Lifetime Probability of Default – Cox Model with Time-Dependent Covariates Using Maximum Likelihood

By

Mark Thackham (41411889)

A thesis submitted to Macquarie University
for the degree of Master of Research
Department of Statistics
10th October 2016

MACQUARIE
University
SYDNEY·AUSTRALIA

Typeset in LaTeX 2_ε.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Mark Thackham (41411889)

Dedication

To Jia and Dorothy, for bringing me so much happiness.

Acknowledgements

I would like to greatly thank my research supervisor, Associate Professor Jun Ma. Thanks Jun for your guidance and patience when explaining technical detail, as well as in-depth feedback on my draft thesis. This year has been both very thought provoking and productive, and I look forward to working with you further.

Thanks to Dr Maurizio Manuguerra, who very kindly shared his `R` code translation for the `survival_mpl` package.

I am also appreciative of several anonymous reviewers for providing feedback on my draft thesis, whose constructive comments were of tremendous assistance and graciously received. I also acknowledge the data provider.

Finally, thank-you to my family, for their encouragement, support and for putting up with my late nights and industrious weekends.

Despite these above gratitudes (and the countless weekends and evenings foregone), all work, opinions and details (including any remaining errors or omissions) remain my own, and do not represent the views or opinions of any other individual or institution.

Abstract

Credit granting institutions need to estimate the probability of default, the chance a customer fails to make repayments as promised (BIS (2006) and IASB (2014)). The Cox model with time-varying covariates (Cox (1972), Crowley and Hu (1977)) is a technique often applied due to its substantial benefits beyond classification approaches (such as logistic regression) whilst achieving similar accuracy (Lessmann et al. (2015), Bellotti and Crook (2009)).

However partial likelihood estimation of this model has two short comings that remain unaddressed in the literature: (1) the baseline hazard is not estimated, so calculating probabilities requires a further estimation step; and (2) a covariance matrix for both regression parameters and the baseline hazard is not produced.

We address these by developing a maximum likelihood method that jointly estimates regression coefficients and the baseline hazard using constrained optimisation that ensures the baseline hazard's non-negativity. We show in a simulation our technique is more accurate in moderate sized samples and when applied to real home loan data it produces a smoother estimate of the baseline hazard than the Breslow (1972) estimator. Our model could be used predict life-time probability of default, required under the International Financial Reporting Standard (IFRS) 9 accounting standard.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and Aims

The aim of survival analysis is to model the time to an event of interest as a random variable. Take for example: the time to death of a patient in a clinical trial; the cumulative usage until the mechanical break-down of an industrial dishwasher; or the time until financial default of a home loan. The common theme in these three examples here is that subjects are under observation for a period of time in anticipation that an event of interest will occur. To help explain such data using available covariates in a regression setting, the semi-parametric Cox model (1972, 1975) has gained much popularity (see for example, Ren and Zhou (2011)). The model has also been extended to cater for time-varying covariates, as introduced by Crowley and Hu (1977) and discussed in Cox and Oakes (1984).

This ability of the Cox model to cater for time-varying covariates provides substantial benefits, as it allows information not available when a subject enters a study to be used in the model. For example, we may be interested in the survival of patients with kidney disease. In a study design that recruits patients at diagnosis, some patients may subsequently (after entering the study) receive a transplant. This information is likely to be very important as transplant recipients are expected to have longer survival times. This information can be included as a time-varying covariate in the model, indicating when recipients during the period of observation received their transplant.

While its earliest and most wide-spread application is in biomedical science and industrial life testing (Kalbfleisch and Prentice, 2002), survival analysis (and particularly the Cox model) is a prevailing method in developing probability of default models for bank loans (Lessmann et al., 2015). Credit-granting institutions such as banks are interested in such models to help estimate the probability that a customer fails to repay in a timely manner the monies they contractually owe (including principle, interest and fees). Estimating probability of default over a one-year period is a necessary input for banks to calculate their minimum capital required under the Basel Accords (BIS, 2006).

Survival models also have the added benefit (beyond classification techniques like logistic regression) as they can be used to determine not only if but when a customer is likely to default (see for example: Bellotti and Crook (2009), Stepanova and Thomas (2002) and Tong et al. (2012)). This feature is particularly beneficial in estimating multi-year and/or conditional probabilities. For example one can estimate the probability a customer defaults within 1 year, or within 2 years or origination. In addition, one can estimate the conditional probability that a customer does not default in the first year of a loan, but defaults anytime during the second year of the loan. In this survival context, one could view loans "surviving" until an event of interest (in this case default). This is demonstrated in the stylised "survival" function in figure 1.1. This feature of survival analysis could prove useful in estimating the lifetime probability of default, a key input for banks to calculate their expected credit losses required under International Financial Reporting Standard (IFRS) 9 accounting standard (IASB, 2014).

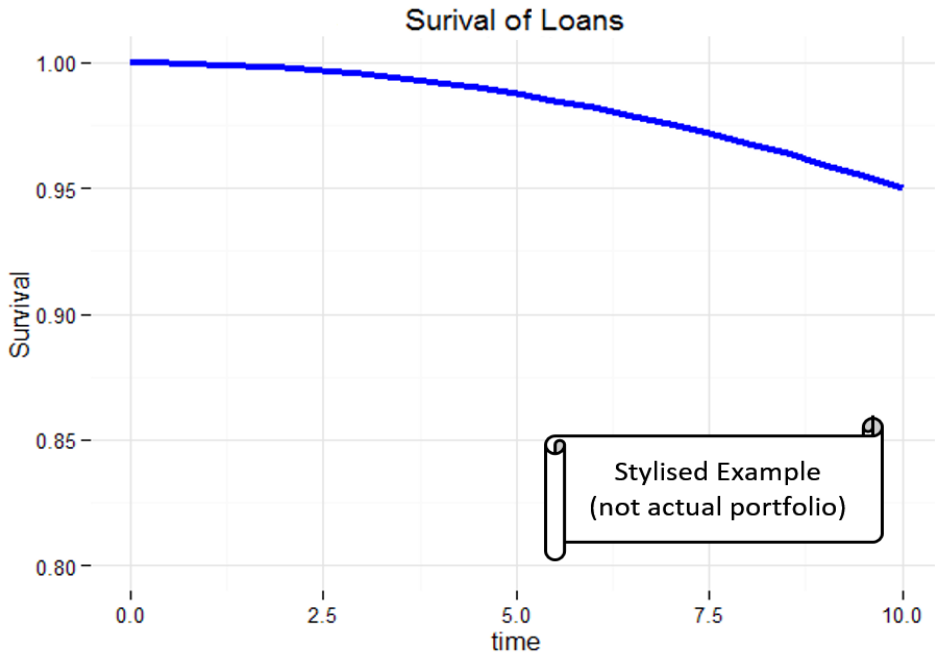

Figure 1.1: Stylised Example of the Survival of Loans

From a broader risk-management perspective, an accurate estimate of probability of default can help banks appropriately manage the risk of their loan book. Regulators too are keen to encourage sound risk management practices by banks in order to promote stability in the financial system, with the events of the Global Financial Crisis in 2008 showing how rapidly instabilities can spread to other parts of the economy. In an Australian context, the Australian Prudential Regulation Authority (APRA) provides supervision to 156 authorised deposit-taking institutions (either banks, building societies or credit unions) with a mission to *"...ensure that in all reasonable circumstances, financial promises made by institutions we supervise are met within a stable, efficient and competitive financial system"*.

These 156 authorised deposit-taking institutions consist of a few very large and many small institutions. The four largest comprise over 80% of all loan balances while the 10 largest comprise over 90% , and the 20 largest comprise over 95% (APRA, 2016). In addition, for each of these institutions homes loans are a dominant asset class.

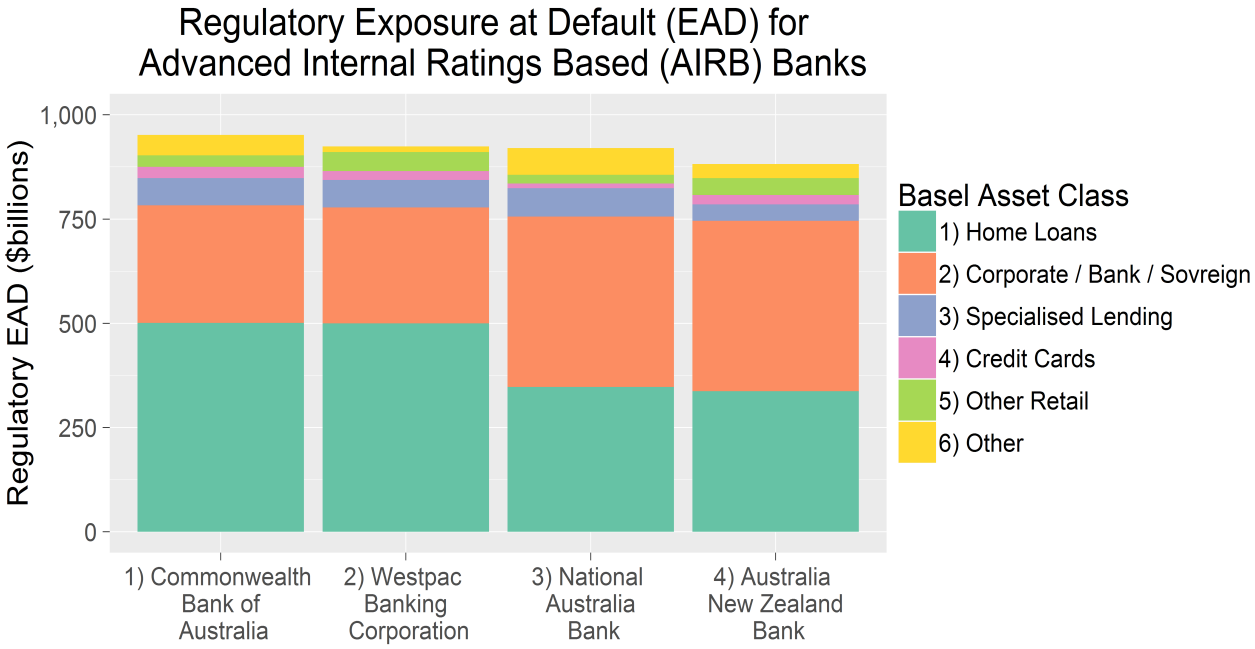

Figure 1.2: Exposure for the Four Largest Australian Banks
(Source: Commonwealth Bank of Australia (2016), Westpac Banking Corporation (2016), National Australia Bank (2016), Australia and New Zealand Banking Group (2016)).

Focussing on the four largest banks, figure 1.2 displays the regulatory exposure (as at 31st March 2016), split by Basel Asset Class (BIS, 2006). At an overall level, the total exposure of $3.2 trillion for these four banks put together is twice the 2015 Australian Gross Domestic Product (GDP) of $1.6 trillion (Australian Bureau of Statistics, 2016). Looking

at the distribution of exposure by Basel Asset Class, one can see that home loans (coloured green) account for a substantial proportion of all the four bank's exposure. It thus stands to reason that a clear measurement and understanding of the risk of home loans is important for an individual bank, as well as for stability of the financial system.

Despite its many applications, the Cox model extended for time-varying covariates when estimated using partial likelihood has two critical shortcomings.

1. The first shortcoming is that while it allows for estimation of regression coefficients, it does not directly provide an estimate for the baseline hazard, instead treating it as an arbitrary function of time (Cox and Oakes, 1984). This is not a problem if the only aim of the analysis is to draw inferences using the estimated coefficients (for example, determining that the risk of a disease increases with age) or to use hazard ratios (for example, finding that smokers are 3 times more likely to contract lung cancer than non-smokers). However, if probabilities are of interest, then an estimate of the baseline hazard is also needed, in addition to the regression coefficients. Available techniques include estimators from Breslow (1972) and Kalbfleisch and Prentice (2002), both of which require as a first step estimated regression coefficients from the Cox model as inputs. In addition to requiring an extra estimation step, these baseline hazard estimators produce estimates that are highly volatile.

2. The second shortcoming is that partial likelihood estimation does not produce a covariance matrix for both the fitted regression parameters and the baseline hazard. This means that joint inferences for the regression coefficients and baseline hazard cannot be drawn. The implication is that point-wise confidence intervals for non-baseline subjects can not be made directly unless a bootstrapping method is used.

We address both identified shortcomings of partial likelihood estimation by instead using maximum likelihood estimation for the Cox model with time-varying covariates. These shortcomings are addressed because: (1) both the regression coefficients and the baseline hazard are estimated jointly in one algorithm; and (2) a single Hessian matrix is produced covering all parameters (regression coefficients and the baseline hazard) allowing for joint inferences.

Note however that simply substituting estimation techniques (from partial likelihood to maximum likelihood) is inappropriate because while the regression coefficients are free to take any real value (positive or negative) the baseline hazard must be non-negative. Thus maximum likelihood estimation of the Cox model poses substantial and steep computational challenges. To meet these challenges, we develop a new constrained optimisation algorithm that undertakes maximum likelihood estimation while respecting the non-negativity constraint

of the baseline hazard. This methodology extends earlier work by Ma et al. (2014) to now include time-varying covariates in the Cox model.

1.2 Structure of this Thesis

This thesis contains six chapters including the current chapter (Chapter 1). Chapter 2 provides a background to survival analysis and survival data, covering important topics such as censoring and truncation. Chapter 3 is a literature review, discussing the Cox model in depth. It covers partial likelihood estimation, extensions of the Cox model to cater for time-varying covariates and estimation of the baseline hazard. It also covers the maximum penalised likelihood approach of Ma et al. (2014) for jointly estimating the baseline hazard and the regression coefficients of the Cox model. The chapter concludes by covering some literature relating to how survival models have been applied to estimating the probability of default for bank loans. Chapter 4 contains the novel research component of this thesis. It extends the maximum penalised likelihood model of Ma et al. (2014) discussed in Chapter 3 to also jointly estimate regression coefficients of time-varying covariates. The chapter contains the assumptions, derivation of theory and the constrained optimisation algorithm that we develop specifically for our maximum likelihood approach. The chapter finishes by discussing the implementation of our approach in `R`. Chapter 5 tests our new maximum likelihood method against the partial likelihood method, using both a simulation study and an application to credit risk data. Chapter 6 ends the thesis with conclusions and discussions as well as suggested avenues for intended future research. The thesis is supported with an appendix detailing the proof of the convergence properties of our method, an application of the delta method for point-wise confidence interval estimation of survival probabilities, as well as the full `R` code discussed in Chapter 4 to implement our approach.

2

Background to Survival Analysis

This chapter discusses briefly some important features of survival data and survival analysis. It outlines the three key elements that must be unambiguously defined in survival data, and also discusses two types of missing data that can effect survival analysis: censoring and truncation. It introduces the random variable studied in survival analysis, specifically the time to an event of interest, and finishes with a broad discussion of the various types of survival models available using some simple illustrative examples.

2.1 Survival Data

In survival analysis, we are interested in analysing ***subjects*** whose response variable is the length of ***time*** until the occurrence of a pre-defined ***event*** of interest, typically referred to as a "failure". Cox and Oakes (1984) state that survival data requires three elements unambiguously defined:

1. A time origin, representing when the subject became at risk of the event of interest.
2. A scale measuring the passage of time (for example, days, weeks, months, years).
3. A clearly defined event of interest (also known as a "failure").

To help solidify this terminology, table 2.1 lists some illustrative hypothetical examples that we continue to refer to during the course of this chapter. The examples cover a diverse set of applications of survival analysis, such as: hospital admission; industrial reliability;

clinical trials; survey results; and financial contracts. For each example listed in the table, the subjects have unambiguously defined: a time origin; a time scale; and an event of interest.

Example Study	Subject	Time Origin	Time Scale	Event of Interest
How long do patients in their 50's remain in hospital after admission for a heart attack?	Patients	Hospital admission	Number of days	Hospital discharge
How many hours usage can an industrial dishwasher withstand before failure?	Industrial dishwashers	First usage	Cumulative hours use	Mechanical breakdown
How long do laboratory rats survive after exposure to radiation?	Laboratory rats	Randomisation	Number of weeks	Death
How long do recent university graduates aged 25 and under take to secure their first job?	Graduates	Graduation	Number of months	Securing a job
How long since origination until a residential home loan defaults?	Home loans	Origination	Number of years	Home loan default

Table 2.1: Illustrative Examples of Studies with Survival Data.

Censoring and truncation are two common features of survival data, both of which are particular types of missing data (Klein and Moeschberger, 2003). Both of these can lead to incomplete observation time for a subject if either censoring and/or truncation are present in the study design. As this is a type of missingness, the modelling needs to be adapted appropriately otherwise the results may be biased (Sterne et al., 2009).

Censoring occurs when the event of interest for a subject is only known to occur during a certain period of time rather than being precisely known. There are three types of censoring schemes which can vary between each subject within the study:

- Right censoring, where all that is known is that a subject is still at risk for the event of interest at a certain point. This is further split into:
  - Type I - completely random drop-out and/or the study period ends and the subject has not yet had an event of interest.
  - Type II - the study ends when a fixed number of subjects have an event of interest.
- Left censoring, where all that is known is that a subject has experienced an event prior to entering into the study.

- Interval censoring, where the event of interest has occurred during a known interval, but the precise event time within this interval is not known.

Klein and Moeschberger (2003) state that in many studies, the censoring scheme is random (often referred to as "non-informative") type I right censoring. This assumes that the mechanism for right-censoring is independent from the survival times. It hypothesises that for each subject there is a right censoring time C and the exact survival time T is observed if and only if $T \leq C$. If $T > C$ then the subject has left the study without having an event of interest observed and the survival time for this subject is recorded as C. This data can be represented by the pair of random variables (Y, δ), where $Y = \min(T, C)$ represents a non-negative continuous random variable for the time to the event of interest for a subject, $\delta = 1$ when an event of interest is observed for the subject and $\delta = 0$ when the subject is censored.

In contrast to censoring, truncation is a deliberate decision made when constructing the sample design which determines whether or not a subject enters the study. There are two types of truncation schemes:

- Right truncation, where only subjects that have already experienced the event of interest are entered into the study.
- Left truncation, where subjects that survive a sufficient length of time are entered into the study. This is also known as "delayed entry", as the subject enters the study after a period of delay from the origin.

Truncation means that a subject will only enter the study if their survival time exceeds some threshold value. For those subjects that do enter the study, we observe their corresponding event time; for those that do not enter the study, they are completely unobserved and not even their existence is know (Kalbfleisch and Prentice, 2002).

While there are approaches to cater for complex censoring and truncation regimes, given that uninformative right censoring is in practice the type most commonly encountered in credit risk modelling, we restrict our methodology and discussion in this thesis to this type of censoring. The method we develop also assume there is no left truncation. Future research for our proposed methodology could include extension to (for example) informative right censoring, interval censoring as well as sample designs with truncation.

2.2 Time to Event as a Random Variable

For the random variable T introduced in section 2.1 above, Klein and Moeschberger (2003) outline several related functions describing its distribution, where knowledge of one

function will allow recovery of all the others.

One of the most frequently used methods to describe the distribution of the random variable T is the survival function $S(t)$, which describes the probability of a subject surviving beyond time t. It is also the complement of the cumulative distribution function $F(t)$. The survival function is

$$S(t) = P[T > t] = 1 - F(t) = 1 - \int_0^t f(u)du \tag{2.1}$$

where $f(t)$ denotes the density function of T. The hazard function $h(t)$ is fundamental to survival analysis, and plays a central role in the Cox model. The hazard function represents the instantaneous probability of an event of interest occurring for a subject, conditional on the subject not having experienced the event yet. The mathematical expression for the hazard is

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t < T \leq t + \Delta t | T > t]}{\Delta t} = \frac{f(t)}{S(t)}. \tag{2.2}$$

This leads into the cumulative hazard function $H(t) = \int_0^t h(u)du$ which represents the accumulated exposure to the event of interest.

2.3 Types of Survival Analysis

Censored time-to-event data was first studied and analysed by actuaries (Fisher and Lin, 1999), but now there are a wide variety of survival analysis techniques. These leverage the functions introduced in section 2.2 (which assumed homogeneous survival times for subjects) by extended them to compare survival between two or more groups or conditionally explaining survival using available covariates in a regression setting. There are broadly three categories of analysis techniques (see for example Hosmer et al. (2008)) which we outline below and highlight with examples.

The first are non-parametric techniques, for which no functional form is assumed for the distribution of failure times. Prominent methods include the Kaplan-Meier (1958) survival function estimator which estimates $S(t)$ and the Nelson-Aalen cumulative hazard estimator (Nelson (1969, 1972) and Aalen (1978)) which estimates $H(t)$. For instance, returning to the examples we introduced in table 2.1, these models could help answer the following questions:

- Do male heart attack patients remain in hospital longer than female patients?
- Does brand “A” dishwasher withstand longer cumulative usage than brand “B”?
- Do the rats who receive a new chemotherapy treatment live longer than rats who receive the existing treatment?

- Do art graduates take longer to find employment than science graduates?
- Do new customers who have no previous relationship with the bank default on home loans sooner than customers who have a previous affiliation with the bank?

The distinguishing feature here is that there are two (or more) discrete groups that we wish to analyse survival time for. Tests for significant differences in survival between groups can be undertaken, for example using the log-rank test (Mantel (1966), see also Peto and Peto (1972)).

The second type of survival analysis are semi-parametric techniques, which assumes a functional form for the covariates using a regression approach, but make no additional assumptions for the distribution of survival times. The Cox model (1972, 1975) is a widely used semi-parametric model for $h(t)$.

The third type of survival analysis are parametric techniques, which assumes a functional form for both covariates and the distribution of survival times using a regression approach. Examples include accelerated failure time models that use the Weibull, exponential and log-normal distributions for survival times (Kalbfleisch and Prentice, 2002).

Both semi-parametric and parametric survival analysis make use of a richer set of available candidate explanatory covariates that are thought to be related to survival. Hosmer et al. (2008) cites the attraction of such regression techniques is that plausible models may be easily fit, evaluated, and interpreted. Again referring to table 2.1, with additional covariates available, these models could help answer the following questions:

- How does patient age and hospital district impact duration of stay in hospital?
- How does the amount of soap and number of dishes impact the cumulative usage of dishwashers?
- How does chemotherapy dose and rat weight impact survival?
- How does GPA and geographic location affect job prospects for graduates?
- How does loan-to-value ratio and customer age impact the time to home loan default?

The distinguishing feature here is that there are multiple and (typically) continuous covariates available to help explain survival.

Often there are covariates whose value for a given individual may change over time during their observation in the study (Cox and Oakes, 1984). These types of variables are called time-varying (or time-dependent) covariates and differ fundamentally from baseline

covariates, whose values are measured only once for each subject at entry to the study and either do not change or do change but are not tracked over time. Referring to table 2.1, with the addition of time-varying covariates, we could ask the following questions regarding survival times:

- How does a patient's daily blood pressure impact duration of stay in hospital?
- What effect does maintenance during the observational period have on the cumulative usage of dishwashers?
- How does an increase in chemotherapy dose during the period of observation impact survival of rats exposed to radiation?
- How does a change of city of residence during observation affect job prospects of graduates?
- How does average house price movements over time impact the time to home loan default?

The common theme for all of these examples is that for some subjects, additional information becomes available after they enter the study. The Cox model extended for time-varying covariates allows this subsequently available information to be used in the model. We elaborate deeper in later chapters, but the most common approach to estimate the regression parameters of the Cox model extended for time-varying covariates uses the partial likelihood. This estimation approach has several deficiencies which our alternative maximum likelihood estimation approach (developed in this thesis) addresses.

3

Literature Review

The literature review in this chapter covers three main topics. The first is an in-depth technical background and discussion of one of the work-horse models of survival analysis: the Cox (1972, 1975) model. The discussion covers both the proportional hazards version (which contains only baseline covariates), as well as the extension by Crowley and Hu (1977) to cater for time-varying covariates. It also discusses techniques to estimate the baseline hazard (in particular those by Breslow (1972) and Kalbfleisch and Prentice (2002)). The second section of this chapter covers recent work by Ma et al. (2014) who develop an approach to jointly estimate the regression coefficients and the baseline hazard for the Cox model with only baseline covariates. The third and final section contains a literature review of key papers that apply survival analysis to probability of default modelling, with a particular focus on applications that include time-varying covariates.

3.1 The Cox Model

3.1.1 Baseline Covariates

The Cox (1972, 1975) model is the corner-stone of modern survival analysis (Zheng and Lin, 2007). Kalbfleisch and Prentice (2002) describe one of its chief benefits is that it allows an intuitive explanation of the hazard conditional on explanatory covariates. Suppose we have p explanatory covariates arranged in the vector $\underline{x}_i^T = [x_1, ..., x_p]$ available for each subject $i = 1, ..., n$ which are thought to explain time to an event of interest. The covariate

vector $\underline{x}_i^T$ may contain variables that can help to answer the simple example questions in section 2.3, including quantitative variables (such as customer age or loan balance), qualitative variables (such as treatment group or product type) as well as potentially interactions between covariates.

In the seminal papers in 1972 and 1975, the Cox model is specified as

$$h_i(t_i|\underline{x}_i) = h_0(t_i)e^{\underline{x_i}^T \underline{\beta}} \tag{3.1}$$

where $\underline{\beta}^T = [\beta_1, ..., \beta_p]$ are regression coefficients for the covariate vector $\underline{x_i}^T$, and $h_0(t)$ is an arbitrary unspecified function of time called the baseline hazard that combines multiplicatively with the covariate effects to act on the hazard function. This leads to the conditional survival function $S_i(t_i|\underline{x}_i) = S_0(t_i)^{[e^{\underline{x}_i^T \underline{\beta}}]}$, where $S_0(t)$ is the baseline survival function.

With only baseline (fixed-time) covariates, the Cox model is commonly referred to as the "proportional hazards" model, because hazards between subjects are proportional (Kalbfleisch and Prentice, 2002). That is, the ratio of hazards between two different subjects i and k

$$\frac{h_i(t|\underline{x}_i)}{h_k(t|\underline{x}_k)} = \frac{h_0(t)e^{\underline{x}_i^T \underline{\beta}}}{h_0(t)e^{\underline{x}_k^T \underline{\beta}}} = e^{(\underline{x}_i^T - \underline{x}_k^T)\underline{\beta}} \tag{3.2}$$

is constant over time. This value is called the relative risk (Klein and Moeschberger, 2003) or the hazard ratio (Hosmer et al., 2008). It is important to test that the assumption of proportional hazards holds when applying the Cox model to a dataset. This can be conducted using either scaled Schoenfeld (1981) residuals or by plotting log-log survival curves (Cox and Oakes, 1984). If these tests detect a violation of the proportional hazards assumption, then one remedy can be to include a time-varying covariate in the model (see Hosmer et al. (2008) for further details).

To estimate the effects $\underline{\beta}^T = [\beta_1, ..., \beta_p]$ in equation (3.1), using observed data $(T_i, \delta_i, \underline{x}_i)$, one uses Cox's partial likelihood (Cox 1972, 1975). The Cox partial likelihood can be derived as a profile likelihood using the likelihood adapted for censored data (Klein and Moeschberger (2003), Johansen (1983)).

Rather than detailing the derivation of the partial likelihood, we instead outline an intuitive derivation from Cox and Oakes (1984). In the absence of tied failure times, let $\tau_1 < \tau_2 < ... < \tau_d$ be the ordered failure times amongst the n subjects in the sample, where d is the total number of subjects in the sample ever observed to experience the event of interest and $(n-d)$ is the total number of censored subjects. Let $\mathscr{R}(\tau_j)$ be the risk set which is comprised of all subjects still under observation in the study (ie: either not yet censored and not yet

encountered an event of interest) just before the j^{th} failure time, $j = 1, ..., d$. Figure 3.1 below demonstrates this setup for four subjects who participate in a total of three risk sets between them. The risk set at the first ordered failure time τ_1 is $\mathscr{R}(\tau_1) = \{1, 2, 3, 4\}$, which is a risk set that contains all four subjects. This is because just before the time τ_1, all four subjects remain at risk of the event of interest. The risk sets for the second (τ_2) and third (τ_3) failure times are $\mathscr{R}(\tau_2) = \{1, 2\}$, and $\mathscr{R}(\tau_3) = \{2\}$ respectively. Subject 4 appears in the first risk set, but is censored between the ordered events times τ_1 and τ_2 and so does not participate in the second or subsequent risk sets.

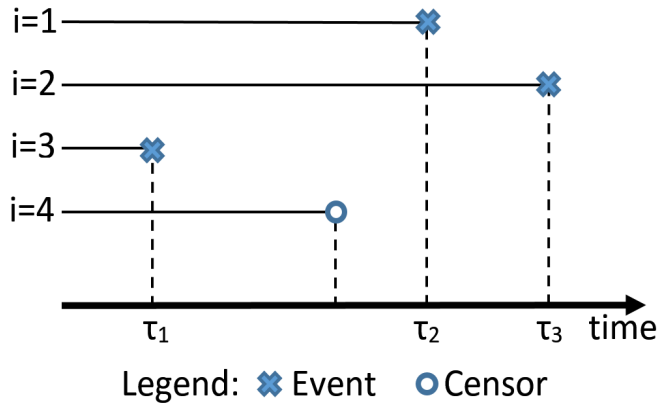

Figure 3.1: Failure of Four Individuals
(Source: Cox and Oakes (1984))

As per Klein and Moeschberger (2003), the conditional probability that subject i fails at τ_j with covariates $\underline{x}_i^T$ given that one individual fails from the risk set $\mathscr{R}(\tau_j)$ is

$$\begin{aligned} &P[\text{subject } i \text{ fails at } \tau_j \mid \text{one failure at } \tau_j] \\ &= \frac{P[\text{subject } i \text{ fails at } \tau_j \mid \text{subject } i \text{ survives to } \tau_j]}{P[\text{one failure at } \tau_j \mid \text{subject } i \text{ survives to } \tau_j]} \\ &= \frac{e^{\underline{x}_i^T \underline{\beta}}}{\sum_{k \in \mathscr{R}(\tau_j)} e^{\underline{x}_k^T \underline{\beta}}}. \end{aligned} \tag{3.3}$$

The partial likelihood is formed by multiplying these conditional probabilities over all d observed failure times from the sample, resulting in the partial likelihood function

$$L(\underline{\beta}) = \prod_{j=1}^{d} \frac{e^{\underline{x}_j^T \underline{\beta}}}{\sum_{k \in \mathscr{R}(\tau_j)} e^{\underline{x}_k^T \underline{\beta}}}. \tag{3.4}$$

While not a true likelihood, the partial likelihood is treated as such with inference carried out by the usual means (Hosmer et al., 2008). The numerator depends only on information from the individual subject who experiences an event of interest at τ_j, whereas the denominator

captures information from all subjects in the risk set (ie: those who have not yet experienced an event of interest and/or have not yet been censored). Estimation is carried out using the log-partial likelihood function

$$l(\underline{\beta}) = \sum_{j=1}^{d} \left[\underline{x}_j^T \underline{\beta} - \ln\left(\sum_{k \in \mathscr{R}(\tau_j)} e^{\underline{x}_k^T \underline{\beta}} \right) \right]. \tag{3.5}$$

The maximum partial likelihood estimator, denoted $\underline{\hat{\beta}}$, is obtained by differentiating equation (3.5) with respect to $\underline{\beta}$ and solving $\partial l(\underline{\beta})/\partial \underline{\beta} = \underline{0}$. The variance estimator is obtained in a similar manner for maximum likelihood, as the inverse of the negative second order derivative of the log-partial likelihood via the observed information $\boldsymbol{I}(\underline{\beta})$. Klein and Moeschberger (2003) discuss how these equations are amended for tied survival times.

3.1.2 Time Varying Covariates

Definitionally, the covariates $\underline{x}_i^T$ are measured only at baseline when the subject enters the study. The Cox model assumes that the values of these covariates remain fixed for each subject during the period of observation. An example is a patient‘s gender in a clinical trial, or the original loan amount requested by the customer when they applied for a home loan. Often there are covariates whose values potentially change for each subject over the period of observation while they are in the study. Examples of time-varying covariates could be a subject's blood-pressure measured at regular intervals (perhaps weekly) during the period of observation or monthly changes in house prices pledged as security for home loans. In this thesis, we define $\underline{z}_i(t_i)^T = [z_1(t_i), \ldots, z_q(t_i)]$ be a tuple of q time varying covariates for the i^{th} subject.

Analysing the famous Stanford heart transplant data, Crowley and Hu (1977) extend the Cox model to cater for time-varying covariates. Their model uses (in addition to baseline covariates) transplant status and transplant age to predict survival of patients. The method developed and applied by Crowley and Hu (1977) estimates the regression coefficients for both baseline and time-varying covariates using an amended version of the partial likelihood. The amendment allows the same subject to have potentially different values for their covariates in different risk sets. This assumes that the covariate process $\underline{z}_i^T(t)$ is known for any time which the subject is under observation (Klein and Moeschberger, 2003). Note however that the partial likelihood only requires values of $\underline{z}_i^T(t)$ to be precisely known at the d failure times (Therneau et al., 2015). Making this amended to the partial likelihood in equation (3.4) to now include time-varying covariates $\underline{z}_i^T(t)$ with regression coefficients $\underline{\gamma}$ results in

$$L(\underline{\beta}, \underline{\gamma}) = \prod_{j=1}^{d} \frac{e^{\underline{x}_j^T \underline{\beta} + \underline{z}_j^T(\tau_j)\underline{\gamma}}}{\sum_{k \in \mathscr{R}(\tau_j)} e^{\underline{x}_k^T \underline{\beta} + \underline{z}_k^T(\tau_k)\underline{\gamma}}}. \tag{3.6}$$

Parameter estimation and inference proceeds as per section 3.1.1 for the case with only baseline covariates.

3.1.3 Baseline Hazard Estimation

In many applications of the Cox model, recovery of survival probabilities (rather than simply just the regression coefficients) are of interest. Royston (2011) opines this should entail an explicit (preferably smooth) estimate of the baseline hazard function $h_0(t)$. For example, van Houwelingen (2000) states:

"It is the duty of the [statistician] involved in reporting the prognostic model to give all the information needed to build further on their model. For Cox models that should also include the baseline hazard or survival rate, if possible smoothed somehow or given in an approximate functional form using fractional polynomials, exponentials, rational functions or something similar."

Given that the partial likelihood specifically does not estimate the baseline hazard $h_0(t)$, recovery of survival probabilities requires an additional estimation step. This will allow estimation of the baseline survival function $S_0(t)$, which can be combined with the estimated effects to recover the survival conditional on the covariates $\underline{x}_i^T$, namely $S(t|\underline{x}_i) = S_0(t)^{exp(\underline{x}_i^T \beta)}$.

There are two common estimators available for this. The unifying theme between the two approaches is that they are both an additional estimation step which requires as input the regression coefficients $\hat{\underline{\beta}}$ (and $\hat{\underline{\gamma}}$ if time-varying covariates are also included) from the Cox model. The first of these is the Breslow (1972) estimator of the cumulative baseline hazard function $H_0(t)$, which can be derived by maximising a profile likelihood conditional on the log-hazard ratio estimates. This estimator has the undesirable feature that it produces point estimates of $h_0(t)$ that are very "noisy" and unstable (Hosmer et al., 2008). This estimator reduces to the non-parametric Nelson-Aalen estimator of the cumulative hazard when there are no covariates present (Klein and Moeschberger, 2003), hence it is sometimes referred to as the Nelson-Aalen-Breslow estimator. The second estimator is the Kalbfleisch and Prentice (2002) estimator of the baseline survival function $S_0(t)$ which mirrors the derivation of the Kapaln-Meier non-parametric estimator. Rodriguez (2005) sketches an outline of how the estimator is constructed. Like the Breslow (1972) estimator, it also produces "noisy" and unstable point estimates of $h_0(t)$. It reduces to the non-parametric Kaplan-Meier estimator when no covariates are present (Hosmer et al., 2008).

There have more recently been developments in the literature beyond the above two estimators for the baseline hazard. In general these later methods rely on a log-transform of $h_0(t)$ which guarantees positivity but will not allow the baseline hazard to ever equal zero. Clearly the baseline hazard $h_0(t)$ won't be zero for all values of time t, but there may be some period of time when the baseline hazard is equal to zero. For example, the Breslow (1972) and Kalbfleisch and Prentice (2002) estimators will provide for $\hat{h}_0(t) = 0$ for some values of t. This shortcoming is addressed by our proposed method (which we outline in the next chapter) by explicitly constraining the baseline hazard to be non-negative – that is, our hazard estimates can be either positive or exactly zero for any value of t.

Royston (2011) proposes a method to approximate the log-baseline hazard (which ensures positivity of the baseline hazard) using fractional polynomials and restricted cubic splines. However, similar to the Breslow (1972) and Kalbfleisch and Prentice (2002) methods, Royston's method requires as input the estimated regression coefficients from the partial likelihood method. This means that Royston does not attempt to estimate both the effects and the baseline hazard jointly.

Cai et al. (2010) estimate the hazard using linear splines to model the log-hazard (which ensures positivity of the hazard) with smoothing parameters estimated by restricted maximum likelihood (REML). The authors recast the problem as a mixed-effects Poisson regression with an offset term, which allows estimation in standard statistical packages, such as SAS or R. Their methodology does not cater for conditional explanation of survival times using covariates. This work extends that from Cai and Betensky (2003), which focussed on estimating the log-hazard using linear splines for interval censored data.

Kneib and Fahrmeir (2004) provide several extensions to the Cox model, calling their model a mixed-hazard regression. Their extensions include modelling log-baseline hazard (to ensure positivity of the baseline hazard) using penalised splines, as well as allowing for time-varying covariates.

3.2 Maximum Penalised Likelihood Estimation for Cox Model

Ma et al. (2014) develop an approach to simultaneously estimate the regression coefficients and the baseline hazard for the Cox model that contains only baseline (non-time-varying) covariates. The approach uses maximum penalised likelihood (MPL), where a penalty is used to impose a degree of smoothness to the baseline hazard. The parameters are fit using constrained optimisation to respect the non-negativity of the baseline hazard. Starting with

the hazard $h_i(t) = h_0(t)e^{x_i^T \underline{\beta}}$, the joint likelihood for all subjects ($i = 1, \ldots, n$) is

$$L(\underline{\beta}, h_0(t)) = \prod_{i=1}^{n} L_i(\underline{\beta}, h_0(t)) \tag{3.7}$$

where for the i^{th} subject with event time t_i, we have $L_i(\underline{\beta}, h_0(t_i)) = [f_i(t_i)]^{\delta_i} \times [S_i(t_i)]^{(1-\delta_i)}$, $\delta_i = 1$ for subjects that are observed to have an event of interest and $\delta_i = 0$ otherwise. The log-likelihood is then

$$l(\underline{\beta}, h_0(t)) = \ln\left(L(\underline{\beta}, h_0)\right) = \sum_{i=1}^{n} \delta_i \ln\left(f_i(t_i)\right) + (1 - \delta_i) \ln\left(S_i(t_i)\right). \tag{3.8}$$

Substituting into equation (3.8) the fact that $\ln(S_i(t_i)) = -H_i(t_i)$ and $f_i(t_i) = h_i(t_i)S_i(t_i)$, the log-likelihood becomes:

$$\begin{aligned} l(\underline{\beta}, h_0(t)) &= \sum_{i=1}^{n} \delta_i \left[\ln\left(h_i(t_i)\right) + \ln\left(S_i(t_i)\right)\right] + (1 - \delta_i) \ln\left(S_i(t_i)\right) \\ &= -\sum_{i=1}^{n} H_i(t_i) + \sum_{i=1}^{n} \delta_i \ln\left(h_i(t_i)\right) \end{aligned} \tag{3.9}$$

After the covariates $\underline{x}_i^T$ are introduced, equation (3.9) becomes

$$l(\underline{\beta}, h_0(t)) = -\sum_{i=1}^{n} H_0(t_i)e^{\underline{x}_i^T \underline{\beta}} + \sum_{i=1}^{n} \delta_i (log\left(h_0(t_i)\right) + \underline{x}_i^T \underline{\beta}). \tag{3.10}$$

Because $h_0(t)$ has infinite dimension, the authors introduce a basis function to approximate the baseline hazard $h_0(t) = \sum_{u=1}^{m} \theta_u \psi_u(t)$, where $\psi_u(t)$ is the basis function of u dimension and θ_u are values that require estimation when fitting the model. The authors impose two conditions on model fitting: (1) a smoothness constraint by adding a penalty term to the log-likelihood, denoted $\lambda J(h_0(t))$; and (2) requiring all values of θ_u to be greater than or equal to zero. This results in needing to undertake constrained maximisation of the penalised log-likelihood function

$$\Phi(\underline{\beta}, h_0(t)) = l(\underline{\beta}, h_0(t)) - \lambda J(h_0(t)). \tag{3.11}$$

Defining the basis for the cumulative baseline hazard as $\Psi_u(t_i) = \int_0^t \phi_u(v)dv$, the cumulative hazard is written as:

$$H_0(t) = \int_0^t h_0(s)ds = \sum_{u=1}^{m} \theta_u \int_0^t \psi_0(s)ds = \sum_{u=1}^{m} \theta_u \Psi_u(t). \tag{3.12}$$

This results in the penalised log-likelihood

$$\Phi(\underline{\beta}, h_0(t)) = -\sum_{i=1}^{n}\sum_{u=1}^{m}\theta_u e^{\underline{x}_i^T\underline{\beta}}\Psi_u(t_i) + \sum_{i=1}^{n}\delta_i\left(\ln\left(\sum_{u=1}^{m}\theta_u\psi_u(t)\right) + \underline{x}_i^T\underline{\beta}\right) - \lambda J(h_0(t)). \quad (3.13)$$

This thesis extends the above work to allow time varying covariates in the Cox model. While the maximum penalised likelihood approach uses a penalty term $\lambda J(h_0(t))$ to help control the smoothness of the baseline hazard, our maximum likelihood approach will instead control this via selecting the number of and distance between knots for the basis function $\psi(t)$.

3.3 Survival Analysis Applied to Credit Risk Modelling

Survival analysis appears widely in the literature as a method to estimate the probability that a bank loan will default (Lessmann et al., 2015). Survival analysis can be used to determine not only if, but when, a customer is likely to default, which is one of the key advantages beyond logistic regression, the most commonly employed method in industry. One of the earliest such applications was by Narian (1992), but several authors have since added to the literature. We review some of the key contributions to the literature that apply survival analysis to estimate probability of loan default.

Banasik et al. (1999) investigate using survival analysis for credit scoring, which is designed to answer the question of how likely an applicant for credit is to default by a given time in the future. The paper compares four models (logistic regression, the Cox model and the Weibull and exponential accelerated lifetime models) finding they all perform similarly given the data in their study. The paper also recognises the competing risk nature of credit risk data, and outlines (but does not fit) an approach that models loan default as the event of interest and successful repayment as a competing risk.

Bellotti and Crook (2009) show that using time-varying macroeconomic variables in the Cox model improves the accuracy of estimating the probability of default compared to both a Cox model (without time-varying covariates) and a logistic regression. The authors also argue that given credit data is typically in a panel format, where new accounts enter, old accounts leave and each account is observed for a sequential period of time, it naturally allows the use of survival models with time varying covariates. They also state an additional advantage of survival analysis is that it provides probability of default estimates over many different horizons, where logistic regression is restricted to a just single horizon.

A masters thesis by Man (2014) develops a probability of default model using consumer and corporate data from Rabobank. Man develops an algorithm to undertake the binning of covariates using the hazard function (rather than the industry standard weight of evidence (Good, 1950)) as well as devising a method to compare predictions from a survival model and a logistic regression. The results confirms that survival models perform similarly to logistic regression, a finding repeated by other authors (see for example Stepanova and Thomas (2002)). Despite these performance similarities, Man (2014) states that survival models have certain advantages over logistic regression, specifically: (1) less data is discarded because survival analysis can utilise censored observations; and (2) logistic regression only estimates the survival probability for a fixed time interval (for example over one year).

Im et al. (2012) develop a Cox model to predict default risk for a sample of United States credit card data. They identify that macroeconomic effects have a marked impact on observed default rates, doubling due to the global financial crisis. Their method does not macro-economic variables as covariates, instead including indicator variables for each calendar quarter, an effect the authors call a "time dependency factor". The authors point out that this is not only a function of the macroeconomic effects but also includes the aggregated effects of all time-dependent factors that are not otherwise accounted for in the remaining predictor variables in the model. Their paper plots the resulting coefficients, which we reproduce in figure 3.2 below. Their model correctly detects the increase in default risk due to the financial crisis in late 2008 and the bursting of the dot-com bubble in early 2004. However the model also detects an apparent increase in the risk of default in the last quarter of 2005, but this was instead subsequently diagnosed by the authors to be the result of a change in collection policy at the bank concerned. The paper concludes with comments that this approach to modelling the macro-economic effects using survival analysis is novel, but would pose substantial challenges if the model were implemented to estimate out-of-time predictions. The authors suggest two alternate methods to counter this shortcoming, both of which use constant extrapolation beyond the in-sample training data.

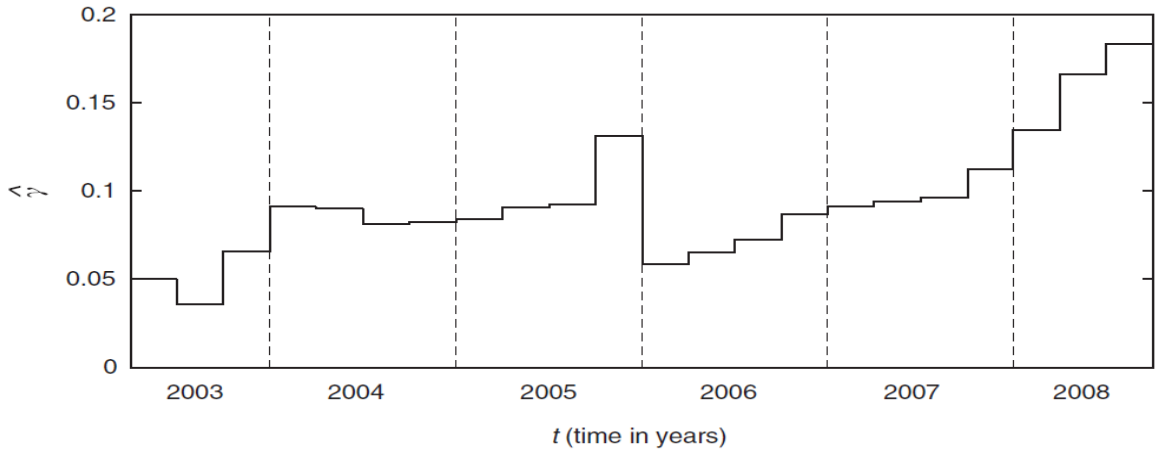

Figure 3.2: Estimated Coefficients from for US Credit Card Data Model
Source: Im et al. (2012) (figure 3)

Tong et al. (2012) constructs a mixture-cure model, which blends a logistic regression and a Cox model to estimate not only if a customer is likely to default (susceptibility) but when they are likely to default given they are susceptible (survival time). The authors state this type of approach explicitly recognises and caters for the competing risk of successful loan repayment, and has been employed previously to model long-term survival of cancer patients for two distinct subpopulations — those cured who will never relapse; and those uncured who remain susceptible to the event. The model is trained using data from a United Kingdom personal loan portfolio, and the results are compared to three other alternative models: (1) a Cox model; (2) logistic regression, targeting accounts that default during the agreed loan term; and (3) logistic regression, targeting default during 1 year cohorts every anniversary from inception (this is the standard technique used in industry for estimating probability of default). Model parameters for the mixture cure model are estimated using the expectation-maximisation (EM) algorithm (Dempster et al., 1977).

4

Maximum Likelihood Estimation for Cox Model with Time-Varying Covariates

We have discussed previously in this thesis that survival analysis is a widely applicable method of analysing censored time-to-event data. The corner-stone of survival analysis is the Cox (1972, 1975) model, which Crowley and Hu (1977) have extended to include time-varying covariates. Our literature review in Chapter 3 also demonstrates that survival analysis is an applicable method for a bank to estimate the probability of default for loans it grants to customers, with several studies having been published in the literature.

In this chapter, we develop the theory, methodology and algorithm for our maximum likelihood estimation approach for the Cox model with time-varying covariates. This technique is specifically developed to correct shortcomings of the partial likelihood estimation approach, the prevailing method used to estimate the Cox model, and adds to the literature in two important ways. The first is joint estimation of both the regression coefficients as well as the baseline hazard (introduced through a basis function) by maximising a full likelihood using a constrained optimisation algorithm that respects the non-negativity constraint of the baseline hazard. The second is the provision of a Hessian matrix that allows calculation of a variance-covariance matrix for both the estimated regression parameters as well as the baseline hazard, meaning joint inferences for all the model parameters can be drawn.

We demonstrate clearly as this chapter progresses that maximum likelihood estimation of

the Cox model with time-varying covariates poses steep computational challenges. Simply exchanging partial likelihood with maximum likelihood estimation is inappropriate unless the estimation is constrained to respect the non-negativity for the baseline hazard. As such, our solution to these computational challenges discussed in this chapter represents the novel research component of this thesis.

This chapter has six sections. The first three sections broadly focuses on the technical derivation (using calculus and matrices) for our maximum likelihood estimation approach. Section 4.1 outlines the technical details to develop the log-likelihood for the Cox model with time-varying covariates. Importantly, we detail how to include an m dimension basis function for the baseline hazard as well as stating a critical assumption that we rely upon to calculate the integral for the cumulative hazard. Section 4.2 takes this log-likelihood from section 4.1 and calculates the gradient vector and Hessian matrix. Section 4.3 defines many helpful matrices related to the gradient vector and Hessian matrix that are particularly beneficial when discussing and implementing the optimisation algorithm in R.

The last three sections broadly focusses on the algorithm and implementation of our maximum likelihood approach. Section 4.4 covers the Newton Multiplicative-Iterative algorithm that undertakes the necessary constrained maximum likelihood estimation to estimate the model parameters involved in the log-likelihood formulated in section 4.1. Section 4.5 proves that our solution using the Newton Multiplicative-Iterative algorithm converges to the correct solution. Section 4.6 discusses the detail of our R implementation.

4.1 Formulation of the Log-Likelihood

In this section, we start with the Cox model with time-varying covariates, and end with the log-likelihood we need to undertake constrained maximisation on to calculate the model parameters.

The Cox model relies on the hazard for the i^{th} subject. Including both baseline (time-invariant) and time-varying covariates, the hazard is

$$h_i(t) = h_0(t)e^{\underline{x}_i^T\underline{\beta}+\underline{z}_i^T(t)\underline{\gamma}} = h_{0i}^*(t)e^{\underline{x}_i^T\underline{\beta}} \tag{4.1}$$

where we define an adjusted baseline hazard as $h_{0i}^*(t) = h_0(t)e^{\underline{z}_i^T(t)\underline{\gamma}}$. Because it has infinite dimension, we introduce a set of m basis functions for the baseline hazard, so that

$$h_0(t) = \sum_{u=1}^{m} \theta_u \psi_u(t). \tag{4.2}$$

The baseline cumulative hazard follows

$$H_0(t) = \sum_{u=1}^{m} \theta_u \Psi_u(t) \tag{4.3}$$

where $\int_0^t \psi_u(s)ds = \Psi_u(t)$. From equation (4.3), the cumulative hazard for the i^{th} subject is given by

$$H_i(t) = H_{0i}^*(t) e^{\underline{x}_i^T \underline{\beta}} \tag{4.4}$$

where

$$H_{0i}^*(t) = \int_0^t h_{0i}^*(s)ds = \int_0^t h_0(s) e^{\underline{z}_i^T(s)\underline{\gamma}} ds. \tag{4.5}$$

Recall that the log-likelihood, as derived in equation (3.9) is

$$l(\underline{\beta}, \underline{\gamma}, \underline{\theta}) = -\sum_{i=1}^{n} H_i(t_i) + \sum_{i=1}^{n} \delta_i \log(h_i(t_i)). \tag{4.6}$$

The next step is to substitute the hazard $h_i(t)$ from equation (4.1) and the cumulative hazard $H_i(t)$ from equation (4.4) into this log-likelihood in equation (4.6). However first we need to perform the integral in equation (4.5), which requires knowledge of the functional form of $e^{\underline{z}_i^T(t)}$.

To do this, we use of the same assumption employed when using partial likelihood to estimate the Cox model with time-varying covariates. Specifically we assume that the value of the time-varying covariates remains constant between the observed event times in our sample data, and we therefore only require the value of time-varying covariates at the observed event times (Therneau et al. (2015)). This assumes that for each subject i, there will be j_i constant values for each of the q time-varying covariates from time $t = 0$ to the observed event time $t = t_i$, measured at the intervening times: $r_{i1}, r_{i2}, ..., r_{ij_i-1}, r_{ij_i} = t_i$.

We elaborate by taking a brief aside to discuss a simple example. Table 4.1 shows a small dataframe with three subjects (numbered 1, 2 and 3) who have three event times (which occur at time 3, 1 and 2 for each subject respectively) at which time each subject had an event of interest (signified by the value of status equal to 1). This gives observed event times of $t_1 = 3$, $t_2 = 1$ and $t_3 = 2$. Because there are three event times in our sample, each subject requires up to three records in the time-varying covariate dataframe, one corresponding to each observed event time detailing the values of all the q time-varying covariates, but only until such time as the subject exits the study (either from an event of interest or by being censored). For subject 1, this requires $j_1 = 3$ records in the dataframe because it leave the study at time $t = 3$. For subject 2, this requires only $j_2 = 1$ record in the dataframe, because it leaves the study at time $t = 1$. For subject 3, $j_3 = 2$ records are required in the datafame because it

leaves the study at time $t = 2$. At every one of these intermediate j_i times, the values of each time-varying covariate is assumed to be known and is required by our maximum likelihood estimation approach.

Subject	Time	Status	r_{ij_i}	$z_1(r_{ij_i})$	...	$z_q(r_{ij_i})$
1	1	0	1	5		1
	2	0	2	4		0
	3	1	3	2		1
2	1	1	1	8		0
3	1	0	1	3		1
	2	1	2	2		1

Table 4.1: Example Time-Varying Covariate Data Frame

Ending the aside, we apply the assumption that the values $\underline{z}_i^T(t)$ are constant between the time periods $[r_{ik}, r_{i(k+1)})$ to equation (4.5). This results in

$$\begin{aligned} H_{0i}^*(t) = &\int_0^{r_{i1}} h_0(s) e^{\underline{z}_i^T(s)\underline{\gamma}} ds \\ &+ \int_{r_{i1}}^{r_{i2}} h_0(s) e^{\underline{z}_i^T(s)\underline{\gamma}} ds \\ &+ \dots \\ &+ \int_{r_{ij_i-1}}^{r_{ij_i}} h_0(s) e^{\underline{z}_i^T(s)\underline{\gamma}} ds \end{aligned} \tag{4.7}$$

and allows us to bring the term $e^{\underline{z}_i^T(t)\underline{\gamma}}$ outside each of the integrals, resulting in

$$\begin{aligned} H_{0i}^*(t) \approx &H_0(r_{i1}) e^{\underline{z}_i^T(r_{i1})\underline{\gamma}} \\ &+ \left(H_0(r_{i2}) - H_0(r_{i1})\right) e^{\underline{z}_i^T(r_{i2})\underline{\gamma}} \\ &+ \dots \\ &+ \left(H_0(r_{ij}) - H_0(r_{ij-1})\right) e^{\underline{z}_i^T(r_{ij})\underline{\gamma}}. \end{aligned} \tag{4.8}$$

Substituting equations (4.1), (4.4) and (4.8) into the log-likelihood given previously in equation (4.6) results in

$$l(\underline{\beta}, \underline{\gamma}, \underline{\theta}) = -\sum_{u=1}^{m} \theta_u \Psi_u^*(t) + \sum_{i=1}^{n} \delta_i \left(\ln\left(h_{0i}^*(t_i)\right) + \underline{x}_i^T \underline{\beta} \right) \tag{4.9}$$

where we define

$$\begin{aligned}\Psi_u^*(t) = \sum_{i=1}^{n} \Big[& \Psi_u(r_{i1}) e^{\underline{z}_i^T(r_{i1})\underline{\gamma}} \\ & + \left(\Psi_u(r_{i2}) - \Psi_u(r_{i1})\right) e^{\underline{z}_i^T(r_{i2})\underline{\gamma}} \\ & + \dots \\ & + \left(\Psi_u(r_{ij_i}) - \Psi_u(r_{ij_i-1})\right) e^{\underline{z}_i^T(r_{ij_i})\underline{\gamma}} \Big] e^{\underline{x}_i^T \underline{\beta}}.\end{aligned} \tag{4.10}$$

4.2 Gradient Vector and Hessian Matrix

In preparation for developing the algorithm to estimate the model parameters, we need to calculate the first, second and cross derivatives of the log-likelihood in equation (4.9) with respect to $\underline{\beta}$, $\underline{\gamma}$ and $\underline{\theta}$. Note that $\underline{\beta}$ and $\underline{\gamma}$ are $p \times 1$ and $q \times 1$ vectors of baseline and time-varying regression coefficients respectively, and that $\underline{\theta}$ is a $m \times 1$ vector of the coefficients of the baseline hazard's basis function. Combining these three vectors into a single vector $\underline{\eta} = [\underline{\beta}^T, \underline{\gamma}^T, \underline{\theta}^T]^T$, we thus need to find the following gradient vector

$$\frac{\partial l}{\partial \underline{\eta}} = \left[\frac{\partial l}{\partial \underline{\beta}^T}, \frac{\partial l}{\partial \underline{\gamma}^T}, \frac{\partial l}{\partial \underline{\theta}^T}\right]^T \tag{4.11}$$

and the following Hessian matrix

$$\frac{\partial^2 l}{\partial \underline{\eta}\, \underline{\eta}^T} = \begin{bmatrix} \frac{\partial^2 l}{\partial \underline{\beta} \partial \underline{\beta}^T} & \frac{\partial^2 l}{\partial \underline{\beta} \partial \underline{\gamma}} & \frac{\partial^2 l}{\partial \underline{\beta} \partial \underline{\theta}} \\ \frac{\partial^2 l}{\partial \underline{\gamma} \partial \underline{\beta}} & \frac{\partial^2 l}{\partial \underline{\gamma} \partial \underline{\gamma}^T} & \frac{\partial^2 l}{\partial \underline{\gamma} \partial \underline{\theta}} \\ \frac{\partial^2 l}{\partial \underline{\theta} \partial \underline{\beta}} & \frac{\partial^2 l}{\partial \underline{\theta} \partial \underline{\gamma}} & \frac{\partial^2 l}{\partial \underline{\theta} \partial \underline{\theta}^T} \end{bmatrix}. \tag{4.12}$$

We begin with finding the first derivatives for the gradient vector:

$$\begin{aligned}\frac{\partial l(\underline{\beta}, \underline{\gamma}, \underline{\theta})}{\partial \beta_c} = & -\sum_{i=1}^{n} \sum_{u=1}^{m} \theta_u \Big[\Psi_u(r_{i1}) e^{\underline{z}_i^T(r_{i1})\underline{\gamma}} \\ & + \left(\Psi_u(r_{i2}) - \Psi_u(r_{i1})\right) e^{\underline{z}_i^T(r_{i2})\underline{\gamma}} \\ & + \dots \\ & + \left(\Psi_u(r_{ij_i}) - \Psi_u(r_{ij-1})\right) e^{\underline{z}_i^T(r_{ij_i})\underline{\gamma}} \Big] e^{\underline{x}_i^T \underline{\beta}} x_{ic} \\ & + \sum_{i=1}^{n} \delta_i x_{ic} \\ & \text{for } c = 1, \dots, p;\end{aligned} \tag{4.13}$$

$$\begin{aligned}\frac{\partial l(\underline{\beta},\underline{\gamma},\underline{\theta})}{\partial \gamma_k} = &-\sum_{i=1}^{n}\sum_{u=1}^{m}\theta_u\Big[\Psi_u(r_{i1})e^{\underline{z}_i^T(r_{i1})\underline{\gamma}}z_{ik}(r_{i1})\\ &+\left(\Psi_u(r_{i2})-\Psi_u(r_{i1})\right)e^{\underline{z}_i^T(r_{i2})\underline{\gamma}}z_{ik}(r_{i2})\\ &+\ldots\\ &+\left(\Psi_u(r_{ij_i})-\Psi_u(r_{ij_i-1})\right)e^{\underline{z}_i^T(r_{ij_i})\underline{\gamma}}z_{ik}(r_{ij_i})\Big]e^{\underline{x}_i^T\underline{\beta}}\\ &+\sum_{i=1}^{n}\delta_i z_{ik}(t_i)\\ &\text{for } k=1,\ldots,q;\ \text{and}\end{aligned} \tag{4.14}$$

$$\begin{aligned}\frac{\partial l(\underline{\beta},\underline{\gamma},\underline{\theta})}{\partial \theta_u} = &-\sum_{i=1}^{n}\Big[\Psi_u(r_{i1})e^{\underline{z}_i^T(r_{i1})\underline{\gamma}}\\ &+\left(\Psi_u(r_{i2})-\Psi_u(r_{i1})\right)e^{\underline{z}_i^T(r_{i2})\underline{\gamma}}\\ &+\ldots\\ &+\left(\Psi_u(r_{ij_i})-\Psi_u(r_{ij_i-1})\right)e^{\underline{z}_i^T(r_{ij_i})\underline{\gamma}}\Big]e^{\underline{x}_i^T\underline{\beta}}\\ &-\sum_{i=1}^{n}\delta_i\frac{\psi_u(t_i)}{\sum_{a=1}^{m}\theta_a\psi_a(t_i)}\\ &\text{for } u=1,\ldots,m.\end{aligned} \tag{4.15}$$

Next, we find the second derivatives:

$$\begin{aligned}\frac{\partial^2 l(\underline{\beta},\underline{\gamma},\underline{\theta})}{\partial \beta_c \beta_f} = &-\sum_{i=1}^{n}\sum_{u=1}^{m}\theta_u\Big[\Psi_u(r_{i1})e^{\underline{z}_i^T(r_{i1})\underline{\gamma}}\\ &+\left(\Psi_u(r_{i2})-\Psi_u(r_{i1})\right)e^{\underline{z}_i^T(r_{i2})\underline{\gamma}}\\ &+\ldots\\ &+\left(\Psi_u(r_{ij_i})-\Psi_u(r_{ij_i-1})\right)e^{\underline{z}_i^T(r_{ij_i})\underline{\gamma}}\Big]e^{\underline{x}_i^T\underline{\beta}}x_{ic}x_{if}\\ &\text{for } c,f=1,\ldots,p;\end{aligned} \tag{4.16}$$

$$\begin{aligned}\frac{\partial^2 l(\underline{\beta},\underline{\gamma},\underline{\theta})}{\partial\gamma_k\partial\gamma_g} = -\sum_{i=1}^{n}\sum_{u=1}^{m}\theta_u\Big[&\Psi_u(r_{i1})e^{\underline{z}_i^T(r_{i1})\underline{\gamma}}z_{ik}(r_{i1})z_{ig}(r_{i1})\\ &+\left(\Psi_u(r_{i2})-\Psi_u(r_{i1})\right)e^{\underline{z}_i^T(r_{i2})\underline{\gamma}}z_{ik}(r_{i2})z_{ig}(r_{i2})\\ &+\ldots\\ &+\left(\Psi_u(r_{ij})-\Psi_u(r_{ij_i-1})\right)e^{\underline{z}_i^T(r_{ij_i})\underline{\gamma}}z_{ik}(r_{ij_i})z_{ig}(r_{ij_i})\Big]e^{\underline{x}_i^T\underline{\beta}}\\ &\text{for } k,g=1,\ldots,q;\ \text{and}\end{aligned} \tag{4.17}$$

$$\frac{\partial^2 l(\underline{\beta},\underline{\gamma},\underline{\theta})}{\partial\theta_u\partial\theta_v} = \sum_{i=1}^{n}\delta_i\frac{\psi_u(t)\psi_v(t)}{\left(\sum_{s=1}^{m}\theta_s\psi_s(t)\right)^2} \qquad \text{for } u,v=1,\ldots,m. \tag{4.18}$$

The final step is finding the cross-derivatives:

$$\begin{aligned}\frac{\partial^2 l(\underline{\beta},\underline{\gamma},\underline{\theta})}{\partial\beta_c\partial\gamma_k} = -\sum_{i=1}^{n}\sum_{u=1}^{m}\theta_u\Big[&\Psi_u(r_{i1})e^{\underline{z}_i^T(r_{i1})\underline{\gamma}}z_{ik}(r_{i1})\\ &+\left(\Psi_u(r_{i2})-\Psi_u(r_{i1})\right)e^{\underline{z}_i^T(r_{i2})\underline{\gamma}}z_{ik}(r_{i2})\\ &+\ldots\\ &+\left(\Psi_u(r_{ij_i})-\Psi_u(r_{ij-1})\right)e^{\underline{z}_i^T(r_{ij_i})\underline{\gamma}}z_{ik}(r_{ij_i})\Big]e^{\underline{x}_i^T\underline{\beta}}x_{ic}\\ &\text{for } c=1,\ldots,p\,,\ k=1,\ldots,q;\end{aligned} \tag{4.19}$$

$$\begin{aligned}\frac{\partial^2 l(\underline{\beta},\underline{\gamma},\underline{\theta})}{\partial\beta_c\partial\theta_u} = -\sum_{i=1}^{n}\Big[&\Psi_u(r_1)e^{\underline{z}_i^T(r_{i1})\underline{\gamma}}\\ &+\left(\Psi_u(r_{i2})-\Psi_u(r_{i1})\right)e^{\underline{z}_i^T(r_{i2})\underline{\gamma}}\\ &+\ldots\\ &+\left(\Psi_u(r_{ij_i})-\Psi_u(r_{ij_i-1})\right)e^{\underline{z}_i^T(r_{ij_i})\underline{\gamma}}\Big]e^{\underline{x}_i^T\underline{\beta}}x_{ic}\\ &\text{for } c=1,\ldots,p\,,\ u=1,\ldots,m;\ \text{and}\end{aligned} \tag{4.20}$$

$$\begin{aligned}\frac{\partial^2 l(\underline{\beta},\underline{\gamma},\underline{\theta})}{\partial\gamma_k\partial\theta_u} = -\sum_{i=1}^{n}\Big[&\Psi_u(r_{i1})e^{\underline{z}_i^T(r_{i1})\underline{\gamma}}z_{ik}(r_{i1})\\ &+\left(\Psi_u(r_{i2})-\Psi_u(r_{i1})\right)e^{\underline{z}_i^T(r_{i2})\underline{\gamma}}z_{ik}(r_{i2})\\ &+\ldots\\ &+\left(\Psi_u(r_{ij})-\Psi_u(r_{ij_i-1})\right)e^{\underline{z}_i^T(r_{ij_i})\underline{\gamma}}z_{ik}(r_{ij_i})\Big]e^{\underline{x}_i^T\underline{\beta}}\\ &\text{for } k=1,\ldots,q\,,\ u=1,\ldots,m.\end{aligned} \tag{4.21}$$

4.3 Helpful Matrix Notation

In order to help implement our model in R (a matrix language), we outline some helpful vectors and matrices.

Let $\underline{1}_n$ be an $n\times 1$ vector of 1's, and $\underline{\xi}$ be an $n\times 1$ vector with the i^{th} element set to 1 if the subject is observed to have an event of interest, and 0 otherwise. In addition, let A be an $n\times n$ diagonal matrix with diagonal elements $H_{01}^*(t_1)e^{\underline{x}_1^T\underline{\beta}},\ldots,H_{0n}^*(t_n)e^{\underline{x}_n^T\underline{\beta}}$. For the baseline (time-invariant) covariate data, let X be an $n\times p$ matrix of covariate observations

$$X=\begin{bmatrix}x_{11} & x_{12} & \ldots & x_{1p}\\ x_{21} & x_{22} & \ldots & x_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ x_{n1} & x_{n2} & \ldots & x_{np}\end{bmatrix}. \tag{4.22}$$

Then we can write

$$\frac{\partial l}{\partial\underline{\beta}} = X^T(-A\,\underline{1}_n+\underline{\xi}) \tag{4.23}$$

and

$$\frac{\partial l^2}{\partial\underline{\beta}\partial\underline{\beta}^T} = -X^TAX. \tag{4.24}$$

Similarly, let $\underline{1}_N$ be an $N\times 1$ vector of 1's, where $N=\sum_i j_i$. In addition, let $\underline{\zeta}$ be an $N\times 1$ vector with the elements $(j_1, j_1+j_2,\ldots,N)$ set to 1 if the i^{th} subject is observed to have an event of interest and zero otherwise. All other remaining elements of $\underline{\zeta}$ are set to zero. In

addition, let B be an $N \times N$ diagonal matrix

$$B = \text{diag}\begin{bmatrix} \sum_{u=1}^{m} \theta_u \Big(\Psi_u(r_{11}) - \Psi_u(r_{10})\Big) e^{\underline{z}_1^T(r_{11})\underline{\gamma}} e^{\underline{x}_1^T\underline{\beta}} \\ \sum_{u=1}^{m} \theta_u \Big(\Psi_u(r_{12}) - \Psi_u(r_1)\Big) e^{\underline{z}_1^T(r_{12})\underline{\gamma}} e^{\underline{x}_1^T\underline{\beta}} \\ \vdots \\ \sum_{u=1}^{m} \theta_u \Big(\Psi_u(r_{1j}) - \Psi_u(r_{1j-1})\Big) e^{\underline{z}_1^T(r_{1j})\underline{\gamma}} e^{\underline{x}_1^T\underline{\beta}} \\ \sum_{u=1}^{m} \theta_u \Big(\Psi_u(r_{21}) - \Psi_u(r_{20})\Big) e^{\underline{z}_2^T(r_{21})\underline{\gamma}} e^{\underline{x}_2^T\underline{\beta}} \\ \sum_{u=1}^{m} \theta_u \Big(\Psi_u(r_{22}) - \Psi_u(r_{21})\Big) e^{\underline{z}_2^T(r_{22})\underline{\gamma}} e^{\underline{x}_2^T\underline{\beta}} \\ \vdots \\ \sum_{u=1}^{m} \theta_u \Big(\Psi_u(r_{2j}) - \Psi_u(r_{2j-1})\Big) e^{\underline{z}_n^T(r_{2j})\underline{\gamma}} e^{\underline{x}_2^T\underline{\beta}} \\ \sum_{u=1}^{m} \theta_u \Big(\Psi_u(r_{n1}) - \Psi_u(r_{n0})\Big) e^{\underline{z}_n^T(r_{n1})\underline{\gamma}} e^{\underline{x}_n^T\underline{\beta}} \\ \sum_{u=1}^{m} \theta_u \Big(\Psi_u(r_{n2}) - \Psi_u(r_{n1})\Big) e^{\underline{z}_n^T(r_{n2})\underline{\gamma}} e^{\underline{x}_n^T\underline{\beta}} \\ \vdots \\ \sum_{u=1}^{m} \theta_u \Big(\Psi_u(r_{nj}) - \Psi_u(r_{nj-1})\Big) e^{\underline{z}_n(r_{nj})\underline{\gamma}} e^{\underline{x}_n^T\underline{\beta}} \end{bmatrix}. \tag{4.25}$$

For the time-varying covariate data, let Z be an $N \times q$ matrix of data where

$$Z = \begin{bmatrix} z_1(r_{11}) & z_2(r_{11}) & \dots & z_q(r_{11}) \\ z_1(r_{12}) & z_2(r_{12}) & \dots & z_q(r_{12}) \\ \vdots & \vdots & \ddots & \vdots \\ z_1(r_{1j_1}) & z_2(r_{1j_1}) & \dots & z_q(r_{1j_1}) \\ z_1(r_{21}) & z_2(r_{21}) & \dots & z_q(r_{21}) \\ z_1(r_{22}) & z_2(r_{22}) & \dots & z_q(r_{22}) \\ \vdots & \vdots & \ddots & \vdots \\ z_1(r_{2j_2}) & z_2(r_{2j_2}) & \dots & z_q(r_{2j_2}) \\ \vdots & \vdots & \dots & \vdots \\ z_1(r_{n1}) & z_2(r_{n1}) & \dots & z_q(r_{n1}) \\ z_1(r_{n2}) & z_2(r_{n2}) & \dots & z_q(r_{n2}) \\ \vdots & \vdots & \ddots & \vdots \\ z_1(r_{nj_n}) & z_2(r_{nj_n}) & \dots & z_q(r_{nj_n}) \end{bmatrix}. \tag{4.26}$$

Then we can write

$$\frac{\partial l}{\partial \underline{\gamma}} = Z^T(-B\,\underline{1}_N + \underline{\zeta}) \tag{4.27}$$

and

$$\frac{\partial^2 l}{\partial \underline{\gamma}\partial \underline{\gamma}^T} = -Z^T B Z. \tag{4.28}$$

For estimation of $\underline{\theta}$, the necessary notation is given next in section 4.4. For the Hessian matrix, the following additional definitions are helpful.

$\frac{\partial^2 l}{\partial \underline{\theta} \partial \underline{\theta}^T} = P_1^T C_1 P_1$, where P_1 is an $n \times m$ matrix with each row containing the basis function $\psi_u(t)$ for the i^{th} subject, and C_1 is an $n \times n$ diagonal matrix with elements $\xi_1/h_0(t_1)^2, \ldots, \xi_n/h_0(t_n)^2$.

$\frac{\partial^2 l}{\partial \underline{\beta} \partial \underline{\theta}^T} = X^T C_2 P_2$, where P_2 is an $n \times m$ matrix with each row containing the integral of the basis function $\Psi_u(t)$ for the i^{th} subject, and C_2 is an $n \times n$ diagonal matrix with elements $e^{\underline{x_1}^T \underline{\beta}}, \ldots, e^{\underline{x_n}^T \underline{\beta}}$. The X matrix was outlined previously in equation (4.22).

$\frac{\partial^2 l}{\partial \underline{\beta} \partial \underline{\gamma}^T} = X_{rep}^T B Z$, where X_{rep} is an $N \times p$ matrix that replicates the rows in the X matrix j_i times for each of the i subjects. The B and Z matricies were outlined previously in equations (4.25) and (4.29).

$\frac{\partial^2 l}{\partial \underline{\gamma} \partial \underline{\theta}^T} = Z^T C_3 P_3$, where C_3 is an $N \times N$ matrix that replicates the rows in C_2 j_i times for each of the i subjects. P_3 is the following $N \times m$ matrix

$$P_3 = \begin{bmatrix}
\Big(\Psi_1(r_{11}) - \Psi_1(r_{10})\Big)e^{\underline{z_1}^T(r_{11})\underline{\gamma}} & \ldots & \Big(\Psi_m(r_{11}) - \Psi_m(r_{10})\Big)e^{\underline{z_1}^T(r_{11})\underline{\gamma}} \\
\Big(\Psi_1(r_{12}) - \Psi_1(r_1 1)\Big)e^{\underline{z_1}^T(r_{12})\underline{\gamma}} & \ldots & \Big(\Psi_m(r_{12}) - \Psi_m(r_1 1)\Big)e^{\underline{z_1}^T(r_{12})\underline{\gamma}} \\
\vdots & \ddots & \vdots \\
\Big(\Psi_1(r_{1j}) - \Psi_1(r_{1j-1})\Big)e^{\underline{z_1}^T(r_{1j})\underline{\gamma}} & \ldots & \Big(\Psi_m(r_{1j}) - \Psi_m(r_{1j-1})\Big)e^{\underline{z_1}^T(r_{1j})\underline{\gamma}} \\
\Big(\Psi_1(r_{21}) - \Psi_1(r_{20})\Big)e^{\underline{z_2}^T(r_{21})\underline{\gamma}} & \ldots & \Big(\Psi_m(r_{21}) - \Psi_m(r_{20})\Big)e^{\underline{z_2}^T(r_{21})\underline{\gamma}} \\
\Big(\Psi_1(r_{22}) - \Psi_1(r_{21})\Big)e^{\underline{z_2}^T(r_{22})\underline{\gamma}} & \ldots & \Big(\Psi_m(r_{22}) - \Psi_m(r_{21})\Big)e^{\underline{z_2}^T(r_{22})\underline{\gamma}} \\
\vdots & \ddots & \vdots \\
\Big(\Psi_1(r_{2j}) - \Psi_1(r_{2j-1})\Big)e^{\underline{z_n}^T(r_{2j})\underline{\gamma}} & \ldots & \Big(\Psi_m(r_{2j}) - \Psi_m(r_{2j-1})\Big)e^{\underline{z_n}^T(r_{2j})\underline{\gamma}} \\
\vdots & \ldots & \vdots \\
\Big(\Psi_1(r_{n1}) - \Psi_1(r_{n0})\Big)e^{\underline{z_n}^T(r_{n1})\underline{\gamma}} & \ldots & \Big(\Psi_m(r_{n1}) - \Psi_m(r_{n0})\Big)e^{\underline{z_n}^T(r_{n1})\underline{\gamma}} \\
\Big(\Psi_1(r_{n2}) - \Psi_1(r_{n1})\Big)e^{\underline{z_n}^T(r_{n2})\underline{\gamma}} & \ldots & \Big(\Psi_m(r_{n2}) - \Psi_m(r_{n1})\Big)e^{\underline{z_n}^T(r_{n2})\underline{\gamma}} \\
\vdots & \ddots & \vdots \\
\Big(\Psi_1(r_{nj}) - \Psi_1(r_{nj-1})\Big)e^{\underline{z_n}(r_{nj})\underline{\gamma}} & \ldots & \Big(\Psi_m(r_{nj}) - \Psi_m(r_{nj-1})\Big)e^{\underline{z_n}(r_{nj})\underline{\gamma}}
\end{bmatrix}. \quad (4.29)$$

4.4 Newton Multiplicative-Iterative Algorithm

In constrained optimisation, the Karush–Kuhn–Tucker (KKT) conditions (Karush (1939), Kuhn and Tucker (1951)) represent first-order necessary conditions for an optimal solution (Luenberger and Ye, 2008). For our constrained optimisation problem, the KKT conditions for the estimation of $\underline{\beta}$, $\underline{\gamma}$ and $\underline{\theta}$ are

$$\frac{\partial l}{\partial \beta_c} = 0, \text{ for } c = 1, ..., p, \tag{4.30}$$

$$\frac{\partial l}{\partial \gamma_k} = 0, \text{ for } k = 1, ..., q, \text{ and} \tag{4.31}$$

$$\frac{\partial l}{\partial \theta_u} = 0 \text{ if } \theta_u > 0 \quad \text{or} \quad \frac{\partial l}{\partial \theta_u} < 0 \text{ if } \theta_u = 0, \quad \text{for } u = 1, ..., m. \tag{4.32}$$

Using the derivations of the first-order and second-order derivatives in the section 4.2, we set up the following 3-step maximisation scheme. Beginning with estimates of $\underline{\beta}^{(s)}$, $\underline{\gamma}^{(s)}$ and $\underline{\theta}^{(s)}$ at step (s), we adopt the updating algorithm:

Step 1: use $l\left(\underline{\beta}^{(s)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\right)$ to update $\underline{\beta}^{(s+1)}$ so that $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\right) \geq l\left(\underline{\beta}^{(s)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\right)$

Step 2: use $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\right)$ to update $\underline{\gamma}^{(s+1)}$ so that $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s)}\right) \geq l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\right)$

Step 3: use $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s)}\right)$ to update $\underline{\theta}^{(s+1)}$ so that $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s+1)}\right) \geq l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s)}\right)$.

This algorithm continues iterating until reaching either: (1) a desired level of tolerance in the difference between the parameter estimates at step $(s + 1)$ and step (s); or (2) a maximum number of iterations.

To estimate $\underline{\beta}$, we employ the Newton method (Luenberger and Ye, 2008). One iteration of the Newton algorithm for solving equation (4.30) for $\underline{\beta}$, starting with $\underline{\beta}^{(s)}$ and using a line-search $\omega_1^{(s)}$ which guarantees an increase in log-likelihood, is given by

$$\underline{\beta}^{(s+1)} = \underline{\beta}^{(s)} + \omega_1^{(s)} (X^T A^{(s)} X)^{-1} X^T (-A^{(s)} \underline{1}_n + \underline{\xi}). \tag{4.33}$$

We employ a near-identical application of the Newton method (Luenberger and Ye, 2008) to estimate $\underline{\gamma}$. One iteration of the Newton algorithm for solving equation (4.31) for $\underline{\gamma}$, starting with $\underline{\gamma}^{(s)}$ and using a line-search $\omega_2^{(s)}$ which guarantees an increase in log-likelihood, is give by

$$\underline{\gamma}^{(s+1)} = \underline{\gamma}^{(s)} + \omega_2^{(s)} (X^T B^{(s)} X)^{-1} X^T (-B^{(s)} \underline{1}_{\sum n_{j_i}} + \underline{\zeta}). \tag{4.34}$$

For $\underline{\theta}$, we need to solve relationship (4.32) while respecting the non-negativity constraint $\underline{\theta} \geq \underline{0}$, thus we need a constrained optimisation. For this, we develop the multiplicative-iterative (Ma, 2010) component of our algorithm. First, we set equation (4.15) to zero, and then re-write the equation such that the left-hand and right-hand sides are strictly positive, we also multiply both sides of the equation by (the non-negative) parameter θ_u, giving

$$\begin{aligned}\theta_u\left\{\sum_{i=1}^{n}\delta_i\frac{\psi_u(t_i)}{\sum_{a=1}^{m}\theta_a\psi_a(t_i)}\right\} = \theta_u\Bigg\{\sum_{i=1}^{n}\Big[&\Psi_u(r_{i1})e^{\underline{z}_i^T(r_{i1})\underline{\gamma}}\\ &+\left(\Psi_u(r_{i2})-\Psi_u(r_{i1})\right)e^{\underline{z}_i^T(r_{i2})\underline{\gamma}}\\ &+\ldots\\ &+\left(\Psi_u(r_{ij_i})-\Psi_u(r_{ij_i-1})\right)e^{\underline{z}_i^T(r_{ij_i})\underline{\gamma}}\Big]e^{\underline{x}_i^T\underline{\beta}}\Bigg\}.\end{aligned} \tag{4.35}$$

Given that both the right and left hand sides of equation (4.35) are positive, then the ratio in equation(4.36) is also positive. Ma (2010) suggests this ratio as a natural updating scheme that begins with the intermediate step

$$\begin{aligned}\theta_u^{(s+\frac{1}{2})} = \frac{\theta_u^s \times \left(\sum_{i=1}^{n}\delta_i\frac{\psi_u(t_i)}{h_{0i}(t_i)}\right)+\epsilon_u}{\sum_{i=1}^{n}\Big[&\Psi_u(r_{i1})e^{\underline{z}_i^T(r_{i1})\underline{\gamma}}+\left(\Psi_u(r_{i2})-\Psi_u(r_{i1})\right)e^{\underline{z}_i^T(r_{i2})\underline{\gamma}}+\ldots\\ &+\left(\Psi_u(r_{ij})-\Psi_u(r_{ij-1})\right)e^{\underline{z}_i^T(r_{ij})\underline{\gamma}}+\Big]e^{\underline{x}_i^T\underline{\beta}}+\epsilon_u}\end{aligned} \tag{4.36}$$

where ϵ_u is a small positive value included in both the numerator and denominator to avoid division by zero.

Following the details outlined in Ma et al. (2014), we extend to account for the additional set of regression parameters $\underline{\gamma}$ for the time-varying covariates. When $\theta_u^{(s)} > 0$ then updates $\theta_u^{(s+\frac{1}{2})}$ given by equation (4.36) are all non-negative. Moreover, if $\theta_u^{(s)} > 0$, then $\theta_u^{(s+\frac{1}{2})} = 0$ only if $\sum_{i=1}^{n}\delta_i\psi_u(t_i)/h_{0i}(t_i) = 0$ and $\epsilon_u = 0$. Although it maintains $\theta_u^{(s+1)} \geq 0$, the iteration in equation (4.36) may fail to increase $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s)}\right)$, leading to possible divergence. To rectify this problem, we use a line-search step to give $\theta_u^{(s+1)}$ such that $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s+1)}\right) \geq l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s)}\right)$. First, we rewrite equation (4.36) as

$$\theta_u^{(s+\frac{1}{2})} = \theta_u^{(s)} + v_u^{(s)}\ \frac{\partial l(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s)})}{\partial\theta_u} \tag{4.37}$$

where $v_u^{(s)} = \theta_u^{(s)}/w_u^{(s)}$ with $w_u^{(s)}$ defined as the denominator in the right hand side of equation (4.36). Clearly, when $\theta_u^{(s)} \neq 0$ we have $v_u^{(s)} \geq 0$. When $\theta_u^{(s)} = 0$ we set $v_u^{(s)} = 0$ only if $\partial l(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s)})/\partial\theta_u < 0$, since $\theta_u^{(s)}$ has already satisfied the KKT condition in this

case. Otherwise, we set $v_u^{(s)} = c/w_u^{(s)}$, where c is a small constant such as 10^{-5}. Equation (4.37) means that $\theta_u^{(s+\frac{1}{2})}$ emanates from $\theta_u^{(s)}$ in the gradient direction of $\underline{\theta}$ with a non-negative size step $v_u^{(s)}$. In the context of a line search, the search direction is $\theta_u^{(s+\frac{1}{2})} - \theta_u^{(s)}$. Letting $\alpha^{(s)} > 0$ be the search step size, then $\theta_u^{(s+1)}$ is obtained using

$$\theta_u^{(s+1)} = \theta_u^{(s)} + \alpha^{(s)}(\theta_u^{(s+\frac{1}{2})} - \theta_u^{(s)}). \tag{4.38}$$

We only require that $\alpha^{(s)} \leq 1$ as this guarantees $\theta_u^{(s+1)} \geq 0$.

4.5 Convergence Properties

We now discuss the convergence properties of the Newton Multiplicative-Iterative algorithm defined in section 4.4. This discussion very closely mirrors that from Ma et al. (2014).

Theorem 1. Consider the Newton Multiplicative-Iterative Algorithm developed in section 4.4. Assume that $\underline{\beta} \in \mathbb{R}^p$ and $\underline{\gamma} \in \mathbb{R}^q$, which are respectively p and q dimensional real spaces. In addition, assume that $\underline{\theta} \in \mathbb{R}^m_{\geq 0}$ is a non-negative m-dimension real space. Let M_1, M_2 and M_3 be respectively be the iteration mappings of the algorithm defined in section 4.4, so that

(1) $\underline{\beta}^{(s+1)} = M_1(\underline{\beta}^{(s)}; \underline{\gamma}^{(s)}, \underline{\theta}^{(s)})$,

(2) $\underline{\gamma}^{(s+1)} = M_2(\underline{\gamma}^{(s)}; \underline{\beta}^{(s+1)}, \underline{\theta}^{(s)})$, and

(3) $\underline{\theta}^{(s+1)} = M_3(\underline{\theta}^{(s)}; \underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)})$.

Assume that the matrix A (which is a function of $\underline{\beta}$) and B (which is a function of $\underline{\gamma}$) as defined in section 4.2, both satisfy the condition that $A^{1/2}X$ and $B^{1/2}Z$ have full column rank for $\underline{\beta} \in R^p$ and $\underline{\gamma} \in \mathbb{R}^q$ respectively and that $\underline{\theta} \in \mathbb{R}^m_{\geq 0}$. With starting values $\underline{\beta}^{(0)} \in \mathbb{R}^p$ and $\underline{\gamma}^{(0)} \in R^q$ and $\underline{\theta}^{(0)} \in \mathbb{R}^m_{\geq 0}$, then the algorithm produces a sequence $\{\underline{\beta}^{(s)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\}$ which converge to a solution satisfying the KKT conditions outlines in equations (4.30, 4.31, 4.32).

The **proof** is available in appendix A.1.

We make the following comments regarding the Newton Multiplicative-Iterative Algorithm developed in section 4.4.

1. From equation (4.24) and equation (4.28), $\frac{\partial l^2}{\partial \underline{\beta} \partial \underline{\beta}^T}$ and $\frac{\partial l^2}{\partial \underline{\gamma} \partial \underline{\gamma}^T}$ are positive definite if both $A^{1/2}X$ and $B^{1/2}Z$ have full column rank for $\underline{\beta} \in R^p$ and $\underline{\gamma} \in R^q$ and $\underline{\theta}^{(0)} \in \mathbb{R}^m_{\geq 0}$.

2. We use an indicator basis function for $\psi(t)$, such that the width for each basis encompasses (approximately) the same fixed number of observed events.

3. Our algorithm converges quickly for small to moderate sized datasets, typically in a few seconds for about $10,000$ subjects and about 10 variables.

4.6 Model Implementation in R

This chapter's previous five sections stepped through several important components needed to implement our model. It started with the setup of the log-likelihood (section 4.1), calculation of the gradient vector and Hessian matrix (section 4.2) as well as additional helpful matrix notation (section 4.3). It next developed a Newton Multiplicative-Iterative algorithm to undertake constrained optimisation for parameter estimation (section 4.4), and showed it converges to the correct solution (section 4.5). In the final section of this chapter, we discuss the model implementation of the Newton Multiplicative-Iterative Algorithm for constrained optimisation. We implement our model in the R programming language (R Core Team, 2016), with the full computer code provided in appendix B.1.

A key requirement of our implementation is the need for two input dataframes: one each for the baseline and time-varying covariates. The best manner in which to demonstrate this is via a small worked example. Suppose we have $n = 3$ subjects, for which have $p = 2$ baseline and $q = 2$ time-varying covariates. Suppose further that for our data there are three event times at $t = 3$, 1 and 2 respectively for subjects 1, 2 and 3. The first of the two input dataframes governs baseline covariates, and contain one row for each subject. The columns detail the subject identifier, event time, and status (1 for observed event, 0 otherwise). The remaining $p = 2$ columns contain the data for the baseline covariates. This is demonstrated in table 4.2 below.

Subject	Time	Status	x_1	x_2
1	3	1	104	1
2	1	1	55	0
3	2	1	23	1

Table 4.2: An Example Baseline Dataframe

The second of the two input dataframes governs the time-varying covariates, containing multiple rows for each subject (as outlined previously in section 4.1). Based on the unique observed event times from the data (in our case, these are the times $t = 1$, 2 and 3), each subject will have a record for each of these observed event times, up until when the given subject leaves the study. Table 4.3 demonstrates this for our example. Subject $i = 1$ has an

event time of $t_1 = 3$, so it has three records in the dataframe. Subject $i = 2$ has an event time of $t_2 = 1$, so it has only one record while subject $i = 3$ has an event time of $t_3 = 2$, so it has two records. The columns for this dataframe begin with a subject identifier, as well as a variable recording the j_i time-points (r_{ij_i}) when the time-varying covariates are measured for each of the i subjects. The final $q = 2$ columns contain the values of the time-varying covariates. Note that the key feature of this dataframe is that the values of the time-varying covariates can potentially (but not necessarily) change over time for a given subject.

Subject	r_{ij_i}	$z_1(r_{ij_i})$	$z_2(r_{ij_i})$
1	1	5	1
1	2	4	0
1	3	2	1
2	1	8	0
3	1	3	1
3	2	2	1

Table 4.3: Example Time-Varying Covariate Data Frame

A key difficulty to overcome in the model implementation is how to undertake calculations for this expanded time-varying covariate dataframe. One natural method to do this in R is to separate the Z matrix into a list of n matrices, one each for every individual subject, and use functions such as `lapply` and `mapply` to undertake the necessary calculations. To demonstrate this for the data in tables 4.2 and 4.3, the baseline covariate matrix (X) and time-varying covariate matrix (Z), as defined (in the general case) earlier in section 4.2 are

$$X = \begin{bmatrix} x_{11} = 104 & x_{12} = 1 \\ x_{21} = 55 & x_{22} = 0 \\ x_{31} = 23 & x_{32} = 1 \end{bmatrix} \tag{4.39}$$

and

$$Z = \begin{bmatrix} z_1(r_{11}) = 5 & z_2(r_{11}) = 1 \\ z_1(r_{12}) = 4 & z_2(r_{12}) = 0 \\ z_1(r_{13}) = 2 & z_2(r_{13}) = 1 \\ z_1(r_{21}) = 8 & z_2(r_{21}) = 0 \\ z_1(r_{31}) = 3 & z_2(r_{31}) = 1 \\ z_1(r_{32}) = 2 & z_2(r_{32}) = 1 \end{bmatrix}. \tag{4.40}$$

Our implementation separates the time-varying covariate matrix Z into a list of matrices $Z.list = [Z_1, Z_2, Z_3]$, with one component matrix for each subject defined as

$$Z_1 = \begin{bmatrix} z_1(r_{11}) = 5 & z_2(r_{11}) = 1 \\ z_1(r_{12}) = 4 & z_2(r_{12}) = 0 \\ z_1(r_{13}) = 2 & z_2(r_{13}) = 1 \end{bmatrix}, \tag{4.41}$$

$$Z_2 = \begin{bmatrix} z_1(r_{21}) = 8 & z_2(r_{21}) = 0 \end{bmatrix}, \text{ and} \tag{4.42}$$

$$Z_3 = \begin{bmatrix} z_1(r_{31}) = 8 & z_2(r_{31}) = 0 \\ z_1(r_{31}) = 3 & z_2(r_{31}) = 1 \\ z_1(r_{32}) = 2 & z_2(r_{32}) = 1 \end{bmatrix}. \tag{4.43}$$

Again, the full model implementation R code is available in appendix B.1, but the below R code shows how we split the Z matrix into the required list of matrices

```
# Make a list storing one z matrix for each subject
Z.list=lapply(split(Z[,var.z],Z[,1]),FUN=as.matrix).
```

An identical process is used to separate other matrices into lists of matrices, for example the R code below shows how to create the a list of n matrices for the basis function $\psi(t)$

```
# Make a list storing one phi matrix for each subject
psi=lapply(Zsplit.time,
FUN=basis_mpl,
knots=knots,
basis=control$basis,
order=control$order,
which=1).
```

This approach works well but given R is memory resident, this can cause some performance issues. For example, testing our implementation using a $12 \times 100{,}000$ baseline dataframe and a $2 \times 5{,}000{,}000$ time-varying dataframe takes 20 minutes and about 20 iterations to converge to a tolerance of 1×10^{-4}. We use an Intel i5 CPU with 3.2GHz, a solid-state SD hard-drive and 32GB of SD3-RAM. An alternate implementation could use a single matrix but control subject identification in R using indices. We plan this as an avenue for future research.

As the user interface to our model implementation we create the `cox_mle_tvc()` function, which will jointly estimate regression coefficients and the baseline hazard of the Cox model with time-varying covariates. The function is called in the following manner:

cox_mle_tvc (formula , data , formula.z , data.z , riji , subject , control , ...)

A call to the *cox_mle_tvc()* function requires the following six mandatory arguments, as well as allowing for additional optional arguments which are passed to the *cox_mle_tvc.control()*:

formula A formula object, with the response on the left of a ~ operator, and baseline covariates on the right separated by a "+" sign. The response must be a survival object as returned by the *survival::Surv()* function. A value of status=1 signifies that subject i was observed to have an event, while status=0 signifies the subject was censored. For example: *Surv (time , status)* ~ $x1 + x2$.

data The baseline data.frame that contains the baseline covariate information, as well as the event time and status indicator used in the formula object. The dataframe also requires each entry to be signified by a unique subject identifier for each subject i.

formula.z A linear predictor for the time-varying covariates. This should not be a formula object, but covariates need to be preceded by a ~ operator prior to listing the time-varying covariates on the right separated by a "+" sign. For example: ~ $z1 + z2$.

data.z The time-varying covariate data.frame, that has been expanded so that there is one record for every event time for every subject, up to the time that the subject leaves the study. For subject i, there are j_i records in the dataframe. One method to achieve this is to use the function *survival::survSplit()*

riji The j_i time-points at which the time varying covariates are observed for subject i.

subject The unique identifier for subject i.

... Other (optional) arguments which are passed to the control function $cox\_mle\_tvc.control()$.

The function returns an object of the class cox_mle_tvc, which is a list containing the results of the fitting algorithm. Example code to call the function is

```
fit.MPLt <– cox_mle_tvc(formula = Surv(time, status) ~ x1 + x2,
                        data = baseline,
                        formula.z = ~ z1 + z2,
                        riji = time_reps,
                        subject = id,
                        data.z = time_varying).
```

The *cox_mle_tvc.control()* function has the following optional arguments that allow the user to control various aspects of the model fit. Any arguments not supplied by the user will be

assigned the indicated default values.

max.iter The maximum number of iterations for the Newton Multiplicative-Iterative algorithm. The default value is 10,000.

n.events_basis The number of observed events to include in each basis of the baseline hazard. With n.obs representing the number of observed events, the default is calculated using round(3.5 log(n.obs) - 7.5).

tol The convergence tolerance value, which is the smallest change in the parameter estimates between iterations that when achieved indicates convergence has occurred. The default value is 1×10^{-6}.

min.theta The size for the individual elements in the $\underline{\theta}$ vector that are considered indistinguishable from zero. The default value is 1×10^{-10}.

kappa The initial step-size in the line search. The default value is 0.6.

5 Results

This chapter comprises two sections, each containing a test problem on which we evaluate and apply our method. The first test problem covers a simulation study, where we compare our maximum likelihood method against the partial likelihood method for estimating the parameters of the Cox model with time-varying covariates. We use this simulation study to focus solely on comparing the ability of the methods to recover regression coefficients across different sample sizes and censoring proportions. We don't use this simulation study to compare the ability of the methods to estimate baseline hazards or provide a full variance-covariance matrix for the model parameters, instead leaving this to the second section of this chapter.

The second section of this chapter applies both the maximum likelihood and partial likelihood methods to the second test problem, which models the time to default for home loan data. We compare the regression coefficients but additionally we also compare the estimated baseline hazard from the maximum likelihood method to that estimated using the Breslow (1972) estimator, which relies on regression coefficients estimated using the partial likelihood. We also discuss, using the delta method (Xu and Scott-Long, 2005), how the full variance-covariance matrix can be used to construct point-wise confidence intervals for survival probabilities for non-baseline subjects.

5.1 Test Problem 1: Simulation Study

For the first test problem for our maximum likelihood method, we simulate survival data from a distribution with the hazard function

$$h_i(t) = h_0(t)e^{\beta x_i + \gamma z_i(t)} \tag{5.1}$$

where we select the true population effects to be known values of $\beta = -3.3$ and $\gamma = 4.0$. The baseline covariate x is a continuous variable while the time-varying covariate $z(t)$ is a discrete variable. We simulate x and $z(t)$ using the `genTDCM` function available in the `R` package `genSurv` (Araujo et al., 2015). The `R` code to draw one sample is

```
# Draw a sample of 2000 for survival data with time-varing covariate
tdcmdata <- genTDCM(n=2000, dist="weibull", corr=1,
                    dist.par=c(2,3,2,3),model.cens="uniform",
                    cens.par=2.5, beta=c(-3.3,4), lambda=1).
```

Our simulation study involves drawing $M = 500$ Monte Carlo simulations across two sample sizes ($n = 100$ and $n = 2000$) and two approximate censoring proportions ($\pi = 20\%$ and $\pi = 80\%$). For each of these, we estimate the baseline and time-dependent effects using our maximum likelihood method ($\hat{\beta}_{ML}$, $\hat{\gamma}_{ML}$) and compare these to estimates from the partial likelihood method ($\hat{\beta}_{PL}$, $\hat{\gamma}_{PL}$). Uninformative right censoring times are drawn from a uniform distribution whose upper bound is selected to approximately achieve the desired sample censoring proportion. The baseline hazard is assumed to be exponential distributed with rate = 1. Table 5.1 below summarises the bias, standard deviation (Sd) and mean-square error (MSE) for the regression coefficients calculated from the simulation study.

			n=100 $\pi = 20\%$	n=100 $\pi = 80\%$	n=2000 $\pi = 20\%$	n=2000 $\pi = 80\%$
$\hat{\gamma}$	PL	Bias	0.909	1.49	0.006	0.007
		Sd	3.915	4.808	0.15	0.18
		MSE	16.122	25.295	0.022	0.033
	ML	Bias	0.153	0.832	0.044	0.046
		Sd	1.089	1.555	0.153	0.181
		MSE	1.207	3.104	0.025	0.035
$\hat{\beta}$	PL	Bias	-0.056	-0.258	-0.006	-0.022
		Sd	0.333	0.866	0.061	0.118
		MSE	0.114	0.815	0.004	0.014
	ML	Bias	-0.091	-0.443	-0.017	-0.045
		Sd	0.293	0.84	0.062	0.117
		MSE	0.094	0.901	0.004	0.016

Table 5.1: Comparing the Estimated Effects for $\hat{\gamma}$ and $\hat{\beta}$ in a Simulation Study using Maximum Likelihood (ML) and Partial Likelihood (PL) Estimation

The results show that while both methods can recover the effect for the baseline covariate (β) with comparable accuracy, the maximum likelihood method has superior accuracy

recovering the effect for the time-varying covariate (γ) for $n = 100$. This is evidenced by the bias, standard deviation and mean-square error all being smaller for the maximum likelihood estimation of $\hat{\gamma}$. Of particular note are the mean-square error for $\hat{\gamma}$ which are between eight and twelve times greater for the partial likelihood approach than for the maximum likelihood approach.

These conclusions are further substantiated in the histograms in figures 5.1 and 5.2, which compare the simulated results for $\hat{\gamma}$ with $n = 100$. Overlaid are vertical lines representing the true effect of $\gamma = 4.0$ (blue line) and the calculated of mean (red line) of $\hat{\gamma}$ from the $M = 500$ sample simulation study. Across both sampling proportions (of 20% and 80%) the histograms confirm that the maximum likelihood estimation recovers the true population parameter for the time-varying covariate more accurately and with lower bias than the partial likelihood method.

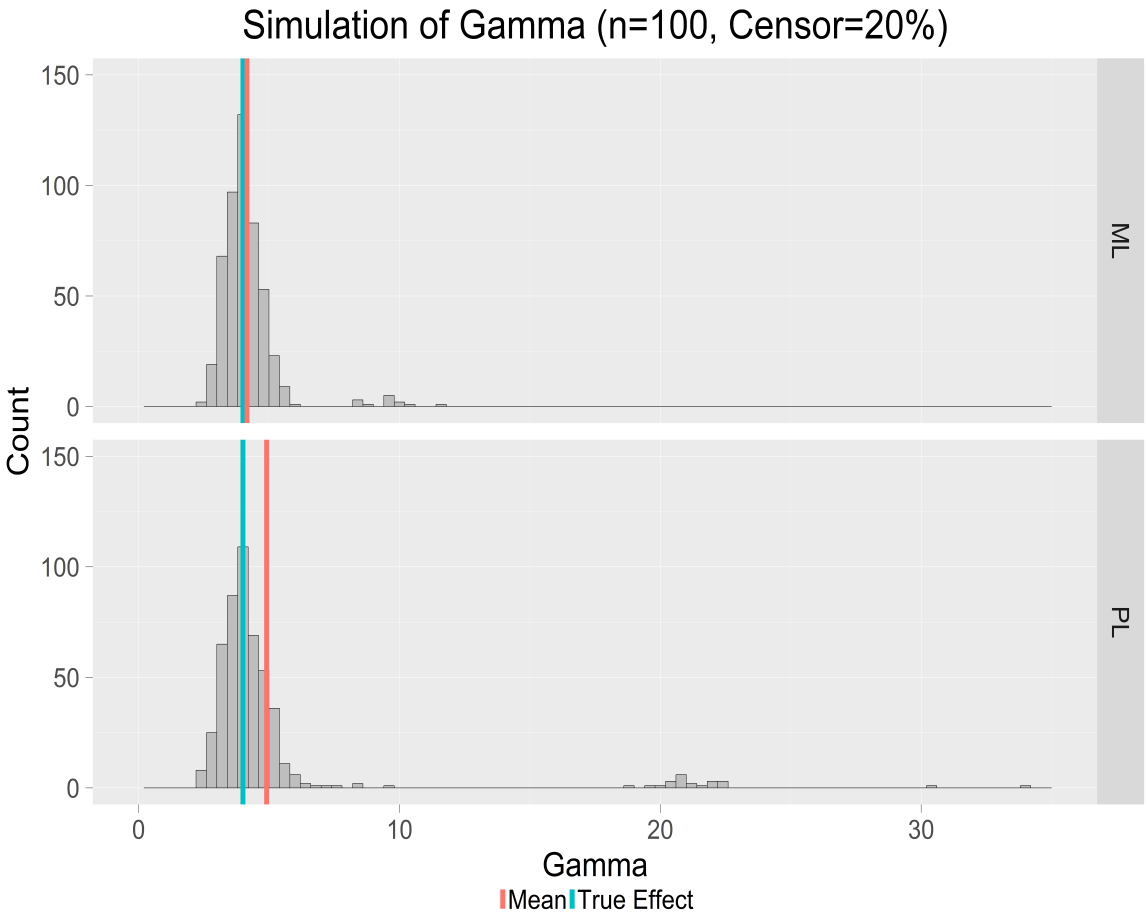

Figure 5.1: Histogram of Simulation Results for $\hat{\gamma}$ (with a censoring proportion of 20%), Comparing Maximum Likelihood (ML - upper panel) and Partial Likelihood (PL - lower panel) Estimation

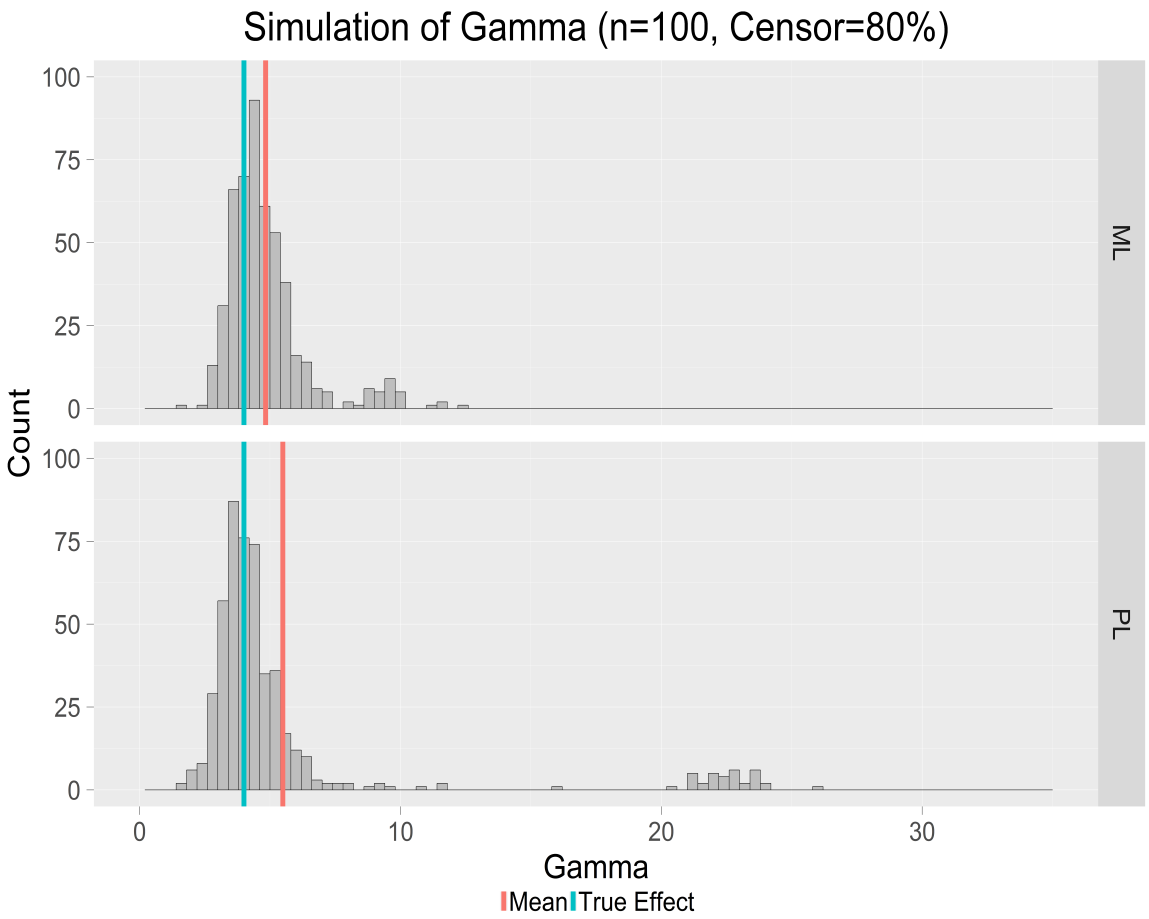

Figure 5.2: Histogram of Simulation Results for $\hat{\gamma}$ (with a censoring proportion of 80%), Comparing Maximum Likelihood (ML - upper panel) and Partial Likelihood (PL - lower panel) Estimation

5.2 Test Problem 2: Application to Credit Risk Data

The second test problem for our maximum likelihood method is to compare it against the partial likelihood method using an applied dataset. We use a credit risk dataset consisting of a sample of approximately 10 years of defaulted and non-defaulted home loans. The data contains approximately $n = 100,000$ loans with a censoring rate of over 98%. The event of interest is the default at any time during the life for each home loan. For this sample, left-truncated observations are removed.

Data validation and filtering is conducted using SAS 9.3 (SAS Institute Inc, 2016). We construct a model with the following 11 baseline covariates and 2 time-varying covariates which we briefly introduce below (note that covariates listed with a * have been mean-corrected).

1. Mortgage Insured – Does the home loan have lenders mortgagee insurance? (Yes /

No)?

2. Borrower's Occupation – Professional / Trades / Sales / Other
3. Borrowers – How many borrowers are there for this home loan (either 1 or 2+)?
4. Salary Credits – Do the borrowers deposit salary directly into the home loan account (Yes/No)?
5. Credit Card – Do the borrowers have a credit card? (Yes/No)
6. Personal Loan – Do the borrowers have a personal loan? (Yes/No)
7. Repayment Method – Are repayments principle and interest (P&I) or interest only (IO)?
8. Repayment Frequency – Are repayments made monthly or fortnightly/other
9. Borrower Tenure* – How many months has the borrower been a customer of the bank (integer 0,1,2...)?
10. Total Home Loans* – Total number of home loans the borrowers have (integer 1, 2, ...)
11. Opening Balance* – The original balance when the home loan opened ($millions)
12. Dynamic Loan to Value Ratio* – The current loan balance divided by the current estimate of the value of the home securing the loan. This is a time-varying covariate, as both the numerator and denominator of this ratio vary over time.
13. Worst Delinquency in Last 6 Months – The highest number of missed monthly payments within the last 6 months, lagged by a period of 12 months. This variable can take integer values of 0, 1, 2, ... This is a time-varying covariate because delinquency status (the number of repayments behind a customer is) definitionally begins at zero at origination but can change to values greater than zero over time.

Prior to presenting results of the model fitting, we add some general comments regarding these covariates. The first 11 covariates are baseline covariates, which means that their values are measured only at the beginning of the loan and their values either: (1) don't change over time (for example, whether lenders mortgagee insurance is in force, or the number of borrowers for a home loan); or (2) do change over time but their values are not tracked (for example, the borrower's occupation, or other product holdings). The final two variables are time-varying covariates, whose values are tracked and potentially change over time.

Moving to the model results, table 5.2 displays that the parameter estimates from the maximum likelihood and partial likelihood methods. These results show that the estimated

regression coefficients and standard errors from the models are very similar, and both result in intuitive parameter estimates. That is, positive parameter estimates suggest an increasing risk of default while negative parameter estimates suggest a decreasing risk of default.

		Partial Likelihood		Maximum Likelihood	
Name	Level	Estimate	Std Err	Estimate	Std Err
Mortgage Insured	Yes	0.1615	0.0557	0.1599	0.0556
Borrower's Occupation	Profession	−0.5542	0.0647	−0.5543	0.0647
	Trades	0.0036	0.0595	0.0034	0.0595
	Other	−0.3496	0.0953	−0.3478	0.0953
	Sales	0	–	0	–
Borrowers	2+	−0.3184	0.0484	−0.3183	0.0484
Salary Credits	Yes	−0.3240	0.2123	−0.3225	0.2123
Credit Card	Yes	−0.2775	0.0515	−0.2778	0.0515
Personal Loan	Yes	0.2666	0.1019	0.2670	0.1019
Repayment Method	IO	0.1720	0.0522	0.1729	0.0522
Repayment Frequency	Monthly	0.1728	0.0600	0.1715	0.0560
Customer Tenure (Months)		−0.0021	0.0003	−0.0021	0.0003
Total Home Loans (Count)		−0.1112	0.0186	−0.1111	0.0186
Opening Balance ($millions)		0.5131	0.1183	0.5124	0.1183
Dynamic Loan to Value Ratio	(t)	2.7924	0.1303	2.7917	0.1303
Worst Delinquency in Last 6 Months	(t-12)	2.9979	0.06328	2.9983	0.0632

Table 5.2: Comparison of Parameter Estimates of the Eleven Baseline and Two Time-Varying Covariates Using Maximum Likelihood and Partial Likelihood Estimation

To help interpret the estimated effects from the fitted models, we group them into 6 categories of risk drivers and make the following conclusions.

1. Contractual – Mortgage insured home loans have a higher risk of default. This is sensible as typically banks require borrowers whose loans have high loan-to-value ratio to obtain lenders mortgagee insurance.

2. Borrower – The riskiest occupations are trades closely followed by sales with professionals being the least risky. This may be detecting "professionals" have more stable income than "trades" and "sales".

3. Depth of Relationship – Home loans held by two or more borrowers, borrowers with multiple home loans, borrowers who have longer affiliation, borrowers who deposit salary into their home loan and borrowers with credit cards but not personal loans all have lower risk. Borrowers with whom the bank have a deep and enduring relationship with are lower risk.

4. Self Selection – Borrowers who elect to pay interest only, or repay monthly have higher risk as these borrowers are less likely to prepay their loans

5. Financial – Home loans with higher original balances are higher risk. Home loans with a higher loan to (house) value ratio have a higher risk, which is an effect that the models caters for changes in over time.

6. Credit quality – Home loans that have been delinquent in the past are much higher risk, which is an effect that the models caters for changes in over time.

We note that the two most important drivers of the model are the time-varying covariates. Dynamic loan to value ratio and delinquency are the most significant variables in the model with z-test statistics respectively of 21.46 (2.79/0.13) and 47.46 (2.99/0.063). This is a fantastic applied example of the power and benefit of using time-varying covariates in a survival model.

The next aspect of comparison for this test problem is to compare the baseline hazards. The previously described shortcoming of the partial likelihood method is that it does not produce an estimate of the baseline hazard. This is precisely one of the key benefits of our maximum likelihood approach, as it jointly estimates regression coefficients and baseline hazard.

In order to compare our estimate of the baseline hazard between the maximum likelihood and partial likelihood methods, we use the Breslow (1972) estimator which relies on partial likelihood regression coefficients as input. Figure 5.3 compares results of the Breslow estimator (blue line); overlaid is the baseline estimate from the maximum likelihood method (red line) along with the associate 95% confidence interval (grey shaded area).

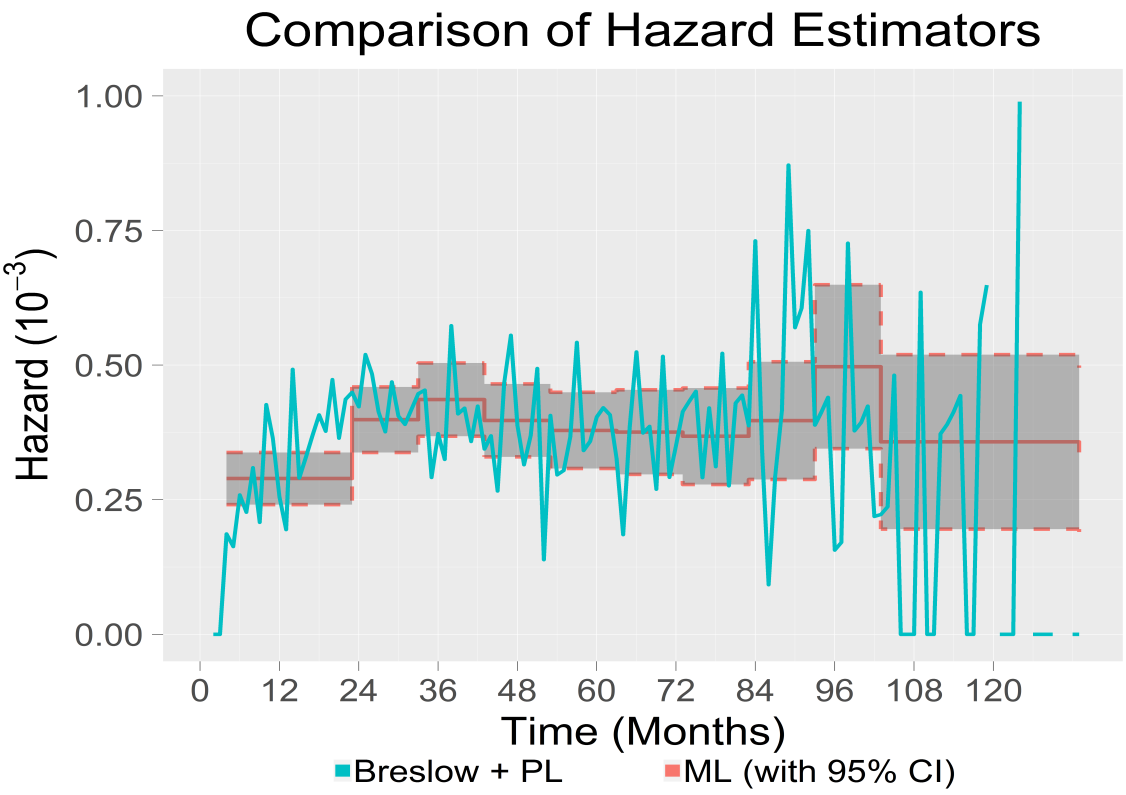

Figure 5.3: Comparison of Baseline Hazard Using:
"Breslow + Partial Likelihood (PL)" verses "Maximum Likelihood (ML)" Estimation

While in general the two estimates tend to follow each other, it is very clear that the

Breslow estimate exhibits a substantially larger degree of volatility than the estimate from the maximum likelihood method. Interpreting the hazard in this applied credit risk setting, we observe that the risk of default begins at a low level up to 24 months from origination, then increases and somewhat plateaus for the remaining time.

The final aspect of comparison is to discuss the additional information available in the variance-covariance matrix using maximum likelihood compared to using partial likelihood. Using maximum likelihood we have estimated $p+q = 15$ regression coefficients (including the additional dummies for categorical variable "borrower occupation") as well as simultaneously estimating $m = 11$ values for $\underline{\theta}$ from the baseline hazard. The resulting 27×27 variance-covariance matrix captures the joint sampling variation of these 27 parameters. On the other hand the partial likelihood only produces a 15×15 variance-covariance matrix, capturing only the joint variation in the regression parameters, and not that of the baseline hazard. Appendix A.2 provides the numerical results for both of these variance-covariance matrices for the home loan data.

Both estimation approaches provide enough information to conduct hypothesis tests and contrasts for the regression coefficients. In addition, there are methods to calculate the standard error of the Breslow baseline hazard estimator (see for example (Kalbfleisch and Prentice, 2002)). Hence both the partial likelihood and maximum likelihood approaches can provide point-wise confidence intervals for the survival function for a baseline subject. However the reduced amount of information available in the variance-covariance matrix obtained under the partial likelihood approach, means it cannot provide confidence intervals for non-baseline subjects. Our maximum likelihood approach corrects this. Appendix A.2 provides further details, but given that the maximum likelihood estimates are asymptotically normally distributed, the confidence intervals can be obtained using the Delta method (Xu and Scott-Long, 2005). Recall that we defined in section 4.2 that $\underline{\beta}$, $\underline{\gamma}$ and $\underline{\theta}$ can be combined into a single vector $\underline{\eta}^T = [\underline{\beta}^T, \underline{\gamma}^T, \underline{\theta}^T]$. Noting that the survival function for a non-baseline subject is a function of both time and $\underline{\eta}$, the Delta method provides

$$Var\left[S(t, \underline{\eta})\right] = \frac{\partial S(t, \underline{\eta})}{\partial \underline{\eta}^T} Var(\underline{\eta}) \frac{\partial S(t, \underline{\eta})}{\partial \underline{\eta}} \tag{5.2}$$

where $\partial S(t, \underline{\eta}) \backslash \partial \underline{\eta}$ is the derivative of the survival function with respect to $\underline{\eta}$. We have not undertaken this work for this thesis, but calculating the variance for non-baseline subjects in this manner remains an avenue of future research, particularly for our R implementation.

6

Conclusion and Discussion

Survival data arises when the random variable under study is the time to an event of interest. Regression techniques have gained popularity as a statistical method to explain variation in survival times using available predictive covariates (Hosmer et al., 2008). For survival analysis, the Cox partial likelihood model (1972, 1975), including its extension by Crowley and Hu (1977) to cater for time-varying covariates, has become the favoured regression technique (Ren and Zhou, 2011). While there are potentially a variety of wide applications for survival analysis, its use has predominantly used in biomedical science and industrial life testing (Kalbfleisch and Prentice, 2002).

Survival analysis (and particularly the Cox model) is a widely adopted method in modelling probability of default by banks (Lessmann et al., 2015), who are interested in understanding the probability that a customer will fail to repay in a timely manner the monies they contractually owe (including principle, interest and fees). An estimate of probability of default is a key input for banks to calculate their minimum capital required under the Basel Accords (BIS, 2006). In the context of estimating probability of default, survival models allow an additional benefit in that they can estimate not only if but also when a customer is likely to default (see fore example: Bellotti and Crook (2009), Stepanova and Thomas (2002) and Tong et al. (2012)), a particularly useful aspect which can be leveraged to estimate the probability of default over multiple horizons rather than simply a single fixed horizon as well as allowing for time-varying covariates. In conjunction with allowing for time-varying covariates, these advantages of survival analysis could both prove useful in estimating the

lifetime probability of default, a key input for banks to calculate their expected credit losses required under International Financial Reporting Standard (IFRS) 9 accounting standard (IASB, 2014).

Despite its many applications, the much famed partial likelihood method used to estimate the Cox model with time-varying covariates contains two distinct short comings: (1) the baseline hazard is not estimated by the model, so that recovery of survival probabilities requires a further estimation step after fitting, such as that provided by either Breslow (1972) or Kalbfleisch and Prentice (2002); and (2) the partial likelihood does not produce a covariance matrix for both fitted parameters and the baseline hazard, meaning joint inferences of the model parameters cannot be made.

This thesis has developed a new methodology to address these two shortcomings. It does this by simultaneously estimating using maximum likelihood both regression coefficients and the baseline hazard for survival data whose sample design include uninformative right-censoring and time-varying covariates. Our approach adds to the literature by: (1) estimating model parameters using maximum likelihood; and (2) providing an estimate of the baseline hazard using a piece-wise constant basis which removes reliance on a secondary estimation step. We develop the necessary theory to estimate our model, including gradient vectors and the Hessian matrix, and implement this in the `R` programming language. Our approach devises a Newton Multiplicative-Iterative method (Ma, 2010) in order to jointly estimate the regression parameters and baseline hazard, which addresses the steep computational challenge of needing to respect the non-negativity constraint of the baseline hazard.

We compare our proposed model with the partial likelihood method in combination with the Breslow (1972) baseline hazard estimator, using both a simulation study and a real-world application to model time to home loan default. The results of the simulation study show superior performance of the maximum likelihood method over the partial likelihood method to recover the true population parameters in small to moderate sized samples. When applied to a sample of home loans, our results show that both baseline and time-varying estimated regression coefficients agree closely between the approaches, however the maximum likelihood estimate of the baseline hazard has markedly lower volatility. We posit that a potential extension of our model could include life-time probability of default prediction for bank loans, a requirement for expected credit loss calculation as per the IFRS 9 accounting standard.

While the work in this thesis advances the literature in survival modelling, there remains key channels for further research. These encompass methodology, comparison, and computational considerations.

Regarding methodology, there are several aspects we intend to research further. The

first would be catering for competing risks. For example, for time to bank loan default, a competing risk is successful full loan repayment. Currently, such competing risks are treated in our methodology as uninformative right-censoring, which may not be fully reflective of the data generating process. A second extension would be allowing for not only time-varying covariates ($z(t)$), but also time-varying effects ($\beta(t)$ and $\gamma(t)$) associated with both the baseline and time-varying covariates. A third avenue of extending the model is to allow for recurrent events. Currently our method is only able to predict the time to the first default, with second and subsequent defaults removed from the training data. A fourth extension would be to adapt the methodology to other censoring and potentially truncation regimes, such as informative right censoring, interval censoring and left-truncation.

Regarding model comparison, we have left as an avenue of future research the explicit comparison of survival probabilities between the maximum likelihood and partial likelihood methods. This is because we have deliberately focussed on joint estimation of the regression coefficients and the baseline hazard, which we have been able to undertake and demonstrate without (thus far) needing to explicitly compare survival probabilities. In our chosen applied setting of estimating the probability of default, this will be a vital area of research. This will likely involve testing the Cox model against the most commonly used model in credit default modelling, the logistic regression.

Regarding computation, we comment that the required non-negativity constraint on the baseline hazard estimation substantially increases the complexity of the computation for model fitting, and some computational aspects of the model implementation in `R` could be refined to help improve the speed of the algorithm. Aspects of the algorithm that require further research and refinement could include using indices rather than lists of matrices, or leveraging non-base `R` packages that are designed for speed optimisation, such as (for example) the `data.table` and the `dplyr` packages. The `R` model implementation deliberately avoided using these to as to reduce the reliance third-party software and packages which may undergo unannounced changes and in turn impact our model implementation. The ultimate aim is to release an `R` package to the `CRAN`.

In closing, we remind the reader that our research provides several enhancements the prevailing available methodologies to estimate the Cox mode with time -varying covariates. These are: (1) joint estimation of regression coefficients and (smoothed) baseline hazard; (2) calculation of a variance-covariance matrix that will allow point-wise confidence intervals for non-baseline survival probabilities. Our methodology also provides more accurate parameters estimates (in smaller sample sizes) and can allow the baseline hazard to equal zero for some values of t.

A

Appendix

A.1 Appendix 1: Proof of Theorem 1

We provide a brief outline of the proof of Theorem 1.

1. The Newton Multiplicative-Iterative Algorithm developed in section 4.4 involves three steps. The first two are Newton steps, one each for updating $\underline{\beta}$ and $\underline{\gamma}$ where their mappings are denoted by M_1 and M_2 respectively. The third step is the Multiplicative-Iterative step for updating $\underline{\theta}$, denoted by its mapping M_3. These mappings M_1, M_2 and M_3 satisfy:

   - $\underline{\beta}^{(s+1)} = M_1(\underline{\beta}^{(s)}; \underline{\gamma}^{s}, \underline{\theta}^{(s)})$ with $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\right) \geq l\left(\underline{\beta}^{(s)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\right)$
   - $\underline{\gamma}^{(s+1)} = M_2(\underline{\gamma}^{(s)}; \underline{\beta}^{s+1}, \underline{\theta}^{(s)})$ with $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s)}\right) \geq l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\right)$
   - $\underline{\theta}^{(s+1)} = M_3(\underline{\theta}^{(s)}; \underline{\beta}^{s+1}, \underline{\gamma}^{(s+1)})$ with $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s+1)}\right) \geq l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s)}\right)$

2. Both $X^T AX$ and $Z^T BZ$ are non-singular if $A^{1/2}X$ and $B^{1/2}Z$ both have full column rank their gradients given in equation (4.23) and equation (4.27) are both finite. Thus the Newton updates in equations (4.33) and (4.33) are bounded. We also assume that $\underline{\beta}^{(s)} \in \mathcal{B}$ and $\underline{\gamma}^{(s)} \in \mathcal{G}$, where $\mathcal{B} = \{\underline{\beta} : |\beta_c| \leq \tilde{B} < \inf, \forall\, c\}$ and $\mathcal{G} = \{\underline{\gamma} : |\gamma_k| \leq \tilde{G} < \inf, \forall\, k\}$.

3. Update $\underline{\theta}^{s+1}$ from equations (4.37) and (4.38) is bounded if $\underline{\theta}^{(s)}$ is bounded, so we can assume $\underline{\theta}^{(s)} \in \mathcal{T}$ where $\mathcal{T} = \{0 \leq \theta_u \leq T < inf, \forall\, u\}$.

4. Letting $M = M_3 \circ M_2 \circ M_1$, which defines the iteration mapping for the The Newton Multiplicative-Iterative Algorithm, such that $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s+1)}\right) = M\left(\underline{\beta}^{(s)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\right)$.

5. Let $\Lambda_{\underline{\beta}} = \{\underline{\hat{\beta}}\}$ be the $\underline{\beta}$ stationary points set (ie: when $\frac{\partial l}{\partial \underline{\beta}} = \underline{0}$ for $c = 1, ..., p$). Let $\Lambda_{\underline{\gamma}} = \{\underline{\hat{\gamma}}\}$ be the $\underline{\gamma}$ stationary points set (ie: when $\frac{\partial l}{\partial \underline{\gamma}} = \underline{0}$ for $k = 1, ..., q$). Let $\Lambda_{\underline{\theta}} = \{\underline{\hat{\theta}}\}$ be the $\underline{\theta}$ stationary points set (ie: $\frac{\partial l}{\partial \theta_u} = 0$ if $\theta_u \neq 0$ and $\frac{\partial l}{\partial \theta_u} < 0$ if $\theta_u = 0$ for $u = 1, ..., m$. Let $\Lambda = \Lambda_{\underline{\beta}} \times \Lambda_{\underline{\gamma}} \times \Lambda_{\underline{\theta}}$.

6. for Mapping M and the Cartesian product set $\mathcal{B} \times \mathcal{G} \times \mathcal{T}$, these satisfy: (1) Mapping M is defined on the set $\mathcal{B} \times \mathcal{G} \times \mathcal{T}$; (2) The set $\mathcal{B} \times \mathcal{G} \times \mathcal{T}$ is closed and bounded, and thus compact; (3) It is easy to verify that the M_1 is continuous and hence closed for all $\underline{\beta} \in R^p$, M_2 is continuous and hence closed for all $\underline{\gamma} \in R^q$, and M_3 is closed for $\underline{\theta} \notin \Lambda$. Thus M is closed for $(\underline{\beta}, \underline{\gamma}, \underline{\theta}) \notin \Lambda$; (4) $l\left(\underline{\beta}, \underline{\gamma}, \underline{\theta}\right)$ is continuous and satisfies $l\left(\underline{\beta}^{(s+1)}, \underline{\gamma}^{(s+1)}, \underline{\theta}^{(s+1)}\right) \geq l\left(\underline{\beta}^{(s)}, \underline{\gamma}^{(s)}, \underline{\theta}^{(s)}\right)$ on $\mathcal{B} \times \mathcal{G} \times \mathcal{T}$, where equality holds only when the maximum penalised likelihood solution is achieved. The Newton Multiplicative-Iterative Algorithm developed in section 4.4 is convergent if the initial values satisfy: $|\beta_c^{(0)} \leq \tilde{B}|, \forall\, c$; $|\gamma_k^{(0)} \leq \tilde{G}|, \forall\, k$; and $0 < \theta_u^{(0)} \tilde{T}, \forall\, u$.

7. Finally, let $\beta^{(s)} \to \tilde{\beta}$, $\gamma^{(s)} \to \tilde{\gamma}$ and $\theta^{(s)} \to \tilde{\theta}$ as $s \to \infty$. $\tilde{\beta}$ satisfies the condition that $\frac{\partial l(\tilde{\beta}, \tilde{\gamma}, \tilde{\theta})}{\partial \underline{\beta}} = 0$ from equation (4.30). $\tilde{\gamma}$ satisfies the condition that $\frac{\partial l(\tilde{\beta}, \tilde{\gamma}, \tilde{\theta})}{\partial \underline{\gamma}} = 0$ from equation (4.31). For $\tilde{\theta}$, if $\tilde{\theta}_u \neq 0$, it must have $\frac{\partial l(\tilde{\beta}, \tilde{\gamma}, \tilde{\theta})}{\partial \underline{\theta}_u} = 0$ as per equations (4.37) and (4.38). If $\tilde{\theta}_u = 0$, since $\theta_u^{(0)} > 0$, it must have $N_u = 0$ and $\frac{\partial J(\tilde{\beta}, \tilde{\gamma}, \tilde{\theta})}{\partial \theta_u} > 0$, thus $\frac{\partial l(\tilde{\beta}, \tilde{\gamma}, \tilde{\theta})}{\partial \underline{\theta}_u} < 0$, ie: $(\tilde{\beta}, \tilde{\gamma}, \tilde{\theta}) \in \Lambda$. ■

A.2 Appendix 2: Point-Wise Confidence Interval for Survival Probabilities

Our maximum likelihood method provides a full variance-covariance matrix for the $p+q+m$ baseline, time-varying and baseline hazard parameters. This can be used in conjunction with the delta method to estimate the confidence intervals of non-baseline subjects. The Delta method provides a relationship for the sampling distribution of a transform of a set of multivariate normally distributed variables. Following Xu and Scott-Long (2005), if $\underline{\eta}$ is a vector of random variables such that

$$\sqrt{n}\left(\underline{\hat{\eta}} - \underline{\eta}\right) \xrightarrow{D} \left(\underline{0}, Var[\underline{\hat{\eta}}]\right) \tag{A.1}$$

then

$$\sqrt{n}\left(S(t,\underline{\hat{\eta}}) - S(t,\underline{\eta})\right) \xrightarrow{D} \left(\underline{0}, \frac{\partial S(t,\underline{\eta})}{\partial \underline{\eta}^T} Var(\underline{\eta}) \frac{\partial S(t,\underline{\eta})}{\partial \underline{\eta}}\right). \tag{A.2}$$

The final part of this appendix prints the variance-covariance matrices for the Cox model for time to home loan default. The first matrix is estimated using maximum likelihood, and thus contains elements for not only the regression coefficients, but also the baseline hazard. The second matrix is estimated using partial likelihood, and thus contains elements only for the regression coefficients. Note that the order of the regression coefficients in both matrices matches that in table 5.2 in section 5.2.

	Beta1	Beta2	Beta3	Beta4	Beta5	Beta6	Beta7	Beta8	Beta9	Beta10	Beta11	Beta12	Beta13	Gamma1	Gamma2
Beta1	3.1E-03	-3.1E-05	-1.0E-04	1.9E-04	-2.3E-06	7.2E-05	9.0E-05	-4.3E-04	-1.8E-04	2.8E-05	1.4E-06	1.3E-04	4.5E-04	-2.4E-03	-3.3E-05
Beta2	-3.1E-05	4.2E-03	1.1E-03	1.1E-03	1.2E-04	-1.7E-05	-1.1E-04	2.9E-05	-9.7E-06	-5.2E-05	3.4E-07	-3.9E-05	3.8E-05	-2.4E-05	2.2E-04
Beta3	-1.0E-04	1.1E-03	3.5E-03	1.2E-03	-1.5E-04	6.1E-05	1.0E-04	-1.1E-04	-1.4E-04	5.4E-05	3.3E-07	5.9E-05	8.9E-04	-1.8E-04	-4.1E-05
Beta4	1.9E-04	1.1E-03	1.2E-03	9.1E-03	8.0E-05	-4.3E-04	-1.7E-04	-5.0E-04	-2.6E-04	-1.3E-04	2.9E-07	2.8E-05	7.1E-04	-3.1E-04	1.4E-04
Beta5	-2.3E-06	1.2E-04	-1.5E-04	8.0E-05	2.3E-03	-2.6E-04	-7.2E-05	-1.6E-04	-9.2E-05	-1.6E-04	-1.1E-07	-1.1E-04	-4.8E-04	4.8E-06	1.8E-04
Beta6	7.2E-05	-1.7E-05	6.1E-05	-4.3E-04	-2.6E-04	4.5E-02	-3.7E-04	-1.0E-04	-5.5E-04	-1.5E-03	-7.0E-08	2.2E-05	2.2E-04	-1.6E-06	4.4E-04
Beta7	9.0E-05	-1.1E-04	1.0E-04	-1.7E-04	-7.2E-05	-3.7E-04	2.6E-03	-3.9E-04	-9.2E-05	-3.5E-05	-4.2E-06	-1.7E-04	-1.8E-04	-4.9E-05	2.0E-04
Beta8	-4.3E-04	2.9E-05	-1.1E-04	-5.0E-04	-1.6E-04	-1.0E-04	-3.9E-04	1.0E-02	-3.9E-04	-2.5E-05	-3.2E-06	-2.6E-05	4.0E-04	-9.4E-04	-3.3E-04
Beta9	-1.8E-04	-9.7E-06	-1.4E-04	-2.6E-04	-9.2E-05	-5.5E-04	-9.2E-05	-3.9E-04	2.7E-03	5.6E-05	-3.2E-07	2.5E-04	1.2E-03	4.4E-04	-2.8E-05
Beta10	2.8E-05	-5.2E-05	5.4E-05	-1.3E-04	-1.6E-04	-1.5E-03	-3.5E-05	-2.5E-05	5.6E-05	3.6E-03	-9.2E-07	-6.6E-06	1.9E-04	1.6E-04	-3.2E-04
Beta11	1.4E-06	3.4E-07	3.3E-07	2.9E-07	-1.1E-07	-7.0E-08	-4.2E-06	-3.2E-06	-3.2E-07	-9.2E-07	9.7E-08	-7.3E-07	-4.1E-07	2.8E-06	9.4E-07
Beta12	1.3E-04	-3.9E-05	5.9E-05	2.8E-05	-1.1E-04	2.2E-05	-1.7E-04	-2.6E-05	2.5E-04	-6.6E-06	-7.3E-07	3.4E-04	1.2E-04	-1.3E-04	6.8E-05
Beta13	4.5E-04	3.8E-05	8.9E-04	7.1E-04	-4.8E-04	2.2E-04	-1.8E-04	4.0E-04	1.2E-03	1.9E-04	-4.1E-07	1.2E-04	1.4E-02	-8.8E-04	-1.8E-04
Gamma1	-2.4E-03	-2.4E-05	-1.8E-04	-3.1E-04	4.8E-06	-1.6E-06	-4.9E-05	-9.4E-04	4.4E-04	1.6E-04	2.8E-06	-1.3E-04	-8.8E-04	1.7E-02	-1.0E-03
Gamma2	-3.3E-05	2.2E-04	-4.1E-05	1.4E-04	1.8E-04	4.4E-04	2.0E-04	-3.3E-04	-2.8E-05	-3.2E-04	9.4E-07	6.8E-05	-1.8E-04	-1.0E-03	4.0E-03

Figure A.1: Variance-Covariance Matrix from Partial Likelihood Estimation

	Beta1	Beta2	Beta3	Beta4	Beta5	Beta6	Beta7	Beta8	Beta9	Beta10	Beta11	Beta12	Beta13	Gamma1	Gamma2	Theta1	Theta2	Theta3	Theta4	Theta5	Theta6	Theta7	Theta8	Theta9	Theta10	Theta11
Beta1	3.1E-03	-3.2E-05	-1.0E-04	1.9E-04	-1.9E-06	7.4E-05	8.9E-05	-4.3E-04	-1.8E-04	2.5E-05	1.4E-06	1.3E-04	4.5E-04	-2.4E-03	-3.8E-05	-7.7E-08	-1.2E-07	-1.7E-07	-1.9E-07	-2.1E-07	-2.2E-07	-1.9E-07	-1.9E-07	-2.0E-07	-1.4E-07	-1.1E-07
Beta2	-3.2E-05	4.2E-03	1.1E-03	1.1E-03	1.2E-04	-1.4E-05	-1.1E-04	2.8E-05	-1.0E-05	-5.1E-05	3.5E-07	-4.0E-05	3.6E-05	-2.3E-05	2.2E-04	-3.3E-07	-4.6E-07	-5.0E-07	-4.6E-07	-4.4E-07	-4.5E-07	-4.4E-07	-4.7E-07	-6.0E-07	-4.4E-07	-4.4E-07
Beta3	-1.0E-04	1.1E-03	3.5E-03	1.2E-03	-1.5E-04	6.2E-05	1.0E-04	-1.1E-04	-1.5E-04	5.4E-05	3.3E-07	5.8E-05	8.9E-04	-1.7E-04	-4.3E-05	-2.6E-07	-3.7E-07	-3.9E-07	-3.8E-07	-3.6E-07	-3.6E-07	-3.4E-07	-3.7E-07	-4.8E-07	-3.3E-07	-3.1E-07
Beta4	1.9E-04	1.1E-03	1.2E-03	9.1E-03	8.0E-05	-4.3E-04	-1.7E-04	-5.1E-04	-2.6E-04	-1.3E-04	3.0E-07	2.7E-05	7.1E-04	-3.1E-04	1.4E-04	-2.6E-07	-3.6E-07	-3.9E-07	-3.6E-07	-3.6E-07	-3.6E-07	-3.4E-07	-3.9E-07	-4.8E-07	-3.4E-07	-3.3E-07
Beta5	-1.9E-06	1.2E-04	-1.5E-04	8.0E-05	2.3E-03	-2.6E-04	-7.1E-05	-1.6E-04	-9.1E-05	-1.6E-04	-1.2E-07	-1.1E-04	-4.8E-04	2.8E-06	1.8E-04	-3.5E-07	-4.7E-07	-5.3E-07	-4.9E-07	-4.8E-07	-4.8E-07	-4.8E-07	-5.3E-07	-6.6E-07	-4.7E-07	-4.5E-07
Beta6	7.4E-05	-1.4E-05	6.2E-05	-4.3E-04	-2.6E-04	4.5E-02	-3.7E-04	-1.0E-04	-5.5E-04	-1.5E-03	-7.8E-08	2.2E-05	2.3E-04	-1.4E-06	4.5E-04	1.0E-07	1.5E-07	1.5E-07	1.4E-07	1.3E-07	1.1E-07	6.7E-08	-6.0E-08	1.7E-08	-4.4E-08	-1.4E-07
Beta7	8.9E-05	-1.1E-04	1.0E-04	-1.7E-04	-7.1E-05	-3.7E-04	2.6E-03	-3.9E-04	-9.2E-05	-3.6E-05	-4.2E-06	-1.7E-04	-1.8E-04	-5.0E-05	2.0E-04	-3.8E-07	-5.3E-07	-5.9E-07	-5.5E-07	-5.2E-07	-5.3E-07	-5.1E-07	-5.5E-07	-7.0E-07	-5.0E-07	-4.7E-07
Beta8	-4.3E-04	2.8E-05	-1.1E-04	-5.1E-04	-1.6E-04	-1.0E-04	-3.9E-04	1.0E-02	-3.9E-04	-2.0E-05	-3.2E-06	-2.6E-05	3.9E-04	-9.4E-04	-3.3E-04	5.9E-08	9.2E-08	9.7E-08	9.0E-08	1.0E-07	9.2E-08	1.1E-07	1.3E-07	1.5E-07	6.3E-08	9.4E-09
Beta9	-1.8E-04	-1.0E-05	-1.5E-04	-2.6E-04	-9.1E-05	-5.5E-04	-9.2E-05	-3.9E-04	2.7E-03	5.5E-05	-3.2E-07	2.5E-04	1.2E-03	4.4E-04	-2.7E-05	-4.0E-07	-5.3E-07	-5.8E-07	-5.2E-07	-5.0E-07	-5.0E-07	-4.9E-07	-5.4E-07	-7.1E-07	-5.3E-07	-4.9E-07
Beta10	2.5E-05	-5.1E-05	5.4E-05	-1.3E-04	-1.6E-04	-1.5E-03	-3.6E-05	-2.0E-05	5.5E-05	3.6E-03	-9.1E-07	-6.9E-06	1.9E-04	1.6E-04	-3.2E-04	-1.7E-07	-2.6E-07	-2.5E-07	-2.4E-07	-2.3E-07	-2.4E-07	-2.4E-07	-2.6E-07	-3.3E-07	-2.1E-07	-2.1E-07
Beta11	1.4E-06	3.5E-07	3.3E-07	3.0E-07	-1.2E-07	-7.8E-08	-4.2E-06	-3.2E-06	-3.2E-07	-9.1E-07	9.7E-08	-7.3E-07	-4.0E-07	2.8E-06	9.3E-07	6.5E-10	9.2E-10	9.7E-10	8.5E-10	8.3E-10	8.5E-10	7.3E-10	8.2E-10	1.1E-09	7.4E-10	7.4E-10
Beta12	1.3E-04	-4.0E-05	5.8E-05	2.7E-05	-1.1E-04	2.2E-05	-1.7E-04	-2.6E-05	2.5E-04	-6.9E-06	-7.3E-07	3.4E-04	1.2E-04	-1.3E-04	6.8E-05	4.8E-08	6.5E-08	7.1E-08	6.5E-08	6.0E-08	6.1E-08	5.9E-08	6.9E-08	9.2E-08	6.2E-08	5.3E-08
Beta13	4.5E-04	3.6E-05	8.9E-04	7.1E-04	-4.8E-04	2.3E-04	-1.8E-04	3.9E-04	1.2E-03	1.9E-04	-4.0E-07	1.2E-04	1.4E-02	-8.8E-04	-1.8E-04	-2.4E-07	-3.4E-07	-3.5E-07	-3.0E-07	-2.9E-07	-2.5E-07	-2.1E-07	-2.0E-07	-1.9E-07	-1.0E-07	-4.4E-08
Gamma1	-2.4E-03	-2.3E-05	-1.7E-04	-3.1E-04	2.8E-06	-1.4E-06	-5.0E-05	-9.4E-04	4.4E-04	1.6E-04	-4.0E-07	-1.3E-04	-8.8E-04	1.7E-02	-9.9E-04	-8.3E-07	-9.9E-07	-9.4E-07	-7.2E-07	-5.5E-07	-4.6E-07	-3.7E-07	-2.8E-07	-1.7E-07	1.2E-08	1.8E-07
Gamma2	-3.8E-05	2.2E-04	-4.3E-05	1.4E-04	1.8E-04	4.5E-04	2.0E-04	-3.3E-04	-2.7E-05	-3.2E-04	9.3E-07	6.8E-05	-1.8E-04	-9.9E-04	4.0E-03	3.2E-08	-2.1E-07	-3.6E-07	-3.5E-07	-3.8E-07	-3.7E-07	-3.6E-07	-4.0E-07	-4.6E-07	-2.9E-07	-3.1E-07
Theta1	-7.7E-08	-3.3E-07	-2.6E-07	-2.6E-07	-3.5E-07	1.0E-07	-3.8E-07	5.9E-08	-4.0E-07	-1.7E-07	6.5E-10	4.8E-08	-2.4E-07	-8.3E-07	3.2E-08	6.1E-10	3.9E-10	4.3E-10	3.9E-10	3.7E-10	3.6E-10	3.5E-10	3.7E-10	4.6E-10	3.2E-10	2.9E-10
Theta2	-1.2E-07	-4.6E-07	-3.7E-07	-3.6E-07	-4.7E-07	1.5E-07	-5.3E-07	9.2E-08	-5.3E-07	-2.6E-07	9.2E-10	6.5E-08	-3.4E-07	-9.9E-07	-2.1E-07	3.9E-10	9.6E-10	6.0E-10	5.5E-10	5.2E-10	5.2E-10	5.0E-10	5.3E-10	6.6E-10	4.6E-10	4.2E-10
Theta3	-1.7E-07	-5.0E-07	-3.9E-07	-3.9E-07	-5.3E-07	1.5E-07	-5.9E-07	9.7E-08	-5.8E-07	-2.5E-07	9.7E-10	7.1E-08	-3.5E-07	-9.4E-07	-3.6E-07	4.3E-10	6.0E-10	1.2E-09	6.1E-10	5.8E-10	5.8E-10	5.6E-10	6.0E-10	7.4E-10	5.2E-10	4.8E-10
Theta4	-1.9E-07	-4.6E-07	-3.8E-07	-3.6E-07	-4.9E-07	1.4E-07	-5.5E-07	9.0E-08	-5.2E-07	-2.4E-07	8.5E-10	6.5E-08	-3.0E-07	-7.2E-07	-3.5E-07	3.9E-10	5.5E-10	6.1E-10	1.2E-09	5.3E-10	5.3E-10	5.1E-10	5.5E-10	6.8E-10	4.8E-10	4.4E-10
Theta5	-2.1E-07	-4.4E-07	-3.6E-07	-3.6E-07	-4.8E-07	1.3E-07	-5.2E-07	1.0E-07	-5.0E-07	-2.3E-07	8.3E-10	6.0E-08	-2.9E-07	-5.5E-07	-3.8E-07	3.7E-10	5.2E-10	5.8E-10	5.3E-10	1.3E-09	5.1E-10	4.9E-10	5.3E-10	6.6E-10	4.6E-10	4.3E-10
Theta6	-2.2E-07	-4.5E-07	-3.6E-07	-3.6E-07	-4.8E-07	1.1E-07	-5.3E-07	9.2E-08	-5.0E-07	-2.4E-07	8.5E-10	6.1E-08	-2.5E-07	-4.6E-07	-3.7E-07	3.6E-10	5.2E-10	5.8E-10	5.3E-10	5.1E-10	1.6E-09	5.0E-10	5.4E-10	6.7E-10	4.7E-10	4.2E-10
Theta7	-1.9E-07	-4.4E-07	-3.4E-07	-3.4E-07	-4.8E-07	6.7E-08	-5.1E-07	1.1E-07	-4.9E-07	-2.4E-07	7.3E-10	5.9E-08	-2.4E-07	-3.7E-07	-4.5E-07	3.5E-10	5.0E-10	5.6E-10	5.1E-10	4.9E-10	5.0E-10	2.1E-09	5.2E-10	6.5E-10	7.1E-10	6.5E-10
Theta8	-1.9E-07	-4.7E-07	-3.7E-07	-3.9E-07	-5.3E-07	-6.0E-08	-5.5E-07	1.3E-07	-5.4E-07	-2.6E-07	8.2E-10	6.9E-08	-2.0E-07	-4.0E-07	-4.0E-07	3.7E-10	5.3E-10	6.0E-10	5.5E-10	5.3E-10	5.4E-10	5.2E-10	3.1E-09	7.1E-10	5.0E-10	4.7E-10
Theta9	-2.0E-07	-6.0E-07	-4.8E-07	-4.8E-07	-6.6E-07	1.7E-08	-7.0E-07	1.5E-07	-7.1E-07	-3.3E-07	1.1E-09	9.2E-08	-1.9E-07	-1.7E-07	-4.6E-07	4.6E-10	6.6E-10	7.4E-10	6.8E-10	6.6E-10	6.7E-10	6.5E-10	7.1E-10	6.0E-09	6.3E-10	5.9E-10
Theta10	-1.4E-07	-4.4E-07	-3.3E-07	-3.4E-07	-4.7E-07	-4.4E-08	-5.0E-07	6.3E-08	-5.3E-07	-2.1E-07	7.4E-10	6.2E-08	-1.0E-07	1.2E-08	-2.9E-07	3.2E-10	4.6E-10	5.2E-10	4.8E-10	4.6E-10	4.7E-10	4.6E-10	5.0E-10	6.3E-10	6.8E-09	4.2E-10
Theta11	-1.1E-07	-4.1E-07	-3.1E-07	-3.3E-07	-4.5E-07	-1.4E-07	-4.7E-07	9.4E-09	-4.9E-07	-2.1E-07	7.4E-10	5.3E-08	-4.4E-08	1.8E-07	-3.1E-07	2.9E-10	4.2E-10	4.8E-10	4.4E-10	4.3E-10	4.3E-10	4.2E-10	4.7E-10	5.9E-10	4.2E-10	6.4E-09

Figure A.2: Variance-Covariance Matrix from Maximum Likelihood Estimation

B

Supplementary Material

B.1 Supplementary 1: Model Implementation R Code

Here we present our the R code implementation of our model. The code leverages the existing R package `survival_mpl` (by Couturier et al. (2014)), whose documentation states that is "inspired" but the functionality of the base package `survival` (by Therneau (2015)). Both packages are available from the CRAN. The R code to implement the `survival_mpl` was translated and kindly shared by Dr Maurizio Manuguerra.

```
#============================================
cox_mle_tvc=function(formula,data,formula.z,riji,subject,data.z,
subset,na.action,control,...){
#  get and organise information
# (same tests as in coxph(), thanks to the survival package)
mc = match.call(expand.dots = FALSE)
m  = match(c("formula","data","subset","na.action"),names(mc),0)
mc.orig = mc
mc = mc[c(1,m)]

if (m[1]==0){stop("A formula argument is required")}
data.name = if(m[2]!=0){mc[m[2]][[1]]}else{"-"}
mc[[1]] = as.name("model.frame")
mc$formula = if(missing(data)) terms(formula)
else              terms(formula, data=data)
mf = eval(mc,parent.frame())
if (any(is.na(mf))) stop("Missing observations in the model variables")
if (nrow(mf) ==0) stop("No (non-missing) observations")
mt = attr(mf,"terms")
# extract response
```

```
y    = model.extract(mf, "response")
type = attr(y, "type")

if(!inherits(y, "Surv")){stop("Response must be a survival object")}
if(type!="right"&&type!="counting"){
stop(paste("Cox model doesn't support \"", type, "\"
survival data",sep = ""))
}
t_i         = y[,1L]
observed    = y[,2L]==1L
n           = length(t_i)
n.obs       = sum(observed)

# control arguments
extraArgs <- list(...)
if (length(extraArgs)) {
controlargs <- names(formals(cpox_mle_tvc.control))
m <- pmatch(names(extraArgs), controlargs, nomatch=0L)
if (any(m==0L))
stop(gettextf("Argument(s) %s not matched", names(extraArgs)[m==0L]),
domain = NA, call. = FALSE)
}
if (missing(control)) control <- cpox_mle_tvc.control(n.obs, ...)
# ties
t_i.obs  = t_i[observed]
ties     = duplicated(t_i.obs)
if(any(ties)){
if(control$ties=="epsilon"){
if(length(control$seed)>0){
old <- .Random.seed
on.exit({.Random.seed <<- old})
set.seed(control$seed)
}
t_i.obs[ties] = t_i.obs[ties]+runif(sum(ties),-1e-11,1e-11)
t_i[observed] = t_i.obs
}else{
t_i.obs = t_i.obs[!ties]
n.obs   = length(t_i.obs)
}
}
# X and centered X matrix
X           = model.matrix(mt, mf, contrasts)
X           = X[,!apply(X, 2, function(x) all(x==x[1])), drop=FALSE]
if(ncol(X)==0){
X   = matrix(0,n,1)
noX = TRUE
}else{  noX = FALSE}
p           = ncol(X)
mean_j      = apply(X, 2, mean)
#XC          = X - rep(mean_j, each=n)
XC=X
mean_j=rep(0,p)

# knot sequence and psi matrices
knots  = knots_mpl(control, t_i.obs, range(t_i))
m      = knots$m
psi    = basis_mpl(t_i,knots,control$basis,control$order,which=1)
PSI    = basis_mpl(t_i,knots,control$basis,control$order,which=2)
R      = penalty_mpl(control,knots)
```

```
# Z variables
var.z=all.names(formula.z)
var.z=var.z[-which(var.z %in% c('~','+','.'))]

# Narrow the Z matrix
var.riji=mc.orig[[which(names(mc.orig)=='riji')]]
var.subj=mc.orig[[which(names(mc.orig)=='subject')]]
var.riji=as.character(var.riji)
var.subj=as.character(var.subj)
Z=data.z[,c(var.subj, var.riji, var.z)]
noZ=FALSE
q=ncol(Z)

# Z matricies, with appropriate individual psi and PSI matricies
# make a list storing z matrix for each subject
Zsplit=lapply(split(Z[,var.z],Z[,1]),FUN=as.matrix)

# Create matrix of last values of z for each subject
last.z=lapply(Zsplit,
FUN=tail,
n=1)
last.z=do.call(rbind, last.z)
rownames(last.z)=NULL

# make a list storing times for z matrix for each subject
# a list of n maxricies of dimension j_{i} x q
Zsplit.time=lapply(split(Z[,var.riji],Z[,1]),FUN=as.matrix)

# psi for each z matrix
# a list of n maxricies of dimension j_{i} x m
Npsi=lapply(Zsplit.time,
FUN=basis_mpl,
knots=knots,
basis=control$basis,
order=control$order,
which=1)

# PSI for each z matrix, set minimum to zero
# a list of n maxricies of dimension j_{i} x m
NPSI=lapply(Zsplit.time,
FUN=basis_mpl,
knots=knots,
basis=control$basis,
order=control$order,
which=2)
NPSI=lapply(NPSI, FUN=function(x) {ifelse(x<0,0,x)})

# Difference in zPSI - First differences, but retaining the first element
# Result is a list of n vectors j_{i} x m
NPSIdiff=lapply(NPSI,
FUN=function(val) {rbind(val[1,] ,val[-1L,] - val[1:dim(val)[1]-1,])})

# Initial value for Gamma
q=length(var.z)
Gamma=as.matrix(rep(0,q))

#Most narrow version of Z
Z=as.matrix(data.z[,c(var.z)])
```

```
# Create status
status=matrix(c(as.integer(observed),
seq(1,length(observed))),
ncol=2)
status.list=lapply(split(status[,1],status[,2]),FUN=as.matrix)

# Create long status
longstatus.list=mapply(FUN=function(val1,val2) {
matrix(c(rep(0,length(val1)-1),val2), ncol=1)},
val1=Zsplit.time,
val2=status.list,
SIMPLIFY = FALSE)
longstatus=do.call(rbind,longstatus.list)
#==============
lambda     = control$smooth
Beta       = rep(0,p)
Theta      = rep(1,knots$m)
correction = 1
full.iter  = 0
this.max.iter=control$max.iter[1]
this.max.iter=1
for(iter in 1:this.max.iter){
fit <- coxphfit(
status = as.integer(observed), longstatus=longstatus,
 X=XC, meanX = mean_j, R = R,
psi  = psi, PSI  = PSI,
Z=Z, Zsplit=Zsplit, Zsplit.time=Zsplit.time, last.z=last.z,
Npsi=Npsi, NPSI=NPSI, NPSIdiff=NPSIdiff,
Beta0 = Beta, Theta0 = Theta/correction, Gamma0=Gamma,
lambda = as.double(lambda), kappa = control$kappa,
convVal = control$tol, minTheta = control$epsilon,
maxiter = control$max.iter[2])
}
#================
if(control$max.iter[1]>1) control$smooth = lambda
M_theta_m1 = fit$coef$Theta
H      = as.matrix(fit$matricies$H) ;rownames(H)=colnames(H)=NULL
p      = length(fit$coef$Beta)
m      = length(fit$coef$Theta)
q      = length(fit$coef$Gamma)
Minv_2 = Hinv = matrix(0,p+q+m,p+q+m)

pos = c(rep(TRUE,p), rep(TRUE,q), !(abs(M_theta_m1)<1E-5) & !(fit$GradTheta< -1E-2))

# Hessian
temp = try(solve(H[pos,pos]),silent=TRUE)
if(class(temp)!="try-error"){
Hinv[pos,pos] = temp
cov_NuNu_H    = Hinv
se.Eta_H      = suppressWarnings(sqrt(diag(cov_NuNu_H)))
}else{
cov_NuNu_H  = matrix(NA,p+m,p+m)
se.Eta_H    = rep(NA,p+m)
}

# Graph data
se.Theta=se.Eta_H[(p+q+1):(p+q+m)]
se.Theta2=se.Theta[c(1,1:m)]
```

```
Theta2=fit$coef$Theta[c(1,1:m)]
graphData=data.frame(Alpha=knots$Alpha,
Theta=Theta2,
se.Theta=se.Theta2,
low=Theta2-1.96*se.Theta2,
high=Theta2+1.96*se.Theta2)

# output
fit$knots   = knots
fit$control = control
fit$call    = match.call()
fit$dim     = list(n = n, n.obs = sum(observed), n.ties = sum(ties),
p = p, q = q, m = knots$m)
fit$data    = list(time = t_i, observed = observed, X = X, Z = Z,
name = data.name, graphData=graphData)
fit$matricies=list(H=H, cov_NuNu_H=cov_NuNu_H, se.Eta_H=se.Eta_H)
class(fit)  = "cox_mle_tvc"
fit
}
#============================================

#============================================
# Function to update H0star using Gamma and Theta
calc<-function(NPSIdiff, Zsplit, thisGamma, thisTheta)
{
# Multiply 2 lists of n matricies j_{i} x q and q x 1
# Result n vectors j_{i} x 1
eZGamma=lapply(Zsplit, FUN= function(val) {exp(val%*%thisGamma)})

# replicate the number of columns for exp(zT.Gamma)
thism=dim(thisTheta)[1]
eZGamma.repm=lapply(eZGamma, FUN=function(val){
matrix(rep(val, time=thism), ncol=thism)})

# Element wise multiplication
NPSIdiff_by_eZGamma=mapply(FUN=function(val1,val2){val1 * val2},
val1=NPSIdiff,
val2=eZGamma.repm,
SIMPLIFY = FALSE)

# Create PSIstar - nXm
PSIstar.list=lapply(NPSIdiff_by_eZGamma, FUN=colSums)
PSIstar=do.call(rbind,PSIstar.list)

# Create H0star - nX1
H0star=PSIstar %*% thisTheta

return(list(H0star=H0star,PSIstar=PSIstar,
NPSIdiff_by_eZGamma=NPSIdiff_by_eZGamma))
}
#============================================

#============================================
coxphfit <- function(status, longstatus, X, meanX, R, psi,
PSI, Z, Zsplit, Zsplit.time, last.z, Npsi, NPSI, NPSIdiff,
Beta0, Gamma0, Theta0, lambda, kappa, convVal, minTheta,
maxiter){
p = ncol(X)
n = nrow(X)
```

```
m = ncol(R)
N = dim(Z)
l = p+m
# Initialise
Theta0 = as.matrix(Theta0) # mx1
PsiTheta = PSI %*% Theta0  # nxm x mx1 = nx1
psiTheta = psi %*% Theta0  # nxm x mx1 = nx1
Mu = exp(X %*% Beta0)      # nxp x px1 = nx1
PsiThetaMu <- PsiTheta*Mu  # element-wise nx1
psiThetaMu <- psiTheta*Mu  # element-wise nx1
RTheta = R %*% Theta0      # mxm x mx1 = mx1
#==============
# Initialise
q=ncol(Z)
l = p+m+q
Zu=exp(last.z%*% Gamma0) # nxq x qx1 = nx1

# Initialise H0star using Gamma0 and Theta0
calc0=calc(NPSIdiff=NPSIdiff, Zsplit=Zsplit,
thisGamma=Gamma0, thisTheta=Theta0)
H0star=calc0$H0star

# ith component is the ith obs contribution to loglik
loglik = -H0star*Mu + status*log(psiTheta*Mu*Zu)

# 1xn x nx1 = 1x1 scalar
pen <- as.numeric(crossprod(Theta0,RTheta))

# 1x1 scalar, the penalised likelihood
ploglik0 = (1-lambda)*sum(loglik) - lambda*pen

Gamma=Gamma0
Beta=Beta0
Theta=Theta0
#==============
# Save values
ploglikMat=matrix(rep(0,(maxiter+1)*3), ncol=3, nrow=maxiter+1)
ploglikMat[1,]=ploglik0

BetaMat =matrix(rep(0,(maxiter+1)*p), nrow=maxiter+1, ncol=p)
GammaMat=matrix(rep(0,(maxiter+1)*q), nrow=maxiter+1, ncol=q)
ThetaMat=matrix(rep(0,(maxiter+1)*m), nrow=maxiter+1, ncol=m)
BetaMat [1,]=Beta0
GammaMat[1,]=Gamma0
ThetaMat[1,]=Theta0
#==============
# Update Beta
for(iter in 1:maxiter){
# Update beta
StatusMinH0starMu = status-H0star*Mu
GradBeta <- t(X) %*% StatusMinH0starMu
#HessianBeta <- t(X) %*% diag(as.numeric(H0star*Mu)) %*% X
HessianBeta <- as.matrix(t(X) %*% Diagonal(n=n, x=H0star*Mu) %*% X)
StepBeta <- solve(HessianBeta) %*% GradBeta
Beta <- Beta0 + StepBeta
##
Mu = exp(X %*% Beta)
loglik = -H0star*Mu + status*log(psiTheta*Mu*Zu)
ploglik = (1-lambda)*sum(loglik) - lambda*pen
```

```
# Adapt Newton step if needed
r=0
while(ploglik < ploglik0){
r=r+1
StepBeta = StepBeta/kappa
Beta <- Beta0 + StepBeta
Mu = exp(X %*% Beta)
loglik = -H0star*Mu + status*log(psiTheta*Mu*Zu)
ploglik = (1-lambda)*sum(loglik) - lambda*pen
if (r>500) break
}
ploglik0=ploglik
ploglikMat[1+iter,1]=ploglik0
#==============
# Update gamma
# List of n matricies j_{i} x m
NPSIdiff_by_eZGamma=calc0$NPSIdiff_by_eZGamma

# List of n matricies j_{i} x 1
NPSIdiff_by_eZGamma_by_Theta=lapply(NPSIdiff_by_eZGamma,
FUN=function(val1){val1 %*% Theta})

# List of n vectors j_{i} x 1
NPSIdiff_by_eZGamma_by_Theta_by_Mu=mapply(FUN=function(val1,val2) {
val1*val2},
val1=NPSIdiff_by_eZGamma_by_Theta,
val2=Mu,
SIMPLIFY = FALSE)
# Elements matrix B
# Vector  N x 1 (N= sum of j_{i})
Belement=do.call(rbind, NPSIdiff_by_eZGamma_by_Theta_by_Mu)

# Newton
longStatusMinBelementexpandMu = longstatus-Belement
GradGamma <- t(Z) %*% longStatusMinBelementexpandMu
HessianGamma <- t(Z) %*% Diagonal(n=length(Belement), x=Belement)
%*% as.matrix(Z)
StepGamma <- solve(HessianGamma) %*% GradGamma
Gamma <- Gamma0 + StepGamma
Gamma=as.matrix(Gamma)

# Update H0star using Gamma and Theta0
Zu=exp(last.z%*% Gamma)
calc0=calc(NPSIdiff=NPSIdiff, Zsplit=Zsplit, thisGamma=Gamma,
thisTheta=Theta)
H0star=calc0$H0star

loglik = -H0star*Mu + status*log(psiTheta*Mu*Zu)
# nx1 + element-wise nx1 = nx1
# (ith component is the ith obs contribution to loglik)
# 1x1 scalar, the penalised likelihood
ploglik = (1-lambda)*sum(loglik) - lambda*pen

# Adapt Newton step if needed
r=0
while(ploglik < ploglik0){
r=r+1
StepGamma = StepGamma/kappa
Gamma <- Gamma0 + StepGamma
```

```
Gamma=as.matrix(Gamma)
Zu=exp(last.z%*% Gamma)
calc0=calc(NPSIdiff=NPSIdiff, Zsplit=Zsplit, thisGamma=Gamma, thisTheta=Theta)
H0star=calc0$H0star
loglik = -H0star*Mu + status*log((psi%*%Theta)*Mu*Zu)
# nx1 + element-wise nx1 = nx1
#(ith component is the ith obs contribution to loglik)
# 1x1 scalar, the penalised likelihood
ploglik = (1-lambda)*sum(loglik) - lambda*pen
if (r>50000) break
}
ploglik0=ploglik
ploglikMat[1+iter,2]=ploglik0
#==============
# Update theta
# nXm /(element-wise) nXm matrix
W <- psi/matrix(psi %*% Theta,nrow=nrow(psi),ncol=ncol(psi), byrow=F)
WTstatus <- t(W) %*% status #mXn x nx1, result is mx1

PSIstar=calc0$PSIstar
PSIstarMu <- t(PSIstar) %*% Mu
GradTheta <- (1-lambda)*(WTstatus-PSIstarMu) - 2*lambda*(R %*% Theta)

sTheta <- Theta/((1-lambda)*PSIstarMu +
ifelse(RTheta>0, 2*lambda*(R %*% Theta0), 0) + 0.3)
StepTheta <- GradTheta*sTheta
Theta <- as.matrix(Theta0) + StepTheta
Theta[which(Theta<minTheta)]=minTheta

calc0=calc(NPSIdiff=NPSIdiff, Zsplit=Zsplit, thisGamma=Gamma, thisTheta=Theta)
H0star=calc0$H0star # nX1
psiTheta = psi %*% Theta # nXm x mX1, result is nX1
RTheta = R %*% Theta # mXm x mX1, result is mX1
loglik = -H0star*Mu + status*log(psiTheta*Mu*Zu)
pen <- as.numeric(crossprod(Theta,RTheta))
ploglik = (1-lambda)*sum(loglik) - lambda*pen

# Adapt Newton step if needed
r=0
while(ploglik < ploglik0){
r=r+1
StepTheta <- StepTheta/kappa
Theta <- as.matrix(Theta0) + StepTheta
Theta[which(Theta<minTheta)]=minTheta
calc0=calc(NPSIdiff=NPSIdiff, Zsplit=Zsplit, thisGamma=Gamma,
thisTheta=Theta)
H0star=calc0$H0star
psiTheta = psi %*% Theta
psiThetaMu <- psiTheta*Mu
RTheta = R %*% Theta
loglik = -H0star*Mu + status*log(psiThetaMu)
pen <- as.numeric(crossprod(Theta,RTheta))
ploglik = (1-lambda)*sum(loglik) - lambda*pen
if (r>500) break
}

# Save the penalised likelihood
ploglik0 <- ploglik
ploglikMat[1+iter,3]=ploglik0
```

```
# Check for convergence
varepsilon <- max(c(abs(Beta-Beta0),abs(Gamma-Gamma0),abs(Theta-Theta0)))
if (varepsilon<convVal) break
Beta0 <- Beta
Gamma0 <- Gamma
Theta0 <- Theta

BetaMat [1+iter,]=Beta0
GammaMat[1+iter,]=Gamma0
ThetaMat[1+iter,]=Theta0
#print(iter)
if ( (round(iter/10,6)-floor(iter/10)) == 0) print(iter)
}
#Correction for Theta
correction <- exp(sum(-meanX*Beta))
ploglik <- c(ploglik, correction)
Theta <- Theta*correction
# ================
#Inference for Beta
H0starMu=H0star*Mu
V1 <- - t(X) %*% Diagonal(n=length(H0starMu), x=H0starMu) %*% X
colnames(V1)=rownames(V1)=NULL
HessianBeta <- V1
# ================
# Inference for Gamma
# List of n matricies j_{i} x m
NPSIdiff_by_eZGamma=calc0$NPSIdiff_by_eZGamma

# List of n matricies j_{i} x 1
NPSIdiff_by_eZGamma_by_Theta=lapply(NPSIdiff_by_eZGamma,
FUN=function(val1){val1 %*% Theta})

# List of n vectors j_{i} x 1
NPSIdiff_by_eZGamma_by_Theta_by_Mu=mapply(FUN=function(val1,val2) {
val1*val2},
val1=NPSIdiff_by_eZGamma_by_Theta,
val2=Mu,
SIMPLIFY = FALSE)
# Elements matrix B
# Vector  N x 1 (N= sum of j_{i})
Belement=do.call(rbind, NPSIdiff_by_eZGamma_by_Theta_by_Mu)

V2= - t(Z) %*% Diagonal(n=length(Belement), x=Belement) %*% as.matrix(Z)
V2=as.matrix(V2)
HessianGamma <- V2
# ================
# Inference for Theta
#(mxn)x(nxn)x(nxm)=mxm #B
V3 <- as.matrix(t(psi) %*% Diagonal(n=n, x=as.numeric(status/psiTheta^2))
%*% psi)
HessianTheta <- V3
# ================
# Inference for d2l/dbeta dtheta
V13 = - t(X)  %*% Diagonal(n=length(Mu), x=Mu) %*% calc0$PSIstar
V13=as.matrix(V13)
colnames(V13)=rownames(V13)=NULL
# ================
# Inference for dl/dbeta dgamma
```

```
# Replicate the rows of X j_{i} times each - result is Nxp matrix
timesToRep=do.call(rbind,lapply(Zsplit,dim))
rows <- rep( 1:nrow(X) , timesToRep[ , 1 ] )
Xrep=matrix(X[rows,], ncol=ncol(X))
V12 = - t(Xrep) %*% Diagonal(n=length(Belement), x=Belement) %*% Z
V12=as.matrix(V12)
# ================
# Inference for dl/dgamma dtheta
# Replicate the rows of Mu j_{i} times each - result is Nx1 vector
timesToRep=do.call(rbind,lapply(Zsplit,dim))
rows <- rep( 1:nrow(Mu) , timesToRep[ , 1 ] )
Murep=matrix(Mu[rows,], ncol=1)
V23= - t(Z) %*% Diagonal(n=length(Murep), x=Murep) %*%
do.call(rbind,calc0$NPSIdiff_by_eZGamma)
V23=as.matrix(V23)
# ================
# H matrix
row1=cbind(   V1 ,   V12 , V13)
row2=cbind(t(V12),    V2 , V23)
row3=cbind(t(V13), t(V23), -V3)
Halt=-1*rbind(row1, row2, row3)
H=Halt
M2=Halt
# ================
#Output
return(list(coef=list(Beta=Beta, Gamma=Gamma, Theta=Theta),
loglik=list(iter=iter, ploglik=ploglik[1], correction=ploglik[2],
ploglikMat=ploglikMat),matricies=list(H=H),GradTheta=GradTheta,
History=list(BetaMat=BetaMat, GammaMat=GammaMat, ThetaMat=ThetaMat)))
}
#============================================

#============================================
cpox_mle_tvc.control <- function(...)
... unchanged from survivalMPL::coxph_mpl.control
#============================================
basis.name_mpl <- function(...)
... unchanged from survivalMPL::basis.name_mpl
#============================================
penalty.order_mpl <- function(...)
... unchanged from survivalMPL::penalty.order_mpl
#============================================
knots_mpl <- function(...)
... unchanged from survivalMPL::knots_mpl
#============================================
basis_mpl <- function(...)
... unchanged from survivalMPL::basis_mpl
#============================================
penalty_mpl <- function(...)
... unchanged from survivalMPL::penalty_mpl
#============================================
```

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6:701.

APRA (2016). Statistics quarterly authorised deposit-taking institution performance (http://apra.gov.au/adi/publications/documents/3105-qadips-mar-2016.pdf). [Online; accessed 17/08/2016].

Araujo, A., Meira-Machado, L., and Faria, S. (2015). *genSurv: Generating multi-state survival data. R package version 1.0.3*. R package version 1.0.3.

Australia and New Zealand Banking Group (2016). 2016 basel iii pillar 3 disclosure as at 31 march 2016 aps:330 public disclosure (www.anz.com). [Online; accessed 24/07/2016].

Australian Bureau of Statistics (2016). 1345.0 - key economic indicators, 2016 (abs.gov.au). [Online; accessed 31/07/2016].

Banasik, J., Crook, J., and Thomas, L. (1999). Not if but when will borrowers default. *journal of the Operational Research Society*, 50:1185–1190.

Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60:1699–1707.

BIS (2006). International convergence of capital measurement and capital standards a revised framework comprehensive version. *Bank for International Settlements*.

Breslow, N. E. (1972). Contribution to the discussion of paper by d.r. cox. j r. *Journal of the Royal Statistical Society: Series B*, 34:216–217.

Cai, T. and Betensky, R. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics*, 59:570–579.

Cai, T., Hyndman, R., and Wand, M. (2010). Mixed model-based hazard estimation. *Working Paper*.

Commonwealth Bank of Australia (2016). Basel iii pillar 3 capital adequacy and risk disclosures as at 31 match 2016 (www.commbank.com.au). [Online; accessed 24/07/2016].

Couturier, D.-L., Ma, J., and Heritier., S. (2014). *survivalMPL: Penalised Maximum Likelihood for Survival Analysis Models*. R package version 0.1.1.

Cox, D. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Cox, D. (1975). Partial likelihood. *Biometrika*, 62:269–276.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, New York, NY.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, 72:27–36.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1):1–38.

Fisher, L. and Lin, D. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual Review of Public Health*, 20:145–57.

Good, I. J. (1950). *Probability and the Weighting of Evidence*. Charles Griffith, London.

Hosmer, D., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data, second ed*. Wiley-Interscience, Hoboken, New Jersey.

IASB (2014). International financial reporting standard 9.

Im, J., Apley, D., Qi, C., and Shan, X. (2012). A time-dependent proportional hazards survival model for credit risk analysis. *Journal of the Operational Research Society*, 63:306 –321.

Johansen, S. (1983). An extension of cox's regression model. *International Statistical Review*, 51:165–174.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley and Sons, New York, NY.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *The Journal of the American Statistical Association*, 53 (282):457–481.

Karush, W. (1939). Minima of functions of several variables with inequalities as side constraints. Master's thesis, M.Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis Techniques for Censored and Truncated Data, Second Edition*. Springer, New York, NY.

Kneib, T. and Fahrmeir, L. (2004). A mixed model approach for structured hazard regression. *Sonderforschungsbereich, Paper 400*.

Kuhn, H. and Tucker (1951). Nonlinear programming. *Proceedings of 2nd Berkeley Symposium. Berkeley*, pages 481–492.

Lessmann, S., Baesens, B., Seowd, H., and Thomas, L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *The Journal of the Operational Research Society*, 247:124–136.

Luenberger, D. and Ye, Y. (2008). *Linear and Nonlinear Programming, 3 ed.* Springer, New York, NY.

Ma, J. (2010). Positively constrained multiplicative iterative algorithm for maximum penalized likelihood tomographic reconstruction. *IEEE Transactions on Nuclear Science*, 57:181–192.

Ma, J., Heritier, S., and Lo, S. (2014). On the maximum penalized likelihood approach for proportional hazard models with right censored survival data. *Computational Statistics and Data Analysis*, 74:142 – 156.

Man, R. (2014). Survival analysis in credit scoring a framework for pd estimation. Master's thesis, University of Twente, The Netherlands.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163.

Narian, B. (1992). *Survival analysis and the credit granting decision, in Credit Scoring and Credit Control, L. C. Thomas, J. N. Crook, D. B. Edelman, eds.* Oxford University Press.

National Australia Bank (2016). 2016 pillar 3 report incorporating the requirements of aps330 half year update as at 31 march 2016 (www.nab.com.au). [Online; accessed 24/07/2016].

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1:27.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135(2):185.

R Core Team (2016). R 3.2.3: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. url http://www.r-project.org.

Ren, J. and Zhou, M. (2011). Full likelihood inferences in the cox model: an empirical likelihood approach. *Annals of the Institute of Statistical Mathematics*, 63 (Issue 5):1005–1018.

Rodriguez, G. (2005). Non-parametric estimation in survival models. *Princeton Lecture Notes*.

Royston, P. (2011). Estimating a smooth baseline hazard function for the cox model. *Working Paper*.

SAS Institute Inc (2016). Sas software, version 9.4. cary. url http://www.sas.com.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tets for comparing survival distributions. *Biometrikal*, 68:316–319.

Stepanova, M. and Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, 50(2):277–289.

Sterne, J., White, I., Carlin, J., Spratt, M., Royston, P., Kenward, M., Wood, A., and Carpenter, J. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal (BMJ 2009;338:b2393)*.

Therneau, T., Crowson, C., and Atkinson, E. (2015). Using time dependent covariates and time dependent coefficients in the cox model (https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf). *R CRAN Vignette*.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. R package version 2.38.

Tong, E. N., Mues, C., and Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, 218(1):132 – 139.

van Houwelingen, H. (2000). Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, 19:3401–3415.

Westpac Banking Corporation (2016). Pillar 3 report march 2016 incorporating the requirements of aps330 (www.westpac.com.au). [Online; accessed 24/07/2016].

Xu, J. and Scott-Long, J. (2005). Confidence intervals for predicted outcomes in regression models for categorical outcomes. *The Stata Journal*, 5 (4):537–559.

Zheng, D. and Lin, D. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society B*, 69(Part 4):507–564.