^{n_{t}} \prod_{j=1}^{n_{t}} \frac{1}{\sqrt{2\pi \sigma_{\epsilon}}} \exp \left( -\frac{1}{2\sigma_{\epsilon}^2} \left( Y_{ij} - \theta_{ij} \right)^2 \right) \\
\cdot \frac{S_{\sigma_{\epsilon}}^{n_{t}}}{\Gamma(n_{\epsilon})} \left( \sigma_{\epsilon}^{-2} \right)^{n_{t}-1} e^{-S_{\sigma_{\epsilon}} \sigma_{\epsilon}^{-2}}
\]

\[= G(n_{\epsilon}^*, S_{\epsilon}^*), \]
where
\[ n_\varepsilon^* = n_\varepsilon + \frac{1}{2} \sum_{i=1}^{n} n_i, \]
\[ S_\varepsilon^* = S_\varepsilon + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (Y_{ij} - \theta_{ij})^2. \]

- \( \sigma_\omega^{-2} : \)
\[
\pi(\sigma_\omega | \Phi(-\sigma_\omega), D) \propto \prod_{i=1}^{n} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_\omega}} \exp\left(-\frac{1}{2\sigma_\omega^2}(\theta_{ij} - \theta_{i,j-1} - \gamma_{j-1})^2\right) \\
\frac{S_\omega^{n_\omega}}{\Gamma(n_\omega)} (\sigma_\omega^{-2})^{n_\omega-1} e^{-S_\omega \sigma_\omega^{-2}} \\
= G(n_\omega^*, S_\omega^*),
\]

where
\[ n_\omega^* = n_\omega + \frac{1}{2} \sum_{i=1}^{n} n_i, \]
\[ S_\omega^* = S_\omega + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (\theta_{ij} - \theta_{i,j-1} - \gamma_{j-1})^2. \]

- \( \sigma_\nu^{-2} : \)
\[
\pi(\sigma_\nu | \Phi(-\sigma_\nu), D) \propto \prod_{j=1}^{N-1} \frac{1}{\sqrt{2\pi\sigma_\nu}} \exp\left(-\frac{1}{2\sigma_\nu^2}(\gamma_j - \gamma_{j-1})^2\right) \\
\frac{S_\nu^{n_\nu}}{\Gamma(n_\nu)} (\sigma_\nu^{-2})^{n_\nu-1} e^{-S_\nu \sigma_\nu^{-2}} \\
= G(n_\nu^*, S_\nu^*),
\]

where
\[ n_\nu^* = n_\nu + \frac{N-1}{2}, \]
\[ S_\nu^* = S_\nu + \frac{1}{2} \sum_{j=1}^{N-1} (\gamma_j - \gamma_{j-1})^2. \]
A.2 Joint Model

A.2.1 The Posterior Distribution of \((\psi, \Phi)\)

The full form of the posterior distribution (3.26) for \((\psi, \Phi)\) up to the proportionality constant is:

\[
\pi(\psi, \Phi|D) \propto \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi \sigma_\theta}} \exp \left( -\frac{1}{2\sigma_\theta^2} \left( \theta_{i0} - m_\theta \right)^2 \right) \cdot \left( e^{\alpha_{n_1} + \beta \theta_{i,n_1}} \right)^{n_{i,n_1}} \prod_{j=1}^{n_{i}} \left[ \frac{1}{\sqrt{2\pi \sigma_e}} \exp \left( -\frac{1}{2\sigma_e^2} \left( Y_{ij} - \theta_{ij} \right)^2 \right) \cdot \exp \left( -e^{\alpha_j + \beta \theta_{ij} \Delta t} \right) \right] \cdot \frac{1}{\sqrt{2\pi \sigma_\omega}} \exp \left( -\frac{1}{2\sigma_\omega^2} \left( \theta_{ij} - \theta_{i,j-1} - \gamma_{j-1} \right)^2 \right) \cdot \frac{1}{\sqrt{2\pi \sigma_\eta}} \exp \left( -\frac{1}{2\sigma_\eta^2} \left( \alpha_j - \alpha_{j-1} \right)^2 \right) \right\} \cdot \frac{1}{\sqrt{2\pi \sigma_\nu}} \exp \left( -\frac{1}{2\sigma_\nu^2} \left( \gamma_0 - \gamma \right)^2 \right) \cdot \frac{1}{\sqrt{2\pi \sigma_\alpha}} \exp \left( -\frac{1}{2\sigma_\alpha^2} \left( \alpha_0 - \alpha \right)^2 \right) \cdot \frac{1}{\sqrt{2\pi \xi_\theta}} \exp \left( -\frac{1}{2\xi_\theta^2} \left( \mu_\theta - \mu \right)^2 \right) \cdot \frac{1}{\sqrt{2\pi \xi_\gamma}} \exp \left( -\frac{1}{2\xi_\gamma^2} \left( \mu_\gamma - \mu \right)^2 \right) \cdot \frac{S_{\nu}^{n\nu}}{\Gamma(n_\nu)} \left( \sigma_\nu^{-2} \right)^{n_\nu-1} e^{-S_\nu \sigma_\nu^{-2}} \cdot \frac{S_{\omega}^{n\omega}}{\Gamma(n_\omega)} \left( \sigma_\omega^{-2} \right)^{n_\omega-1} e^{-S_\omega \sigma_\omega^{-2}} \cdot \frac{S_{\gamma}^{n\gamma}}{\Gamma(n_\gamma)} \left( \sigma_\gamma^{-2} \right)^{n_\gamma-1} e^{-S_\gamma \sigma_\gamma^{-2}} \cdot \frac{1}{\sqrt{2\pi \sigma_\theta}} \exp \left( -\frac{1}{2\sigma_\theta^2} \left( \beta - m_\theta \right)^2 \right) \right) \right\} \right)
\]

(A.2)

A.2.2 Full Conditionals for \((\psi, \Phi)\)

The full conditional distributions of almost all the parameters in \(\Phi\) are exactly the same as the ones showed in the previous section for the 2 stage model and therefore
will not be repeated here. Below, the ones that do differ are presented:

- \( \theta_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_{i-1} \):

\[
\pi(\theta_{ij} | \Phi_{(\theta_{ij})}, D) \propto \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp \left( -\frac{1}{2\sigma_\varepsilon^2} \left( Y_{ij} - \theta_{ij} \right)^2 \right) \cdot \exp \left( -e^{\alpha_j + \beta_{ij} \Delta t} \right) \\
\cdot \frac{1}{\sqrt{2\pi}\sigma_j} \exp \left( -\frac{1}{2\sigma_j^2} \left( \theta_{ij} - \theta_{i,j-1} - \gamma_{j-1} \right)^2 \right) \\
\cdot \frac{1}{\sqrt{2\pi}\sigma_j} \exp \left( -\frac{1}{2\sigma_j^2} \left( \theta_{i,j+1} - \theta_{ij} - \gamma_j \right)^2 \right) \\
= N(M_{\theta_{ij}}, \Sigma_{\theta_{ij}}) \cdot \exp \left( -e^{\alpha_j + \beta_{ij} \Delta t} \right),
\]

where

\[
M_{\theta_{ij}} = \Sigma_{\theta_{ij}} \left( Y_{ij} \frac{\theta_{i,j-1} + \gamma_{j-1} + \theta_{i,j+1} - \gamma_j}{\sigma_j^2} \right),
\]

\[
\Sigma_{\theta_{ij}} = \left( \sigma_\varepsilon^{-2} + 2\sigma_j^{-2} \right)^{-1}.
\]

- \( \theta_{i,n_i}, i = 1, \ldots, n \):

\[
\pi(\theta_{i,n_i} | \Phi_{(\theta_{i,n_i})}, D) \propto \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp \left( -\frac{1}{2\sigma_\varepsilon^2} \left( Y_{i,n_i} - \theta_{i,n_i} \right)^2 \right) \\
\cdot \frac{1}{\sqrt{2\pi}\sigma_j} \exp \left( -\frac{1}{2\sigma_j^2} \left( \theta_{i,n_i} - \theta_{i,n_i-1} - \gamma_{n_i-1} \right)^2 \right) \\
\cdot \left( e^{\alpha_{n_i} + \beta_{i,n_i}} \right)^{\delta_{i,n_i}} \cdot \exp \left( -e^{\alpha_{n_i} + \beta_{i,n_i} \Delta t} \right) \\
= N(M_{\theta_{i,n_i}}, \Sigma_{\theta_{i,n_i}}) \cdot \left( e^{\alpha_{n_i} + \beta_{i,n_i}} \right)^{\delta_{i,n_i}} \cdot \exp \left( -e^{\alpha_{n_i} + \beta_{i,n_i} \Delta t} \right),
\]

where

\[
M_{\theta_{i,n_i}} = \Sigma_{\theta_{i,n_i}} \left( Y_{i,n_i} \frac{\theta_{i,n_i-1} + \gamma_{n_i-1}}{\sigma_j^2} \right),
\]

\[
\Sigma_{\theta_{i,n_i}} = \left( \sigma_\varepsilon^{-2} + \sigma_j^{-2} \right)^{-1}.
\]
\[ \pi(\alpha_0 | \Phi_{(\alpha_0)}, D) \propto \exp \left( -\frac{1}{2\sigma_{\eta}^2} (\alpha_0 - \alpha_0)^2 \right) \cdot \exp \left( -\frac{1}{2\sigma^2_{\alpha}} (\alpha_0 - m_\alpha)^2 \right) \]
\[ = N(M_{\alpha_0}, \Sigma_{\alpha_0}) \]

where
\[ M_{\alpha_0} = \Sigma_{\alpha_0} \left( \frac{\alpha_1}{\sigma^2_{\eta}} + \frac{m_\alpha}{\sigma^2_{\alpha}} \right), \]
\[ \Sigma_{\alpha_0} = \left( \sigma^{-2}_{\eta} + \sigma^{-2}_{\alpha} \right)^{-1}. \]

\[ \alpha_j, j = 1, \ldots, N - 1: \]
\[ \pi(\alpha_j | \Phi_{(\alpha_j)}, D) \propto \exp \left( -\frac{1}{2\sigma_{\eta}^2} (\alpha_j - \alpha_{j-1})^2 \right) \cdot \exp \left( -\frac{1}{2\sigma^2_{\alpha}} (\alpha_{j+1} - \alpha_j)^2 \right) \]
\[ \cdot \prod_{i=1}^{R_j} \left( e^{\alpha_j + \beta_{i,j}} \right) \delta_{i,j} \exp \left( -e^{\alpha_j + \beta_{i,j}} \Delta t \right) \]
\[ = N\left( \frac{\alpha_{j-1} + \alpha_{j+1}}{2}, \frac{\sigma_{\eta}}{2} \right) \cdot \exp \left( \alpha_j \sum_{i \in R_j} \delta_{ij} - e^{\alpha_j} \Delta t \sum_{i \in R_j} e^{\beta_{i,j}} \right). \]

\[ \alpha_N: \]
\[ \pi(\alpha_N | \Phi_{(\alpha_N)}, D) \propto \exp \left( -\frac{1}{2\sigma_{\eta}^2} (\alpha_N - \alpha_{N-1})^2 \right) \]
\[ \cdot \prod_{i=1}^{R_N} \left( e^{\alpha_N + \beta_{i,N}} \right) \delta_{i,N} \exp \left( -e^{\alpha_N + \beta_{i,N}} \Delta t \right) \]
\[ = N\left( \frac{\alpha_{N-1}}{2}, \frac{\sigma_{\eta}}{2} \right) \cdot \exp \left( \alpha_N \sum_{i \in R_N} \delta_{i,N} - e^{\alpha_N} \Delta t \sum_{i \in R_N} e^{\beta_{i,N}} \right). \]
• $\beta$:

$$
\pi(\beta_N|\Phi_{(-\beta_N)}, D) \propto \prod_{i=1}^{n} \prod_{j=1}^{n_t} \left( e^{\alpha_j + \beta R_{ij}} \right)^{\delta_{ij}} \exp \left( -e^{\alpha_j + \beta \epsilon_j} \Delta t \right) \cdot \frac{1}{\sqrt{2\pi c_\beta}} \exp \left( -\frac{1}{2c_\beta^2} (\beta - m_\beta)^2 \right) \\
= N(m_\beta, c_\beta) \cdot \exp \left( \beta \sum_{i=1}^{n} \sum_{j=1}^{n_t} \theta_{ij} \delta_{ij} - \sum_{i=1}^{n} \sum_{j=1}^{n_t} e^{\alpha_j + \beta \epsilon_j} \Delta t \right).
$$

A.3 Ibrahim et al. (2001) method

A.3.1 Likelihood

Assuming $\Phi = (b, \mu, \Sigma, \sigma_\epsilon^{-2})$, with $b$, $\mu$, and $\Sigma$ defined in section 4.5 the likelihood for the observed longitudinal data is:

$$
L(\Phi|D) = \prod_{i=1}^{n} \prod_{j=1}^{n_t} N(Y_{ij}|b_{0i} + b_{i1}t_j, \sigma_\epsilon^2) \tag{A.3}
$$

$$
= \prod_{i=1}^{n} \prod_{j=1}^{n_t} \frac{1}{\sqrt{2\pi\sigma_\epsilon}} \exp \left( -\frac{1}{2\sigma_\epsilon^2} (Y_{ij} - (b_{0i} + b_{i1}t_j))^2 \right)
$$

A.3.2 Prior for $\Phi$

The following prior is assumed for $\Phi$:

$$
\pi(\Phi) = \prod_{i=1}^{n} \left[ N(b_i|\mu, \Sigma) \right] N(\mu|\mu^*, \Sigma^*) W(\Sigma^{-1}|\rho, R^{-1}) G(n_\epsilon, S_\epsilon) \tag{A.4}
$$

$$
= \prod_{i=1}^{n} \frac{1}{2\pi|\Sigma|^\frac{1}{2}} \exp \left( -\frac{1}{2} (b_i - \mu)^t \Sigma^{-1} (b_i - \mu) \right) \cdot \frac{1}{2\pi|\Sigma^*|^\frac{1}{2}} \exp \left( -\frac{1}{2} (\mu - \mu^*)^t \Sigma^*^{-1} (\mu - \mu^*) \right) \cdot |R^{-1}|^\frac{\epsilon}{2} |\Sigma^{-1}|^\frac{\epsilon^2}{2} \exp \left( -\frac{1}{2} \text{tr}(R\Sigma^{-1}) \right) \cdot \frac{S_\epsilon^{n_\epsilon}}{\Gamma(n_\epsilon)} (\sigma_\epsilon^{-2})^{n_\epsilon-1} e^{-S_\epsilon \sigma_\epsilon^{-2}}
$$
A.3.3 The Posterior Distribution of $\Phi$

The posterior distribution for $\Phi$ is obtained by simply putting together A.3 and A.4 in a joint product.

$$\pi(\Phi|D) \propto L(\Phi|D)\pi(\Phi).$$

A.3.4 Full Conditionals for $\Phi$

Let us define

$$Z_i = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_{n_i} \end{pmatrix}'.$$

The full conditional distributions for parameters in $\Phi$ are (Wakefield et al. (1994)):

- $b_i$, $i = 1, \ldots, n$:

$$\pi(b_i|\Phi_{(-b_i)}, D) \propto \prod_{i=1}^{n} \left[ \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} (b_i - \mu)'\Sigma^{-1}(b_i - \mu) \right) \right] \frac{1}{\sqrt{2\pi}\sigma^i} \exp \left( -\frac{1}{2\sigma^2} (y_{ij} - (b_0i + b_1t_j))^2 \right)$$

$$= N\left(M_1, C_1\right),$$

where

$$M_1 = C_1 \left( \sigma^{-2}Z_i'y_i + \Sigma^{-1}\mu \right),$$

$$C_1 = \left( \sigma^{-2}Z_i'Z_i + \Sigma^{-1} \right)^{-1}. $$
\begin{itemize}
  \item $\mu$:
  \[
  \pi(\mu|\Phi_{(-\mu)}, D) \propto \prod_{i=1}^{n} \frac{1}{2\pi|\Sigma|^\frac{1}{2}} \exp \left( -\frac{1}{2} (b_i - \mu)' \Sigma^{-1} (b_i - \mu) \right) \\
  \cdot \frac{1}{2\pi|\Sigma^*|^\frac{1}{2}} \exp \left( -\frac{1}{2} (\mu - \mu^*)' \Sigma^*^{-1} (\mu - \mu^*) \right) \\
  = N\left(M_2, C_2\right)
  \]
  where
  \[
  M_2 = C_2 \left( n \Sigma^{-1} \bar{b} + \Sigma^{-1} \mu^* \right)
  \]
  \[
  C_2 = \left( n \Sigma^{-1} + \Sigma^*^{-1} \right)^{-1}
  \]
  with $\bar{b} = \frac{1}{n} \sum_{i=1}^{n} b_i$.

  \item $\Sigma^{-1}$:
  \[
  \pi(\Sigma^{-1}|\Phi_{(-\Sigma^{-1}), D}) \propto \prod_{i=1}^{n} \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} (b_i - \mu)' \Sigma^{-1} (b_i - \mu) \right) \\
  \cdot |R^{-1}|^{\frac{1}{2}} |\Sigma^{-1}|^{\frac{1}{2}} \exp \left( -\frac{1}{2} tr(R\Sigma^{-1}) \right) \\
  = W\left(\rho^*, R^{-1}\right)
  \]
  where
  \[
  \rho^* = n + \rho
  \]
  \[
  R^* = R + \sum_{i=1}^{n} (b_i - \mu)' (b_i - \mu)
  \]
\end{itemize}
\[ \pi(\sigma^{-2}_e | \Phi_{(\sigma^{-2}_e)}, D) \propto \prod_{i=1}^{n} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_e}} \exp \left( -\frac{1}{2\sigma_e^2} \left( Y_{ij} - (b_{0i} + b_{1i}t_{ij}) \right) \right) \]

\[ \frac{S_{e_i}^{n_e}}{\Gamma(n_e)} \left( \sigma^{-2}_e \right)^{n_e-1} e^{-S_e \sigma^{-2}_e} \]

\[ = \ G\left(n_e^*, S_e^*\right) \]  

(A.5)

where

\[ n_e^* = n_e + \frac{n}{2} \]

\[ S_e^* = S_e + \frac{1}{2} \sum_{i=1}^{n} (Y_i - Z_i b_i)' (Y_i - Z_i b_i). \]
Bibliography


