# Discrete-Time Threshold Regression for Survival Data with Time-Dependent Covariates

Stefan Sharkansky

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Kwun Chuen Gary Chan, Chair

Volodymyr Minin

Lurdes Inoue

Program Authorized to Offer Degree:
Statistics

University of Washington

**Abstract**

Discrete-Time Threshold Regression for Survival Data with Time-Dependent Covariates

Stefan Sharkansky

Chair of the Supervisory Committee:
Associate Professor Kwun Chuen Gary Chan
Biostatistics

A natural approach to survival analysis in many settings is to model the subject's "health" status as a latent stochastic process, where the terminal event is represented by the first time that the process crosses a threshold. "Threshold regression" models the covariate effects on the latent process. Much of the literature on threshold regression assumes that the process is a one-dimensional Wiener process, where crossing times have a tractable inverse Gaussian distribution but where the process characteristics are fixed at baseline. This framework is not easily extended to incorporate time-varying covariates or dependent competing risks. We introduce a novel approach for performing threshold regression with time-dependent covariates in a discrete time setting, where the process is a Gaussian random walk, with time-varying drift as a parameterized function of time-varying covariates. This model is then extended to consider dual correlated competing risks. We present methods for estimating model parameters, including an EM algorithm, and outline numerical algorithms for efficiently evaluating the observed and complete data likelihoods and score functions and for estimating standard errors. We discuss results of applying these methods to both simulated data and to the Freddie Mac residential mortgage data set. In the latter case we quantify associations between baseline borrower characteristics and time-varying macroeconomic conditions versus time to mortgage default and prepayment events.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

I am tremendously grateful to my advisor, Gary Chan, as fine a coach and collaborator as I could have hoped for; and to my committee members Lurdes Inoue, Vladimir Minin and Hendrik Wolff for their invaluable insights and comments which improved both the content and presentation of this dissertation; I owe thanks to the many other professors in Statistics, Biostatistics and Economics who have taught and inspired me throughout my course of study; and to Michael Perlman, Thomas Richardson and Ellen Reynolds for their indispensable advice and encouragement.

I thank my family – my parents, Ina and Ira, for instilling in me from an early age the joy and value of lifelong learning; and my wife Irene and our son David for all their love, support, patience and good times.

# DEDICATION

to Irene and David

# Chapter 1

# INTRODUCTION AND LITERATURE REVIEW

> "People, he had said, were always being looked at as points, and they ought to be looked at as lines. There weren't any points, it was false to assume that a person ever was anything. He was always becoming something, always changing, always continuous and moving, like the wiggly line on a machine used to measure earthquake shocks. He was always what he was in the beginning, but never exactly what he was; he moved along a line dictated by his heritage and his environment, but he was subject to every sort of variation within the narrow limits of his capabilities", Wallace Stegner, *The Big Rock Candy Mountain*, 1943.

## *1.1 Background and Motivation*

This dissertation is motivated by the study of credit risk, and specifically the risks of mortgage default and prepayment. Survival analysis methods in general are a good fit for studying credit risk, since a key quantity of interest is the duration from initial observation to termination. Indeed, much literature, as noted in Sec. 1.2, has studied credit risk using familiar survival methods such as the Cox (1972) proportional hazards (PH) model. In this dissertation we explore alternative survival models that follow naturally from commonly used economic models of credit risk, the variants on the structural models of Merton (1974). While credit risk is the original impetus for this research, the methods we develop grow out of established methods that have been applied in a variety of fields. We believe that our methodological contributions will find applications in other domains where survival methods are used.

In essence, the structural credit model hypothesizes that a firm will default on its debt

when the value of the debt exceeds the liquidation value of the firm; where the latter is assumed to evolve stochastically. Analogously, a household may be assumed to default on a mortgage or other debt when the debt exceeds the household net worth and earning prospects. In studies on credit risk of publicly held companies, such as Duffie (2011), the firms' assets and liabilities are disclosed periodically and the market values are broadcast and recorded continuously. In the case of household credit, however, the household financials are generally unobservable, if even ascertainable with any accuracy. Thus the household net worth could be modeled as a latent process where an observable default implies that the net worth has reached zero. However, a household does not have a liquidation value in the same sense that a firm does – e.g. its assets include nonsaleable human capital; and frictional and behavioral considerations may affect timing of a mortgage default. So the "liquidation value" in the household case may be viewed more generally as an abstract stochastic latent distance to default that is influenced by, or associated with, but not strictly determined by quantifiable economic factors. Extending the analogy further, we also model a household's abstract distance from mortgage prepayment as a latent stochastic process. In a natural way, the time of the event of interest is modeled by the first time when the latent process representing a positive distance from the event crosses the threshold of zero. For this reason, threshold models are also variously known in the literature as "first passage time" or "first hitting time" models.

This dissertation has two primary goals. The applied goal is to quantify associations between mortgage default and prepayment behavior and observable factors, both household-level and macroeconomic, some of which may vary with time. Another goal is to develop statistical methods to infer parameters of time-varying latent stochastic processes of the type which in this application correspond to distance from default or prepayment. Similar methods that identify associations between regressors and threshold crossing times have been developed in the literature for a variety of applications. However, as we argue in the literature review, the established methods for threshold regression do not easily incorporate time-dependent covariates or dependent competing risks as needed in a thorough study of

default and prepayment behavior. In order to address these issues for a broad subset of applications, we limit our focus to the discrete time setting. We consider situations where the latent stochastic process is appropriately modeled as a Gaussian random walk with time varying drift. To model dual competing risks, the random walk is a bivariate Gaussian with time-varying drift and arbitrary correlation. We further both of our primary goals by comparing our new methods against other survival methods for analyzing mortgage data as well as simulated data.

### *1.1.1 Outline of the dissertation*

The remainder of this chapter reviews both the literature on survival analysis methods applied to credit risk modeling as well as the literature on threshold regression. Chapter 2 presents the model for a single absorbing event of interest and derives analytic expressions for the likelihood and score functions and an expectation-maximization algorithm for parameter estimation. Chapter 3 extends that model to dual correlated competing risks. Chapter 4 outlines practical computational methods for obtaining parameter estimates. Chapter 5 reports results of applying these methods to both simulated data as well to the Freddie Mac data set of U.S. residential mortgages. Chapter 6 concludes with discussion and suggestions for future work.

## ***1.2 Literature Review***

### *1.2.1 Survival analysis methods in credit risk modeling*

The literature on credit risk modeling is vast. A comprehensive review is beyond the scope of this dissertation. We focus primarily on mortgage risk.

LaCour-Little (2008) offers a lengthy survey of the literature on mortgage termination risk. Much recent work has been based on the Cox PH model and related methods, as early as Green and Shoven (1986), who quantify the intuitive negative association between time-varying interest rates and mortgage prepayment. Schaffer (2011) applies the Cox PH model

and both default and prepayment risks (as individual, non-competing risks under separate models) subject to time-varying macroeconomic covariates, and also finds nonparametric estimates of the baseline hazard functions. Demyanyk and Van Hemert (2011) examine default behavior of subprime mortgages leading up to the crisis of 2007, using a discrete proportional odds (multiperiod logistic) model. Deng et al. (2000) also use a proportional hazards model, incorporating methods of Han and Hausman (1990) and Sueyoshi (1992) to jointly estimate mortgage default and prepayment as dependent competing risks along with nonparametric estimates of baseline hazard rates and unobserved heterogeneity. Deng et al. (2000) report that default (prepayment) rates are strongly associated with the "in-the-moneyness" of the mortgages' embedded put (call) options; that failing to estimate the two risks jointly leads to serious forecast errors; that there is considerable heterogeneity among mortgage borrowers; and that baseline loan-to-value, and time-varying unemployment and divorce rates have significant effects on default and/or prepayment.

Other mortgage studies use a multinomial logit (MNL) model to represent the competing risks of default and prepayment. Clapp et al. (2001) apply both MNL and Cox PH models and innovate by distinguishing prepayments due to refinancing from prepayments due to sale of the property, and estimate the competing risks of default, sale and refinance. Clapp et al. (2006) build on Clapp et al. (2001), incorporating unobserved heterogeneity as in Deng et al. (2000) and report that Cox PH with unobserved heterogeneity has the best model fit and outperforms on out-of-sample tests. Watkins et al. (2013) apply a MNL model to unsecured consumer credit loans, innovatively including in the model that such loans have a finite term, i.e. non-stochastic endpoint, where in earlier studies such terminations are incorrectly modeled as right censored. This correction is at least theoretically important for mortgage modeling, although readily available data sets tend to contain few if any observed maturations.

Bajari et al. (2013) take a very different approach and apply a discrete-time, discrete-choice dynamical structural model to the borrower's decision at each time point whether to default, prepay or make a regularly scheduled payment.

Zhang (2009) explores dual-time survival analysis in relation to credit risk, where one time scale is calendar time of origination ("vintage"), and the other times scale is lifetime, or elapsed time since origination. They describe extending a first-crossing-time model (based on a univariate Wiener process) to incorporate time-varying exogenous effects by a time transformation using a subordinate process. However they do not implement this method.

A substantial body of credit risk literature, influenced by Black and Cox (1976), deals with first passage time models of default. These models are variants of Merton (1974)'s original structural model, where default can occur when firm value passes a non-zero, possibly stochastic threshold. Metzler (2008) provides an extensive review of this literature. There are clear parallels to the work we present here. One major difference though, is that the corporate structural default models, such as the KMV model described by McNeil et al. (2010), generally focus on estimating the value of the firm, which is both quantifiable and may be reasonably estimated from periodically reported accounting statements and market values. The probability of default can then be estimated conditional on the estimated firm value. The firm value is used here as what is often termed in other contexts a "marker" variable, or what Cox (1999) also calls an "intermediate" or "surrogate response" variable (as opposed to a covariate which represents an exogenous influence). In our data set, and in other readily conceivable applications, there are no such markers, or other post-baseline proxies for an individual subject's status until event time. The difference is subtle, but an important design consideration, which we elaborate upon more below.

### *1.2.2 Threshold Regression*

Although threshold regression has not been applied or explored nearly as broadly as other survival methods, its body of literature has grown impressively in recent years. Lee and Whitmore (2006) provide a useful overview of the literature to that date. Aalen and Gjessing (2001) offer this basic foundational argument for using first-passage-time models in addition to, or in place of, the more popular hazard rate models:

> What is however, usually disregarded in the standard approach to survival analysis is that the event is the end point of some process. Apart from pure accidents the events do not happen out of the blue, but there is a development preceding each event. (p. 1)

They also argue that the stochastic process first-passage-time approach can help explain results where the shape of hazard rate function is difficult to interpret or seemingly paradoxical. They review several examples of first-passage-time models, including reversible and irreversible processes and finite state Markov processes as well as the Wiener process which is most common in the literature. All of the following examples are based on the Wiener process unless otherwise noted.

Threshold regression has been applied in a variety of fields. The earliest commonly cited example is Lancaster (1972), who uses the method to model the duration of labor strikes in the United Kingdom. In this model, the value of the process indexes the difference between the two parties in a dispute, where the absorbing threshold of 0 represents "agreement". Whitmore (1979) studies employee turnover, where the process models employee job attachment and threshold of 0 represents separation. Eaton and Whitmore (1977) report that the Wiener model offered the best fit among several alternative models for studying length of hospitalization for schizophrenia. Doksum and Hóyland (1992) model the usable life of cable insulation under varying and increasing levels of electrical stress. Lee et al. (2009) perform a case-control study relating railroad worker mortality to diesel exhaust exposure. They note that the diffusion process model is better suited to this study than the more popular proportional hazards model, as the data clearly fail to support the proportional hazards assumption. Yu et al. (2009) study onset of sexually transmitted infections in young women, where the Wiener process drift is modeled as a nonparametric function of baseline covariates, including age of sexual debut and number of partners. Li and Anderson (2009) apply threshold regression in what they call the "vitality model" to explore classic problems in demography, including the effect of diet restriction on longevity; and the differences in

age-dependent mortality rates between male and female medflies.

In many applications, the model entails a single unobserved stochastic process representing an abstract distance from the event of interest, and which is influenced by (typically baseline) covariates. In other applications the single process may be at least somewhat observable. The model may also include additional observable "marker" processes which, as with firm value in the corporate default models, provide information on the state of the subject, and are assumed to be highly correlated proxies of the latent degradation process. Whitmore et al. (1998) analyze reduction cells in an aluminum smelter. In their model, a latent process represents distance from catastrophic failure, while marker processes indicate measurable iron contamination and mechanical warpage. The objective is to estimate the marker level associated with failure, and the correlation between the marker and latent degradation process for the purpose of predicting residual survival time based on observations of the marker. The marker and degradation are jointly a bivariate Wiener process with unknown but constant drift and covariance. Distinct from many applications with time-dependent covariates, in the training sample the marker process is not observed periodically, but only upon failure or censoring. Lee et al. (2000) add covariates to Whitmore et al. (1998)'s statistical framework and apply the method to data from an AIDS clinical trial, where covariates include treatment dosage and the marker represents CD4 cell count as measured over time.

A number of studies have considered competing risks in the context of threshold regression. Whitmore (1986) represents multiple failure modes using Wiener processes which are conditionally independent on common covariates; and where failure by one mode is treated as non-informative censoring for the other modes. As he acknowledges that he did not possess a solution for computing the distribution of first passage time in the general case of multivariate Brownian motion, he necessarily assumes that the competing risk processes are uncorrelated. (More recently, Metzler (2010) presents a Monte Carlo algorithm for approximating the density of first passage time of a bivariate correlated Brownian motion with arbitrary constant drift). Horrocks and Thompson (2004) and Lindqvist and Skogsrud (2008)

respectively model dependent competing risks as crossings of different thresholds by a single univariate process. Xu et al. (2011) model competing causes of death for human subjects using a Wiener process for a (sole) cause of interest and a Gompertz distribution for time of death by any other cause.

Aalen and Gjessing (2001) suggest that "A further development of the present models to incorporate time-dependent covariates might be of interest" (p. 13). Our search of the literature found few examples of progress toward that goal. Lee et al. (2010b) propose a threshold regression model with time-varying covariates applied to data from the Nurses' Health Study on lung cancer risk. In their model, the infrequently observed covariates determine the subject's expected position (value of the latent degradation process) at observation time, while the post-observation degradation process starting from that expected position is Brownian motion with zero drift and unit variance. Lee et al. (2010b)'s model depends on the assumption that the sequence of observed data (covariates as well as outcome) is Markov, and they acknowledge that this assumption may be difficult to verify in practice.

A Markov assumption might be reasonable in the presence of an observable distance to event or a marker variable. We suggest that the Markov assumption is overly strong for Lee et al. (2010b)'s method to be used in applications where the degradation process is latent and where there is no observable distance or marker process closely associated with the degradation status. In this case the probability of failure at time $t$ given survival at $t-1$ depends on the value of the latent process at $t-1$, the probability distribution of which depends on the value of the process at $t-2$, etc. Thus the outcome process is not Markov, unless it is observable. Moreover, we do not require any probability structure on the time-varying covariates, which may be either subject-specific or environmental and common to all subjects, such as macroeconomic variables. Chen and Hong (2011), for example, list several instances of economic time series which have been shown to be non-Markov, potentially excluding Lee et al. (2010b)'s method from broad econometric applications.

The method we present here does not require a Markov assumption and should apply in general discrete time settings.

Another common question in survival analysis that appears to be as yet little explored under threshold regression is unobserved heterogeneity (frailty). Pennell et al. (2010) address the issue in an analysis of malignant melanoma data. In their model, both the drift of the Wiener process and its initial position are dependent on the same baseline covariates. However, the initial position is also subject to a per-subject random effect. They estimate their parameters using Markov Chain Monte Carlo methods and find by way of simulations that omitting the random effects leads to inaccurate estimates.

The literature contains various methods for model fitting and parameter estimation in a threshold regression. Whitmore (1983) parameterizes the drift as a linear function of baseline covariates and estimates the parameters using an EM algorithm. Li and Lee (2011) propose a model where the Wiener process drift is a semi-parametric function of baseline covariates and time-varying coefficents. They fit the model using nonlinear estimating equations which are solved numerically. Xu et al. (2011) estimate the parameters of their competing risks model using an EM algorithm, along with Louis (1982)'s method for obtaining standard errors.

With the widespread adoption of the Cox PH model, and the relative novelty of threshold regression, there is now some literature comparing the two methods and arguing for the benefits of threshold regression in some applications. Lee et al. (2010a) assert that threshold regression has several benefits relative to Cox PH models, among them: The threshold model is preferred in cases where the proportional hazards assumption appears to be violated; it may help describe the causal mechanisms related to survival; it is more parsimonious in estimating the shape of the hazard function. Furthermore, as Lee and Whitmore (2010) also argue in more detail, any reasonably regular proportional hazards function can be formulated as a threshold model. Thus further insights can be gained while preserving the proportional hazards property.

The literature review reveals that threshold regression is growing in acceptance and application in a number of fields, and there is growing recognition of certain benefits relative to other survival methods. It has not yet however been explored in nearly as much depth

as more established approaches to survival analysis. Specific problems that have been well-studied with other approaches, but as yet not exhaustively under threshold regression include time-varying covariates, frailty, dependent competing risks, interpretation of parameters and model validation.

This dissertation makes the following novel contributions to that literature:

1. A framework for discrete time threshold regression, for applications where events may occur only at fixed times, and therefore the assumption of a continuous distribution of hitting times under a Wiener process does not hold.

2. A general method for incorporating time-varying covariates in threshold regression, without requiring the strong Markov assumption of Lee et al. (2010b).

3. Extension of threshold regression to bivariate correlated competing risks.

4. An analytic representation and efficient algorithm for computing first hitting time probabilities of univariate and bivariate Gaussian random walks with time-varying drift.

5. Application of threshold regression to mortgage termination risk.

# Chapter 2

# METHODS FOR A SINGLE EVENT PROCESS

"Eventlessness has no posts to drape duration on. From nothing to nothing is no time at all", John Steinbeck, *East of Eden*, 1952

## *2.1 Model Basics*

In the case where we are concerned with a single terminating event, the process we use to model a subject's latent, stochastically evolving "health" status is a discrete Gaussian random walk $W$, with constant variance, time-varying drift and positive initial condition, i.e.

$$\begin{aligned} &W_0 > 0, \\ &W_t = W_{t-1} + w_t, \quad (t = 1, 2, \ldots), \\ &w_t \sim N(\mu_t, \sigma^2), \\ &\mu_t = \mathbf{x}_t' \boldsymbol{\beta}, \end{aligned}$$

where $\mathbf{x}_t \in \mathbb{R}^m$ is a vector of time-varying covariates (some of which may be fixed at baseline), $\boldsymbol{\beta} \in \mathbb{R}^m$ and $\sigma$ are unknown parameters to be estimated, and where a status of $W_T < 0$ corresponds to death or other event of interest, i.e. $T = \inf\{t : W_t < 0\}$. The event time may be right-censored, in which case censoring time is $C$. Denote the last time at which the process is observed as $Y$, i.e. $Y = T \wedge C$, and let $\Delta = \mathbf{1}_{T<C} = \mathbf{1}_{W_Y<0}$. Let $\mathbf{X}$ be the $Y \times m$ matrix of all observable covariates, where row $t$ is $\mathbf{x}_t'$.[1]

[1]Implicit in this model is that the innovations $w_t$ occur at uniform unit time intervals. In the next chapter we show how this model is easily extended to allow innovations at arbitrary non-informative time points.

We do not observe the values $W_t$. The only observables are the covariates $\mathbf{X}$, final observation time $Y$ and final status $\Delta$.

For identifiability, at least one of $W_0$ and $\sigma$ must be a fixed constant. In most examples of threshold regression in the literature, $\sigma$ is fixed (arbitrarily) at 1, and $W_0$ is modeled as a function of baseline covariates. However, as Stogiannis and Caroni (2013) allude, it can be difficult to interpret the parameters of such a model. We prefer instead the setup

$$W_0 = 100, \qquad \sigma \text{ is an unknown parameter to be estimated.}$$

We arbitrarily but conveniently choose a scale of 100 for $W_0$, which allows the $\boldsymbol{\beta}$ coefficients to be interpreted in a straightforward and intuitive way in terms of percentage change in the original distance from event. At the same time, $\sigma$ can be seen as a measure of unexplained heterogeneity.

For comparison purposes we also consider after the main exposition models where

$$W_0 = \mathbf{x}_0'\boldsymbol{\gamma}, \qquad \sigma = 1,$$

where $\mathbf{x}_0$ is a vector of baseline covariates and $\boldsymbol{\gamma}$ is an unknown parameter to be estimated.

We define the outcome probabilities as:

$$\psi(\Delta, Y|\mathbf{X}, \boldsymbol{\theta}) = \begin{cases} \mathrm{P}(T > Y \mid \mathbf{X}, \boldsymbol{\theta}) = \mathrm{P}(W_t > 0 \;\forall\; t = 1, \ldots, Y \mid \mathbf{X}, \boldsymbol{\theta}) & \Delta = 0 \\ \mathrm{P}(T = Y \mid \mathbf{X}, \boldsymbol{\theta}) = \mathrm{P}(W_Y < 0; W_t > 0 \;\forall\; t = 1, \ldots, Y-1 \mid \mathbf{X}, \boldsymbol{\theta}) & \Delta = 1 \end{cases}.$$

From here on we may drop the conditioning information from $\psi(\cdot)$ and abbreviate as $\psi(\Delta, Y)$ when the context is clear.

We assume that censoring is non-informative, that

$$\mathrm{P}(C = t \mid x_1, \ldots, x_t, w_1, \ldots, w_t, C \geq t) = \mathrm{P}(C = t \mid x_1, \ldots, x_t, C \geq t),$$

and that the time-varying covariates are external to the subjects. Then the likelihood of the complete sample is

$$L \propto \prod_{i=1}^{N} \psi(1, Y_i \mid \mathbf{X}_i, \boldsymbol{\theta})^{\Delta_i} \times \psi(0, Y_i \mid \mathbf{X}_i, \boldsymbol{\theta})^{1-\Delta_i}.$$

We now show that every $\psi(\Delta, Y)$ can be expressed as the orthant probability of a certain $Y$-dimensional multivariate normal, in a form which facilitates analytic derivation of the score and gradient vectors and expressions for an expectation-maximization algorithm. In Chapter 4 we describe a numerical algorithm for computing these probabilities.

Define $\mathbf{w} = (w_1, w_2, \ldots, w_Y)', \quad \boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_Y)'$. Since each innovation is $w_t \sim N(\mu_t, \sigma^2)$ and $w_s \perp w_t$ for $s \neq t$, then clearly $\mathbf{w} \sim MVN_Y(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_Y)$.

We show that there exist $Y \times Y$ matrices $\boldsymbol{\Sigma}_{Y,\Delta}$ and vectors $\mathbf{b}_{Y,\Delta} \in \mathbb{R}^Y, \Delta = 0$ or 1, such that

$$\mathrm{P}(\Delta, Y \mid \mathbf{X}, \boldsymbol{\beta}, \sigma, W_0) = \mathrm{P}(z_t > 0, \ \forall\ t = 1, \ldots, Y),$$

where $\mathbf{z} = (z_1, \ldots, z_Y)' = \boldsymbol{\Sigma}_{Y,\Delta}\mathbf{w} + \mathbf{b}_{Y,\Delta}$ such that $z \sim MVN_Y(\boldsymbol{\Sigma}_{Y,\Delta}\boldsymbol{\mu} + \mathbf{b}_{Y,\Delta}, \sigma^2 \boldsymbol{\Sigma}_{Y,\Delta}\boldsymbol{\Sigma}'_{Y,\Delta})$.

To show this result, for the event of surviving at $Y$, that is $\Delta = 0$, define

$$\begin{aligned} z_1 &= W_0 + w_1, \\ z_2 &= W_0 + w_1 + w_2, \\ &\cdots \\ z_Y &= W_0 + w_1 + w_2 + \ldots + w_Y. \end{aligned}$$

The event of survival at $Y$ is equivalent to $z_t > 0, \ \forall\ t = 1, \ldots, Y$. Therefore we can write

$$\mathbf{0} \leq \mathbf{z} = \boldsymbol{\Sigma}_{Y,0}\mathbf{w} + \mathbf{b}_{Y,0},$$

where

$$\boldsymbol{\Sigma}_{Y,0} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \cdots & & \cdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}_{Y \times Y,}$$

$$\mathbf{b}_{Y,0} = (W_0, W_0, \ldots, W_0)'_Y = W_0 \mathbf{u}_{Y,0}, \qquad \mathbf{u}_{Y,0} \equiv (1, 1, \ldots, 1)'_Y.$$

Note that

$$\boldsymbol{\Sigma}_{Y,0}\boldsymbol{\Sigma}'_{Y,0} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 2 & 2 & \cdots & 2 & 2 \\ 1 & 2 & 3 & \cdots & 3 & 3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & 3 & \cdots & T-1 & T-1 \\ 1 & 2 & 3 & \cdots & T-1 & T \end{pmatrix},$$

and

$$(\boldsymbol{\Sigma}_{Y,0}\boldsymbol{\Sigma}'_{Y,0})^{-1} = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix}.$$

Computations can be simplified by noting that $(\boldsymbol{\Sigma}_{Y,0}\boldsymbol{\Sigma}'_{Y,0})^{-1}$ is tri-diagonal. Similarly, for the event of failing at $Y$, define

$$\begin{aligned} z_1 &= W_0 + w_1, \\ z_2 &= W_0 + w_1 + w_2, \\ &\cdots \\ z_Y &= -W_0 - w_1 - w_2 - \ldots - w_T. \end{aligned}$$

The event of failing at $Y$ is equivalent to $z_t > 0,\ \forall\ t = 1, \ldots, Y$. Therefore we can write

$$\mathbf{0} \leq \mathbf{z} = \boldsymbol{\Sigma}_{Y,1}\mathbf{w} + \mathbf{b}_{Y,1},$$

where

$$\boldsymbol{\Sigma}_{Y,1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ \cdots & & \cdots & & & \\ 1 & 1 & 1 & \cdots & 1 & 0 \\ -1 & -1 & -1 & \cdots & -1 & -1 \end{pmatrix}_{Y \times Y,}$$

$$\mathbf{b}_{Y,1} = (W_0, W_0, \ldots, W_0, -W_0)'_Y = W_0 \mathbf{u}_{Y,1}, \qquad \mathbf{u}_{Y,1} = (1, 1, \ldots, 1, -1)'_Y.$$

Note that

$$\boldsymbol{\Sigma}_{Y,1}\boldsymbol{\Sigma}'_{Y,1} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & -1 \\ 1 & 2 & 2 & \cdots & 2 & -2 \\ 1 & 2 & 3 & \cdots & 3 & -3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & 3 & \cdots & T-1 & -(T-1) \\ -1 & -2 & -3 & \cdots & -(T-1) & T \end{pmatrix},$$

and

$$(\boldsymbol{\Sigma}_{Y,1}\boldsymbol{\Sigma}'_{Y,1})^{-1} = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 2 & 1 \\ 0 & 0 & \cdots & 0 & 1 & 1 \end{pmatrix}.$$

Observe that every such $\boldsymbol{\Sigma}_{Y,\Delta}$ is non-singular with $\det \boldsymbol{\Sigma}_{Y,\Delta} = (-1)^{\Delta}$.

We proceed to derive expressions for estimating $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$ given a set of observed outcomes $\{Y_i, \Delta_i\}$ and a corresponding set of covariates $\mathbf{X}_i$ for subjects $i = 1, \ldots, N$.

## 2.2 Parameter Estimation

### 2.2.1 Likelihood and Score Functions

We derive expressions for estimating parameters by maximum likelihood using either gradient based methods or an expectation-maximization algorithm. We describe our method for obtaining standard errors for these estimates.

Consider the density of the multivariate normal for a single subject conditioned on observed data and parameters

$$\phi(\mathbf{z} \mid \boldsymbol{\theta}, \mathbf{X}, Y, \Delta) = (2\pi)^{-Y/2} \det(\sigma^2 \boldsymbol{\Sigma}_{Y,\Delta} \boldsymbol{\Sigma}'_{Y,\Delta})^{-1/2} \exp\Big\{ -\frac{1}{2}(\mathbf{z}-\boldsymbol{\nu})'(\sigma^2 \boldsymbol{\Sigma}_{Y,\Delta} \boldsymbol{\Sigma}'_{Y,\Delta})^{-1}(\mathbf{z}-\boldsymbol{\nu})\Big\}. \tag{2.1}$$

where

$$\boldsymbol{\nu} = \boldsymbol{\Sigma}_{Y,\Delta}\mathbf{X}\boldsymbol{\beta} + W_0 \mathbf{u}_{Y,\Delta}.$$

For readability, we will write $\boldsymbol{\Sigma}$ and $\mathbf{u}$ without subscripts when $Y$ and $\Delta$ should be clear from the context. The likelihood function for a single observation is

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{X}, Y, \Delta) &= \int_{\mathbf{z}>0} \phi(\mathbf{z} \mid \boldsymbol{\theta}, \mathbf{X}, Y, \Delta) d\mathbf{z} \\ &= \int_{\mathbf{z}>0} (2\pi)^{-Y/2} \sigma^{-Y} \exp\Big\{ -\frac{1}{2\sigma^2}(\mathbf{z}-\boldsymbol{\nu})'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu})\Big\} d\mathbf{z}. \end{aligned}$$

The corresponding score functions are

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{L(\boldsymbol{\theta})}\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{1}{L(\boldsymbol{\theta})}\int_{\mathbf{z}>\mathbf{0}} \frac{\partial}{\partial \boldsymbol{\theta}} \phi(\mathbf{z} \mid \boldsymbol{\theta}, \mathbf{X}, Y, \Delta) d\mathbf{z},$$

then

$$\begin{aligned} \frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \frac{1}{L(\boldsymbol{\theta})} \int_{\mathbf{z}>\mathbf{0}} (2\pi\sigma^2)^{-Y/2} \frac{1}{\sigma^2} \mathbf{X}'\boldsymbol{\Sigma}'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu}) \\ &\qquad \times \exp\Big\{ -\frac{1}{2\sigma^2}(\mathbf{z}-\boldsymbol{\nu})'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu})\Big\} d\mathbf{z} \\ &= \int_{\mathbf{z}>\mathbf{0}} \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\nu}) \frac{\phi(\mathbf{z} \mid \boldsymbol{\theta}, \mathbf{X}, Y, \Delta)}{L(\boldsymbol{\theta})} d\mathbf{z} \\ &= \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{\Sigma}^{-1} E(\mathbf{z}-\boldsymbol{\nu} \mid \boldsymbol{\theta}, \mathbf{X}, Y, \Delta, \mathbf{z}>\mathbf{0}). \end{aligned}$$

Therefore, given the constraints on $\mathbf{z}$ defined by the observed outcomes, the score of $\boldsymbol{\beta}$ invokes the mean of the truncated multivariate normal distribution. Appendix A provides moments of a truncated multivariate normal. We use the result from (A.1) that for $\mathbf{z} \sim MVN(\boldsymbol{\nu}, \boldsymbol{\Omega}),\ E(\mathbf{z} \mid \mathbf{z} > 0) = \boldsymbol{\Omega}\mathbf{f} + \boldsymbol{\nu}$, where $\mathbf{f} \in \mathbb{R}^Y$ is the vector of univariate marginal densities for $\mathbf{z}$, evaluated at $z_t = 0,\ (t = 1, \ldots, Y)$. In this case

$$E(\mathbf{z} - \boldsymbol{\nu} \mid \mathbf{z} > 0) = \boldsymbol{\Omega}\mathbf{f} = (\sigma^2 \boldsymbol{\Sigma}\boldsymbol{\Sigma}')\mathbf{f}.$$

Thus,

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\sigma^2(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')\mathbf{f} = \mathbf{X}'\boldsymbol{\Sigma}'\mathbf{f}. \tag{2.2}$$

Similarly,

$$\begin{aligned}
\frac{\partial \log L}{\partial \sigma} &= \frac{1}{L(\boldsymbol{\theta})} \int_{\mathbf{z}>\mathbf{0}} \Big\{ (2\pi)^{-Y/2}(-Y\sigma^{-Y-1}) + (2\pi\sigma^2)^{-Y/2}\frac{1}{\sigma^3}(\mathbf{z}-\boldsymbol{\nu})'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu}) \Big\} \\
&\qquad \times \exp\Big\{ -\frac{1}{2\sigma^2}(\mathbf{z}-\boldsymbol{\nu})'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu}) \Big\} d\mathbf{z} \\
&= \int_{\mathbf{z}>\mathbf{0}} \Big\{ \frac{-Y}{\sigma} + \frac{1}{\sigma^3}(\mathbf{z}-\boldsymbol{\nu})'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu}) \Big\} \frac{\phi(\mathbf{z} \mid \boldsymbol{\theta}, \mathbf{X}, Y, \Delta)}{L(\boldsymbol{\theta})} d\mathbf{z} \\
&= \frac{-Y}{\sigma} + \frac{1}{\sigma^3} E\{(\mathbf{z}-\boldsymbol{\nu})'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu}) \mid \boldsymbol{\theta}, \mathbf{X}, Y, \Delta, \mathbf{z} > \mathbf{0}\}.
\end{aligned}$$

To evaluate the expectation term, use the identity that for a quadratic form in constant matrix $\mathbf{A}$ and random vector $\mathbf{x}$,

$$E(\mathbf{x}'\mathbf{A}\mathbf{x}) = \mathbf{tr}\{\mathbf{A}E(\mathbf{x}\mathbf{x}')\} = \mathbf{tr}\{\mathbf{A}\mathrm{cov}(\mathbf{x})\} + E(\mathbf{x})'\mathbf{A}E(\mathbf{x}). \tag{2.3}$$

This follows easily from the facts that the trace of a product is invariant under cyclic permutation of the factors; and that the expectation of a trace is equal to the trace of the expectation. Appendix A derives the result that

$$E\{(\mathbf{z}-\boldsymbol{\nu})'\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu}) \mid \mathbf{z} > 0\} = Y - \boldsymbol{\nu} \cdot \mathbf{f},$$

where $\boldsymbol{\nu} \cdot \mathbf{f}$ is the scalar product of those vectors. Thus,

$$\begin{aligned}
\frac{\partial \log L}{\partial \sigma} &= \frac{-Y}{\sigma} + \frac{1}{\sigma^3}\sigma^2 E\{(\mathbf{z}-\boldsymbol{\nu})'\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu}) \mid \boldsymbol{\theta}, \mathbf{X}, Y, \Delta, \mathbf{z} > \mathbf{0}\} \\
&= -\frac{1}{\sigma}\boldsymbol{\nu} \cdot \mathbf{f}.
\end{aligned} \tag{2.4}$$

Therefore the score function for an entire sample of subjects $i = 1, \ldots, N$ is

$$\sum_{i=1}^{N} \begin{pmatrix} \mathbf{X}_i' \boldsymbol{\Sigma}_i' \mathbf{f}_i \\ -\boldsymbol{\nu}_i \cdot \mathbf{f}_i / \sigma \end{pmatrix}.$$

### *2.2.2 Expectation-maximization algorithm*

We derive expressions for estimating the unknown parameters $(\boldsymbol{\beta}, \sigma)$ using the expectation-maximization algorithm of Dempster et al. (1977).

*Expectation step*

For a single observation the complete data likelihood, $L(\boldsymbol{\theta}; \mathbf{z}, \mathbf{X}, Y, \Delta) = P(\mathbf{z}, Y, \Delta \mid \mathbf{X}, \boldsymbol{\theta})$ is the multivariate normal density given above in (2.1). The conditional likelihood of the latent data given the observed information is simply the complete data likelihood normalized by dividing by the probability of the observed,

$$p(\mathbf{z} \mid \boldsymbol{\theta}, \mathbf{X}, Y, \Delta) = \frac{L(\boldsymbol{\theta}; \mathbf{z}, \mathbf{X}, Y, \Delta)}{\mathrm{P}(Y, \Delta \mid \boldsymbol{\theta}, \mathbf{X})};$$

i.e. the denominator is appropriately conditioned $\psi(\Delta, Y)$.

Then the expectation step for a single observation is (again dropping the subscripts $Y, \Delta$ from $\boldsymbol{\Sigma}$ when the context is clear):

$$\begin{aligned}
&E_{\theta_n}\{\log L(\boldsymbol{\theta}; \mathbf{z}, \mathbf{X}, Y, \Delta)\} \\
&= \int_{\mathbf{0}<\mathbf{z}\in\mathbb{R}^Y} \log L(\boldsymbol{\theta}; \mathbf{z}, \mathbf{X}, Y, \Delta) p(\mathbf{z} \mid \boldsymbol{\theta}_n, \mathbf{X}, Y, \Delta) d\mathbf{z} \\
&= \int_{\mathbf{0}<\mathbf{z}} \Big( -\frac{Y}{2}\log(2\pi) - Y \log\sigma - \frac{1}{2\sigma^2}(\mathbf{z}-\boldsymbol{\nu})'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu})\Big) p(\mathbf{z} \mid \boldsymbol{\theta}_n, \mathbf{X}, Y, \Delta) d\mathbf{z}.
\end{aligned}$$

Letting $\boldsymbol{\theta}_n = (\boldsymbol{\beta}_n, \sigma_n)$ be the estimates computed in step $n$ and $\boldsymbol{\nu}_n, \mathbf{f}_n$ be the respective quantities derived therefrom, with the unsubscripted variables representing the corresponding values computed in the subsequent step, for a single observation:

$$E_{\theta_n}\{\log L(\boldsymbol{\theta}; \mathbf{z}, \mathbf{X}, Y, \Delta)\} = -\frac{Y}{2}\log(2\pi) - Y\log\sigma - \frac{1}{2\sigma^2} E_{\boldsymbol{\theta}_n}\Big\{(\mathbf{z}-\boldsymbol{\nu})'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu})\Big\}. \tag{2.5}$$

Writing $\mathbf{z} - \boldsymbol{\nu} = \mathbf{z} - \boldsymbol{\nu}_n + \boldsymbol{\nu}_n - \boldsymbol{\nu}$ and expanding,

$$\begin{aligned}
E_{\theta_n}\{\log L(\boldsymbol{\theta}; \mathbf{z}, \mathbf{X}, Y, \Delta)\} &= -\frac{Y}{2}\log(2\pi) - Y\log\sigma - \frac{1}{2\sigma^2}E_{\boldsymbol{\theta}_n}\Big\{(\mathbf{z}-\boldsymbol{\nu}_n)'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu}_n) \\
&\qquad + (\boldsymbol{\nu}-\boldsymbol{\nu}_n)'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\boldsymbol{\nu}-\boldsymbol{\nu}_n) + 2(\boldsymbol{\nu}_n-\boldsymbol{\nu})'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\mathbf{z}-\boldsymbol{\nu}_n)\Big\} \\
&= -\frac{Y}{2}\log(2\pi) - Y\log\sigma - \frac{\sigma_n^2}{2\sigma^2}(Y - \boldsymbol{\nu}_n\cdot\mathbf{f}_n) - \frac{2\sigma_n^2}{2\sigma^2}(\boldsymbol{\nu}_n-\boldsymbol{\nu})\cdot\mathbf{f}_n \\
&\qquad - \frac{1}{2\sigma^2}(\boldsymbol{\nu}-\boldsymbol{\nu}_n)'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}(\boldsymbol{\nu}-\boldsymbol{\nu}_n) \qquad (2.6) \\
&= -\frac{Y}{2}\log(2\pi) - Y\log\sigma - \frac{\sigma_n^2}{2\sigma^2}\{Y + \mathbf{f}_n'\boldsymbol{\Sigma}\mathbf{X}(\boldsymbol{\beta}_n - 2\boldsymbol{\beta})\} \\
&\qquad - \frac{1}{2\sigma^2}\{\boldsymbol{\Sigma}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\}'(\boldsymbol{\Sigma}\boldsymbol{\Sigma}')^{-1}\{\boldsymbol{\Sigma}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\} \\
&= -\frac{Y}{2}\log(2\pi) - Y\log\sigma - \frac{\sigma_n^2}{2\sigma^2}\{Y + \mathbf{f}_n'\boldsymbol{\Sigma}\mathbf{X}(\boldsymbol{\beta}_n - 2\boldsymbol{\beta})\} - \frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_n). \qquad (2.7)
\end{aligned}$$

Where (2.6) holds because of the trace result (A.7) given in Appendix A. Summing over all observations $i = 1, \ldots, N$ yields

$$\begin{aligned}
Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n) &= -\Big(\log\sigma + \frac{1}{2}\log(2\pi)\Big)\sum_{i=1}^{N} Y_i - \frac{\sigma_n^2}{2\sigma^2}\Big\{\sum_{i=1}^{N} Y_i + \mathbf{f}_{n,i}'\boldsymbol{\Sigma}_i\mathbf{X}_i(\boldsymbol{\beta}_n - 2\boldsymbol{\beta})\Big\} \\
&\qquad - \frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)'\Big(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\Big)(\boldsymbol{\beta}-\boldsymbol{\beta}_n). \qquad (2.8)
\end{aligned}$$

*Maximization Step*

The maximization step computes $\boldsymbol{\theta}_{n+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$, i.e. $0 = \partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)/\partial\boldsymbol{\theta}$. Differentiating both sides of (2.8) gives

$$0 = \frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)}{\partial\boldsymbol{\beta}} = \frac{\sigma_n^2}{\sigma^2}\sum_{i=1}^{N}\mathbf{X}_i'\boldsymbol{\Sigma}_i'\mathbf{f}_{n,i} - \frac{1}{\sigma^2}\Big(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\Big)(\boldsymbol{\beta}-\boldsymbol{\beta}_n).$$

Therefore

$$\boldsymbol{\beta}_{n+1} = \boldsymbol{\beta}_n + \sigma_n^2\Big(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\Big)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\boldsymbol{\Sigma}_i'\mathbf{f}_{n,i}.$$

Likewise,

$$\begin{aligned} 0 &= \frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)}{\partial \sigma} \\ &= -\frac{1}{\sigma}\sum_{i=1}^{N} Y_i + \frac{\sigma_n^2}{\sigma^3}\Big\{\sum_{i=1}^{N} Y_i + (\boldsymbol{\nu}_{n,i} - 2\boldsymbol{\nu}_i)\cdot \mathbf{f}_{n,i}\Big\} + \frac{1}{\sigma^3}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)'\Big(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\Big)(\boldsymbol{\beta} - \boldsymbol{\beta}_n). \end{aligned}$$

Therefore

$$\sigma_{n+1}^2 = \sigma_n^2 + \frac{\sigma_n^2\Big\{\sum_{i=1}^{N}(\boldsymbol{\nu}_{n,i} - 2\boldsymbol{\nu}_{n+1,i})\cdot \mathbf{f}_{n,i}\Big\} + (\boldsymbol{\beta}_{n+1} - \boldsymbol{\beta}_n)'\Big(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\Big)(\boldsymbol{\beta}_{n+1} - \boldsymbol{\beta}_n)}{\sum_{i=1}^{N} Y_i}.$$

### *2.2.3 Obtaining Standard Errors of Parameter Estimates*

Several methods have been proposed to obtain standard errors of parameters estimated via the expectation-maximization algorithm. Most of these methods (e.g. Oakes (1999) ), if applied here, would require second and higher order moments of the truncated multivariate normal. While we have analytical expressions for second order moments, we do not otherwise need to calculate them in the expectation-maximization algorithm, since we only need to compute the trace of a function of the covariance matrix. The extra computation of the higher moments would be at least $O(Y^4)$ in the number of convolutions and less than desirable in this setting. Straightforward calculation of the Hessian of the log likelihood would similarly require second and higher order moments.

We therefore adopt a suggestion from McLachlan and Krishnan (2007) to obtain standard errors from the sample estimate of the expected information, i.e.

$$\hat{\mathcal{I}}(\hat{\boldsymbol{\theta}}) = N^{-1}\Big\{\sum_{i=1}^{N}\mathbf{s}_i(\hat{\boldsymbol{\theta}})\mathbf{s}_i(\hat{\boldsymbol{\theta}})'\Big\} - \bar{\mathbf{s}}(\hat{\boldsymbol{\theta}})\bar{\mathbf{s}}(\hat{\boldsymbol{\theta}})',$$

where $\mathbf{s}_i(\hat{\boldsymbol{\theta}})$ is the score vector for observed data $i$ evaluated at maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ (such as computed in the last iteration of the expectation-maximization algorithm) and $\bar{\mathbf{s}} = N^{-1}\sum \mathbf{s}_i(\hat{\boldsymbol{\theta}})$.

### *2.2.4 Alternative Parameterizations*

We consider two alternative versions of the primary parameterization we have chosen to focus on, one where the variance is a parameterized function of baseline covariates, the other where the variance is fixed and the initial position is a parameterized function of baseline covariates.

#### *Parameterized Variance*

While we are primarily concerned here with a model where the variance is common to all subjects, we can extend the model by parameterizing the variance for subject $i$ as:

$$\sigma_i = g(\boldsymbol{\alpha}, \mathbf{x}_{0,i})$$

for some link function $g$, baseline covariate vector $\mathbf{x}_{0,i}$ and parameter $\boldsymbol{\alpha}$.

Proceeding from (2.4),

$$\frac{\partial \log L}{\partial \alpha} = \frac{\partial \log L}{\partial \sigma}\frac{\partial \sigma}{\partial \alpha} = -\boldsymbol{\nu} \cdot \mathbf{f}\frac{g'(\alpha)}{g(\alpha)}.$$

A reasonable link function would be the exponential function, $g(\boldsymbol{\alpha}, \mathbf{x}_0) = \exp(\mathbf{x}_0'\boldsymbol{\alpha})$, in which case

$$\frac{\partial \log L}{\partial \alpha} = -\sum_{i=1}^{N}(\boldsymbol{\nu}_i \cdot \mathbf{f}_i)\mathbf{x}_{0,i}. \tag{2.9}$$

A hybrid expectation-maximization algorithm can be performed to estimate $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. For a fixed $\boldsymbol{\alpha}$, one can perform the expectation-maximization algorithm in Section 2.2.2 to compute an estimate of $\boldsymbol{\beta}$. For a fixed $\boldsymbol{\beta}$ one can solve the score equation (2.9) to obtain an estimate of $\boldsymbol{\alpha}$. One can iterate these two steps until convergence.

#### *Fixed Variance, Varying Initial Position*

Another alternative parameterization is to fix the variance arbitrarily but conveniently at 1 and to let the subject's initial position be a parameterized function of baseline covariates,

i.e.

$$W_{0,i} = g(\boldsymbol{\gamma}, \mathbf{x}_{0,i}),$$

for some link function $g$, baseline covariate vector $\mathbf{x}_{0,i}$ and parameter $\boldsymbol{\gamma}$. A common choice in the literature is to use a strictly positive link function to guarantee a positive initial value. We choose a simple linear function

$$W_{0,i} = \mathbf{x}'_{0,i}\boldsymbol{\gamma},$$

which leads to a tractable expectation-maximization algorithm. Furthermore, by imposing a convention that a subject with a negative initial position fails at $t = 1$ with probability 1, we minimize the possibility of parameter estimates that imply negative starting values.

The derivation of the score vector for $\boldsymbol{\gamma}$ follows the derivation of the score vector for $\boldsymbol{\beta}$ at (2.2) and obtains

$$\frac{\partial L}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{N} \mathbf{x}_{0,i}\mathbf{u}'_i\mathbf{f}_i.$$

The derivation of the maximization step for $\boldsymbol{\gamma}$ is similar to the derivation for $\boldsymbol{\beta}$ and the result is

$$\boldsymbol{\gamma}_{n+1} = \boldsymbol{\gamma}_n + \Big(\sum_{i=1}^{N} \mathbf{x}_{0,i}\mathbf{x}'_{0,i}\Big)^{-1}\Big\{\Big(\sigma_n^2 \sum_{i=1}^{N} \mathbf{x}_{0,i}\mathbf{u}'_i\mathbf{f}_{\theta_n,i}\Big) - \Big(\sum_{i=1}^{N} \mathbf{x}_{0,i}\mathbf{u}'_i\boldsymbol{\Sigma}_i^{'-1}\mathbf{X}_i\Big)(\boldsymbol{\beta}_{n+1} - \boldsymbol{\beta}_n)\Big\}.$$

# Chapter 3

# METHODS FOR DUAL COMPETING RISKS

## *3.1 Basic Model*

In this chapter we consider the case where a subject may experience one or both of two terminal events, and where the processes representing progression toward those events may be positively or negatively correlated. The process we use to model the subject's latent, stochastically evolving bivariate "health" status is a discrete 2-dimensional Gaussian random walk $\mathbf{W}$, with constant covariance matrix, time-varying drift and fixed positive initial condition. Labeling the events and associated processes and variables by $a$ and $b$, and using $q$ to represent an arbitrary member of $(a, b)$, $\mathbf{W} = (\mathbf{W}_a, \mathbf{W}_b)$, with $\mathbf{W}_q = (W_{q0}, W_{q1}, \ldots)$. Further define:

$$
\begin{aligned}
W_{a0} > 0, \quad W_{b0} > 0,& \\
\mathbf{W}_t &= \mathbf{W}_{t-1} + \mathbf{w}_t, \quad (t = 1, 2, \ldots), \\
\mathbf{w}_t &\sim MVN_2(\boldsymbol{\mu}_t, \boldsymbol{\Psi}), \\
\boldsymbol{\mu}_t &= (\mu_{at}, \mu_{bt})', \\
\boldsymbol{\Psi} &= \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix}, \\
\mu_{at} &= \mathbf{x}_{at}'\boldsymbol{\beta}_a, \\
\mu_{bt} &= \mathbf{x}_{bt}'\boldsymbol{\beta}_b, \\
\boldsymbol{\beta} &= (\boldsymbol{\beta}_a', \boldsymbol{\beta}_b')'.
\end{aligned}
$$

where $\mathbf{x}_{at} \in \mathbb{R}^{m_a}, \mathbf{x}_{bt} \in \mathbb{R}^{m_b}$ are vectors of time-varying covariates, some of which may be fixed at baseline. $\mathbf{x}_{at}, \mathbf{x}_{bt}$ are possibly equal, but need not be equal or even of equal

length. $\boldsymbol{\beta}_a \in \mathbb{R}^{m_a}, \boldsymbol{\beta}_b \in \mathbb{R}^{m_b}$ and $\sigma_a, \sigma_b, \rho$ are unknown parameters to be estimated, and where a status of $W_{q,T_q} < 0$ corresponds to terminal event of type $q \in \{a, b\}$, i.e $T_q = \inf\{t : W_{q,t} < 0\}$. Once the first terminal event is observed, the process stops and the stopping time is $T = T_a \wedge T_b$. This basic model assumes that all observations/innovations occur at uniform unit time intervals, e.g. the duration between $\mathbf{W}_{t_1}$ and $\mathbf{W}_{t_2}$, with $t_1 < t_2$, is $t_2 - t_1$ time units. Later in this section we extend the model to allow the observations/innovations to occur at arbitrarily and variously spaced time points.

We allow that both terminal events may be observed at the same time, i.e. $T = T_a = T_b$. In Section 3.4 we consider alternative constraints on the event observations. The event times may be right-censored, in which case censoring time is $C$. Denote the last time at which the process is observed as $Y$, i.e. $Y = T \wedge C$, and let

$$\Delta = \begin{cases} 0 & Y = C < T_a \wedge T_b \\ 1 & Y = T_a < T_b \\ 2 & Y = T_b < T_a \\ 3 & Y = T_a = T_b \end{cases} .$$

Let $\mathbf{X}_q$ be the $Y \times m_q$ matrix of all observable covariates corresponding to process $q \in (a, b)$, where row $t$ is $\mathbf{x}'_{qt}$. Let the unsubscripted $\mathbf{X}$ be the block diagonal matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_a & \mathbf{0}_{m_a \times m_b} \\ \mathbf{0}_{m_b \times m_a} & \mathbf{X}_b \end{pmatrix}_{(m_a+m_b)\times(m_a+m_b)} .$$

We do not observe the values $\mathbf{W}_t$. The only observables are the covariates $\mathbf{X}$, final observation time $Y$ and final status value $\Delta$

For identifiability, and for each $q \in (a, b)$, either the variance $\sigma_q^2$ or the initial position $W_{q0}$ must be an *a priori* constant. As in the single event case, for ease of interpretation, we fix the initial position at the arbitrary but convenient value $W_{a0} = W_{b0} = 100$. The covariance matrix $\boldsymbol{\Psi}$ is an unknown parameter, to be estimated from the data.

Define the (implicitly conditional) event likelihood $\psi(\Delta, Y) = P(\Delta, Y | \mathbf{W}_0, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Psi})$.

Consider the vector of innovations through time $Y$, $\mathbf{w} = (w_{a1}, \ldots, w_{aY}, w_{b1}, \ldots, w_{bY})'$. It is a truncated multivariate normal of dimension $2Y$, i.e. $\mathbf{w} \sim MVN_{2Y}(\boldsymbol{\mu}, \boldsymbol{\Psi}_Y)$, subject to certain constraints, described formally below, which ensure that each $\mathbf{W}_t = \mathbf{W}_0 + \sum_{j=1}^{t} \mathbf{w}_j$ is consistent with the observed outcome, and where:

$$\boldsymbol{\mu} = (\mu_{a1}, \ldots, \mu_{aY}, \mu_{b1}, \ldots, \mu_{bY})',$$
$$\boldsymbol{\Psi}_Y = \boldsymbol{\Psi} \otimes \mathbf{I}_Y = \begin{pmatrix} \sigma_a^2 \mathbf{I}_Y & \sigma_{ab} \mathbf{I}_Y \\ \sigma_{ab} \mathbf{I}_Y & \sigma_b^2 \mathbf{I}_Y \end{pmatrix}_{2Y \times 2Y},$$

where $\mathbf{I}_Y$ is the $Y \times Y$ identity matrix and $\otimes$ is the Kronecker product. Although we estimate each of $\sigma_a, \sigma_b, \rho$ directly and not their product as such, some formulas here will be expressed with covariance $\sigma_{ab}$ for compactness of notation.

As in the single event case, in order to facilitate derivation of the score function and expectation-maximization algorithm, it is helpful to express the likelihood of an observation as the orthant probability of a multivariate normal. That is to say the likelihood is the probability mass of the distribution over the region formed by the intersection of the positive half-planes of every dimension. Therefore we find matrices $\boldsymbol{\Sigma}_{Y,\Delta}$ of dimension $2Y \times 2Y$ and vectors $\mathbf{b}_{Y,\Delta} \in \mathbb{R}^{2Y}, \Delta \in \{0, 1, 2, 3\}$, such that

$$\mathbf{z} \equiv (z_{a1}, \ldots, z_{aY}, z_{b1}, \ldots, z_{bY})' \equiv \boldsymbol{\Sigma}_{Y,\Delta} \mathbf{w} + \mathbf{b}_{Y,\Delta},$$

thus

$$\mathbf{z} \sim MVN_{2Y}(\boldsymbol{\Sigma}_{Y,\Delta} \boldsymbol{\mu} + \mathbf{b}_{Y,\Delta}, \boldsymbol{\Sigma}_{Y,\Delta} \boldsymbol{\Psi}_Y \boldsymbol{\Sigma}'_{Y,\Delta}), \tag{3.1}$$

such that $\psi(\Delta, Y) = P\{z_{qt} \geq 0 \; \forall \; 1 \leq t \leq Y, q \in (a, b)\}$.

We also use the notation $\boldsymbol{\nu}_{Y,\Delta} \equiv \boldsymbol{\Sigma}_{Y,\Delta} \boldsymbol{\mu} + \mathbf{b}_{Y,\Delta}$. When $Y$ and $\Delta$ are clear from the context we drop the subscripts from $\boldsymbol{\Sigma}, \mathbf{b}, \boldsymbol{\nu}$.

If the subject has survived event $q$ at time $Y$, then the innovations on the $q$ dimension

are subject to the linear inequality constraints:

$$
\begin{aligned}
0 \le z_{q1} &= W_{q1} = W_{q0} + w_{q1}, \\
0 \le z_{q2} &= W_{q2} = W_{q0} + w_{q1} + w_{q2}, \\
&\cdots \\
0 \le z_{qY} &= W_{qY} = W_{q0} + w_{q1} + w_{q2} + \cdots + w_{qY},
\end{aligned}
$$

i.e. for implicit $Y$, $\mathbf{0} \le \mathbf{z}_q = \mathbf{\Sigma}_s \mathbf{w}_q + W_{q0}\mathbf{u}_s$ where

$$
\mathbf{\Sigma}_s = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \cdots & & \cdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}_{Y \times Y},
$$

$$
\mathbf{u}_s = (1, 1, \ldots, 1)'_Y.
$$

If the subject experiences occurrence of ("failure" by) event $q$ at $Y$, then $W_{qY} < 0$ and all other $W_{qt} \ge 0$, i.e.

$$
\begin{aligned}
0 \le z_{q1} &= W_{q1} = W_{q0} + w_{q1}, \\
0 \le z_{q2} &= W_{q2} = W_{q0} + w_{q1} + w_{q2}, \\
&\cdots \\
0 < z_{qY} &= -W_{qT} = -W_{q0} - w_{q1} - w_{q2} - \cdots - w_{aY}.
\end{aligned}
$$

i.e.

$$
\mathbf{0} \le \mathbf{z}_q = \mathbf{\Sigma}_f \mathbf{w}_q + W_{q0}\mathbf{u}_f,
$$

where

$$\boldsymbol{\Sigma}_f = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \cdots & & \cdots & & \\ 1 & 1 & 1 & \cdots & 0 \\ -1 & -1 & -1 & \cdots & -1 \end{pmatrix}_{Y \times Y},$$

$$\mathbf{u}_f = (1, 1, \ldots, 1, -1)'_Y.$$

Survival by both events produces $\mathbf{z} = \boldsymbol{\Sigma}_{0,Y}\mathbf{w} + \mathbf{b}_{0,Y}$, where

$$\boldsymbol{\Sigma}_{0,Y} = \begin{pmatrix} \boldsymbol{\Sigma}_s & \mathbf{0}_{Y \times Y} \\ \mathbf{0}_{Y \times Y} & \boldsymbol{\Sigma}_s \end{pmatrix}_{2Y \times 2Y},$$

$$\mathbf{b}_{0,Y} = \begin{pmatrix} W_{a0}\mathbf{u}_s \\ W_{b0}\mathbf{u}_s \end{pmatrix}_{2Y \times 1},$$

Failure by event $a$ alone produces $\mathbf{z} = \boldsymbol{\Sigma}_{1,Y}\mathbf{w} + \mathbf{b}_{1,Y}$, where

$$\boldsymbol{\Sigma}_{1,Y} = \begin{pmatrix} \boldsymbol{\Sigma}_f & \mathbf{0}_{Y \times Y} \\ \mathbf{0}_{Y \times Y} & \boldsymbol{\Sigma}_s \end{pmatrix}_{2Y \times 2Y},$$

$$\mathbf{b}_{1,Y} = \begin{pmatrix} W_{a0}\mathbf{u}_f \\ W_{b0}\mathbf{u}_s \end{pmatrix}_{2Y \times 1}.$$

Similarly, failure by event $b$ alone produces $\mathbf{z} = \boldsymbol{\Sigma}_{2,Y}\mathbf{w} + \mathbf{b}_{2,Y}$, where

$$\boldsymbol{\Sigma}_{2,Y} = \begin{pmatrix} \boldsymbol{\Sigma}_s & \mathbf{0}_{Y \times Y} \\ \mathbf{0}_{Y \times Y} & \boldsymbol{\Sigma}_f \end{pmatrix}_{2Y \times 2Y},$$

$$\mathbf{b}_{2,Y} = \begin{pmatrix} W_{a0}\mathbf{u}_s \\ W_{b0}\mathbf{u}_f \end{pmatrix}_{2Y \times 1}.$$

Finally, the event of simultaneous failures $a$ and $b$ at $Y$ can be represented as $\mathbf{z} = \boldsymbol{\Sigma}_{3,Y}\mathbf{w} + \mathbf{b}_{3,Y}$,

where

$$\boldsymbol{\Sigma}_{3,Y} = \begin{pmatrix} \boldsymbol{\Sigma}_f & \mathbf{0}_{Y\times Y} \\ \mathbf{0}_{Y\times Y} & \boldsymbol{\Sigma}_f \end{pmatrix}_{2Y\times 2Y},$$

$$\mathbf{b}_{3,Y} = \begin{pmatrix} W_{a0}\mathbf{u}_f \\ W_{b0}\mathbf{u}_f \end{pmatrix}_{2Y\times 1}.$$

Recalling that the determinant of a triangular matrix is the product of its diagonal entries, note that for any $Y, \det \boldsymbol{\Sigma}_s = 1, \det \boldsymbol{\Sigma}_f = -1$. Consequently, $\det \boldsymbol{\Sigma}_{0,Y} = \det \boldsymbol{\Sigma}_{3,Y} = 1$ and $\det \boldsymbol{\Sigma}_{1,Y} = \det \boldsymbol{\Sigma}_{2,Y} = -1$.

In general, we will write

$$\boldsymbol{\Sigma}_{\Delta,Y} = \begin{pmatrix} \boldsymbol{\Sigma}_a & \mathbf{0}_{Y\times Y} \\ \mathbf{0}_{Y\times Y} & \boldsymbol{\Sigma}_b \end{pmatrix}_{2Y\times 2Y},$$

such that each $\boldsymbol{\Sigma}_q = \boldsymbol{\Sigma}_s$ or $\boldsymbol{\Sigma}_f$ as appropriate. Consequently,

$$\boldsymbol{\Sigma}^{-1}_{\Delta,Y} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1}_a & \mathbf{0}_{Y\times Y} \\ \mathbf{0}_{Y\times Y} & \boldsymbol{\Sigma}^{-1}_b \end{pmatrix},$$

where

$$\boldsymbol{\Sigma}^{-1}_s = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots \\ & \cdots & & & \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix}_{Y\times Y},$$

$$\boldsymbol{\Sigma}^{-1}_f = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots \\ & \cdots & & & \\ 0 & \cdots & 0 & -1 & -1 \end{pmatrix}.$$

The likelihood for a single observation is thus

$$\int_{dz_{bY}=0}^{\infty} \cdots \int_{dz_{b1}=0}^{\infty} \int_{dz_{aY}=0}^{\infty} \cdots \int_{dz_{a1}=0}^{\infty} \phi(\mathbf{z}|Y, \Delta, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Psi}) dz_{a1} \cdots dz_{aY} dz_{b1} \cdots dz_{bY},$$

where $\phi(\cdot)$ is the density of a $2Y$-dimensional multivariate normal distribution, i.e. the likelihood is

$$\int_{\mathbf{z}>0} (2\pi)^{-Y} |\boldsymbol{\Omega}|^{-1/2} \exp\Big\{ -\frac{1}{2}(\mathbf{z}-\boldsymbol{\nu})\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu})' \Big\} d\mathbf{z}, \tag{3.2}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{\Delta,\mathbf{Y}} \boldsymbol{\Psi}_Y \boldsymbol{\Sigma}'_{\Delta,\mathbf{Y}}$ and $\boldsymbol{\nu} = \boldsymbol{\Sigma}_{\Delta,Y} \mathbf{X}\boldsymbol{\beta} + \mathbf{b}_{\Delta,Y}$.

The unknown parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_a, \sigma_b, \rho)'$ can be estimated by numerically maximizing the sample likelihood. In the next sections of this chapter we derive the score vector to be used in gradient-based optimization and for computing standard errors. We also present an expectation-maximization algorithm for finding the maximum likelihood estimator.

Before proceeding to the derivations, we extend the model to consider arbitrary and varying time intervals between successive observations/innovations. We assume there are $Y$ discrete post-baseline observations of a subject's covariates and its status vis-a-vis events $a$ and $b$, where the final, $Y$th observation is the stopping time due to event occurrence or censoring. We further assume, similar to assumptions about independent censoring, that observation times are administrative or otherwise distributed independently of the outcome. Now let $\ell_t$ be the length of the time interval between observation/innovation $t-1$ and $t$, with $1 \leq t \leq Y$. Then if $\boldsymbol{\Psi}$ is the covariance of an innovation over a unit time interval, then the covariance of an innovation over a time interval of length $\ell_t$ is $\ell_t \boldsymbol{\Psi}$. Similarly, the elements of $\boldsymbol{\beta}$ represent the covariates' effects per unit time interval, while the elements of $\mathbf{x}_t$ represent the magnitudes of the underlying phenomena suitably weighted for the time interval of length $\ell_t$. This way, the parameters lend themselves to uniform, per-unit time interpretation for every subject, each of which may have unique and variable observation times.

Letting $\mathbf{L}$ be the $Y \times Y$ diagonal matrix where the $i,i$ element is $\ell_i$, the covariance matrix

of the multivariate normal $\mathbf{w}$ is $\mathbf{\Psi}_Y = \mathbf{\Psi} \otimes \mathbf{L}$ and the covariance matrix for $\mathbf{z}$ is:

$$\mathbf{\Omega} = \mathbf{\Sigma}_{\Delta,Y}(\mathbf{\Psi} \otimes \mathbf{L})\mathbf{\Sigma}'_{\Delta,Y}. \tag{3.3}$$

Clearly, the basic scenario of uniform unit intervals is a special case of (3.3), with $\mathbf{L} = \mathbf{I}_Y$. Henceforth we assume that the observations/innovations are observed at arbitrary time intervals, i.e. every covariance matrix $\mathbf{\Omega}$ is of form (3.3).

### *3.2 Derivation of Score Vector*

Let $L(\boldsymbol{\theta}; \mathbf{X}, \Delta, Y)$ represent the likelihood of the observed outcome for a single subject. The form is the same whether the observed outcome at $Y$ is survival or failure. We drop the $\mathbf{X}, \Delta, Y$ when clear from the context.

$$\begin{aligned}
\frac{\partial \log L}{\partial \boldsymbol{\theta}} &= \frac{1}{L(\boldsymbol{\theta})}\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{1}{L(\boldsymbol{\theta})}\int_{\mathbf{z}>\mathbf{0}} \frac{\partial}{\partial \boldsymbol{\theta}}\phi(\mathbf{z}, \boldsymbol{\theta})d\mathbf{z}, \\
\frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \frac{1}{L(\boldsymbol{\theta})}\int_{\mathbf{z}>\mathbf{0}} (2\pi)^{-Y}|\mathbf{\Omega}|^{-1/2}\frac{\partial \boldsymbol{\nu}}{\partial \boldsymbol{\beta}}\mathbf{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu})\exp\Big[-\frac{1}{2\sigma^2}(\mathbf{z}-\boldsymbol{\nu})'\mathbf{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu})\Big]d\mathbf{z} \\
&= \int_{\mathbf{z}>\mathbf{0}} \mathbf{X}'\mathbf{\Sigma}'\mathbf{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu})\frac{\phi(\mathbf{z}, \boldsymbol{\theta})}{L(\boldsymbol{\theta})}d\mathbf{z} \\
&= \mathbf{X}'\mathbf{\Sigma}'\mathbf{\Omega}^{-1}E[\mathbf{z}-\boldsymbol{\nu}|\boldsymbol{\theta}, \mathbf{z}>\mathbf{0}] \\
&= \mathbf{X}'\mathbf{\Sigma}'\mathbf{\Omega}^{-1}\mathbf{\Omega}\mathbf{f} = \mathbf{X}'\mathbf{\Sigma}'\mathbf{f}.
\end{aligned}$$

The last line follows from the expectation result of Eq. A.1, where $\mathbf{f} = (\mathbf{f}'_a, \mathbf{f}'_b)'$, the vector of length $2Y$, consisting of the $Y$ univariate marginal densities for $a$ followed by those for $b$,

evaluated at $z_{qt} = 0$ for each $q \in (a, b), t = 1 \ldots Y$.

$$\begin{aligned}\frac{\partial \log L}{\partial \sigma_a} &= \frac{1}{L(\boldsymbol{\theta})} \int_{\mathbf{z}>\mathbf{0}} \frac{\partial}{\partial \sigma_a} \Big\{ (2\pi)^{-Y} \Big[ (\sigma_a^2 \sigma_b^2 [1-\rho^2])^Y (\prod_{j=1}^{Y} \ell_j)^2 \Big]^{-1/2} \\ &\qquad \times \exp\Big[ -\frac{1}{2} (\mathbf{z}-\boldsymbol{\nu})' \boldsymbol{\Omega}^{-1} (\mathbf{z}-\boldsymbol{\nu}) \Big] \Big\} d\mathbf{z} \\ &= \int_{\mathbf{z}>\mathbf{0}} \Big\{ \frac{-Y}{\sigma_a} - \frac{1}{2} (\mathbf{z}-\boldsymbol{\nu})' \frac{\partial}{\partial \sigma_a} [\boldsymbol{\Omega}^{-1}] (\mathbf{z}-\boldsymbol{\nu}) \Big\} \frac{\phi(\mathbf{z}, \boldsymbol{\theta})}{L(\boldsymbol{\theta})} d\mathbf{z} \\ &= \frac{-Y}{\sigma_a} - \frac{1}{2} E\Big[ (\mathbf{z}-\boldsymbol{\nu})' \frac{\partial}{\partial \sigma_a} [\boldsymbol{\Omega}^{-1}] (\mathbf{z}-\boldsymbol{\nu}) \Big| \boldsymbol{\theta}, \mathbf{z} > \mathbf{0} \Big].\end{aligned}$$

Similarly

$$\begin{aligned}\frac{\partial \log L}{\partial \sigma_b} &= \frac{-Y}{\sigma_b} - \frac{1}{2} E\Big[ (\mathbf{z}-\boldsymbol{\nu})' \frac{\partial}{\partial \sigma_b} [\boldsymbol{\Omega}^{-1}] (\mathbf{z}-\boldsymbol{\nu}) \Big| \boldsymbol{\theta}, \mathbf{z} > \mathbf{0} \Big], \\ \frac{\partial \log L}{\partial \rho} &= \frac{Y\rho}{1-\rho^2} - \frac{1}{2} E\Big[ (\mathbf{z}-\boldsymbol{\nu})' \frac{\partial}{\partial \rho} [\boldsymbol{\Omega}^{-1}] (\mathbf{z}-\boldsymbol{\nu}) \Big| \boldsymbol{\theta}, \mathbf{z} > \mathbf{0} \Big],\end{aligned}$$

where for $\theta = \sigma_a, \sigma_b, \rho$,

$$\frac{\partial}{\partial \theta} [\boldsymbol{\Omega}^{-1}] = \boldsymbol{\Sigma}'^{-1} \Big[ \frac{\partial \boldsymbol{\Psi}^{-1}}{\partial \theta} \otimes \mathbf{L}^{-1} \Big] \boldsymbol{\Sigma}^{-1}.$$

Define $\mathbf{H}_\theta, \mathbf{H}_{\theta,Y}$ as follows:

$$\begin{aligned}\mathbf{H}_\theta &\equiv \frac{\partial \boldsymbol{\Psi}^{-1}}{\partial \theta}, \\ \mathbf{H}_{\theta,Y} &\equiv \mathbf{H}_\theta \otimes \mathbf{L}^{-1}, \\ \frac{\partial}{\partial \theta} [\boldsymbol{\Omega}^{-1}] &= \boldsymbol{\Sigma}'^{-1} \mathbf{H}_{\theta,Y} \boldsymbol{\Sigma}^{-1}, \\ \mathbf{H}_{\sigma_a} \equiv \frac{\partial \boldsymbol{\Psi}^{-1}}{\partial \sigma_a} &= \frac{1}{1-\rho^2} \begin{pmatrix} -2\sigma_a^{-3} & \rho \sigma_a^{-2} \sigma_b^{-1} \\ \rho \sigma_a^{-2} \sigma_b^{-1} & 0 \end{pmatrix}, \\ \mathbf{H}_{\sigma_b} \equiv \frac{\partial \boldsymbol{\Psi}^{-1}}{\partial \sigma_b} &= \frac{1}{1-\rho^2} \begin{pmatrix} 0 & \rho \sigma_a^{-1} \sigma_b^{-2} \\ \rho \sigma_a^{-1} \sigma_b^{-2} & -2\sigma_b^{-3} \end{pmatrix}, \\ \mathbf{H}_\rho \equiv \frac{\partial \boldsymbol{\Psi}^{-1}}{\partial \rho} &= \frac{2\rho}{(1-\rho^2)^2} \begin{pmatrix} \sigma_a^{-2} & -\rho \sigma_a^{-1} \sigma_b^{-1} \\ -\rho \sigma_a^{-1} \sigma_b^{-1} & \sigma_b^{-2} \end{pmatrix} + \frac{1}{1-\rho^2} \begin{pmatrix} 0 & -\sigma_a^{-1} \sigma_b^{-1} \\ -\sigma_a^{-1} \sigma_b^{-1} & 0 \end{pmatrix} \\ &= \frac{1}{(1-\rho^2)^2} \begin{pmatrix} 2\rho \sigma_a^{-2} & -(1+\rho^2) \sigma_a^{-1} \sigma_b^{-1} \\ -(1+\rho^2) \sigma_a^{-1} \sigma_b^{-1} & 2\rho \sigma_b^{-2} \end{pmatrix}.\end{aligned}$$

Abbreviate, for any $\mathbf{H}_\theta$ when the context is clear

$$\mathbf{H} \equiv \begin{pmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{pmatrix}.$$

Defining $\mathbf{R}^* = E[(\mathbf{z}-\boldsymbol{\nu})(\mathbf{z}-\boldsymbol{\nu})'|\boldsymbol{\theta}, \mathbf{z} > \mathbf{0}]$, and using the result from (2.3)

$$E\Big[(\mathbf{z}-\boldsymbol{\nu})'\frac{\partial}{\partial\theta}[\boldsymbol{\Omega}^{-1}](\mathbf{z}-\boldsymbol{\nu})\Big] = \mathbf{tr}\Big\{\frac{\partial}{\partial\theta}[\boldsymbol{\Omega}^{-1}]E[(\mathbf{z}-\boldsymbol{\nu})(\mathbf{z}-\boldsymbol{\nu})']\Big\},$$

the trace term is:

$$\begin{aligned}\mathbf{tr}(\frac{\partial\boldsymbol{\Omega}^{-1}}{\partial\theta}\mathbf{R}^*) &= \mathbf{tr}(\boldsymbol{\Sigma}^{'-1}[\mathbf{H}_\theta \otimes \mathbf{L}^{-1}]\boldsymbol{\Sigma}^{-1}\mathbf{R}^*) \\ &= \mathbf{tr}([\mathbf{H}_\theta \otimes \mathbf{L}^{-1}]\boldsymbol{\Sigma}^{-1}\mathbf{R}^*\boldsymbol{\Sigma}^{'-1}).\end{aligned}$$

The last step follows from the property that the trace of a product is invariant under cyclic permutation of the factors.

We can partition the product of the rightmost factors into $Y \times Y$ quadrants:

$$\begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix} \equiv \boldsymbol{\Sigma}^{-1}\mathbf{R}^*\boldsymbol{\Sigma}^{'-1}.$$

Similarly, dividing the factor $(\mathbf{H}_\theta \otimes \mathbf{L}_i^{-1})$ into $Y \times Y$ quadrants yields a diagonal matrix in each quadrant, i.e

$$\mathbf{H}_\theta \otimes \mathbf{L}^{-1} \equiv \begin{pmatrix} h_{11}\mathbf{L}^{-1} & h_{12}\mathbf{L}^{-1} \\ h_{12}\mathbf{L}^{-1} & h_{22}\mathbf{L}^{-1} \end{pmatrix}.$$

Expanding from identity (A.6) that $\mathbf{R}^* = \boldsymbol{\Omega} - \boldsymbol{\Omega}\mathbf{K}\mathbf{Q}\boldsymbol{\Omega} + \boldsymbol{\Omega}\mathbf{F}\boldsymbol{\Omega} - \boldsymbol{\Omega}\mathrm{diag}(\mathbf{F}\boldsymbol{\Omega})\mathbf{Q}\boldsymbol{\Omega}$,

$$\begin{aligned}\boldsymbol{\Sigma}^{-1}\mathbf{R}^*\boldsymbol{\Sigma}^{'-1} &= \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Omega} - \boldsymbol{\Omega}\mathbf{K}\mathbf{Q}\boldsymbol{\Omega} + \boldsymbol{\Omega}\mathbf{F}\boldsymbol{\Omega} - \boldsymbol{\Omega}\mathrm{diag}(\mathbf{F}\boldsymbol{\Omega})\mathbf{Q}\boldsymbol{\Omega})\boldsymbol{\Sigma}_i^{'-1} \\ &= (\boldsymbol{\Psi}\otimes\mathbf{L}) - (\boldsymbol{\Psi}\otimes\mathbf{L})\boldsymbol{\Sigma}'\mathbf{K}\mathbf{Q}\boldsymbol{\Sigma}(\boldsymbol{\Psi}\otimes\mathbf{L}) + (\boldsymbol{\Psi}\otimes\mathbf{L})\boldsymbol{\Sigma}'\mathbf{F}\boldsymbol{\Sigma}(\boldsymbol{\Psi}\otimes\mathbf{L}) \\ &\quad - (\boldsymbol{\Psi}\otimes\mathbf{L})\boldsymbol{\Sigma}'\mathrm{diag}(\mathbf{F}\boldsymbol{\Omega})\mathbf{Q}\boldsymbol{\Sigma}(\boldsymbol{\Psi}\otimes\mathbf{L}).\end{aligned}$$

Since $(\mathbf{H}_\theta \otimes \mathbf{L}^{-1})(\mathbf{\Psi} \otimes \mathbf{L}) = (\mathbf{H}_\theta \mathbf{\Psi}) \otimes \mathbf{I}$,

$$\begin{aligned}&(\mathbf{H}_\theta \otimes \mathbf{L}^{-1})\mathbf{\Sigma}^{-1}\mathbf{R}^*\mathbf{\Sigma}'^{-1} \\ &= (\mathbf{H}_\theta \mathbf{\Psi} \otimes \mathbf{I}_Y)\Big\{\mathbf{I}_Y - \mathbf{\Sigma}'\mathbf{K}\mathbf{Q}\mathbf{\Sigma}(\mathbf{\Psi} \otimes \mathbf{L}) + \mathbf{\Sigma}'\mathbf{F}\mathbf{\Sigma}(\mathbf{\Psi} \otimes \mathbf{L}) - \mathbf{\Sigma}'\mathrm{diag}(\mathbf{F}\mathbf{\Omega})\mathbf{Q}\mathbf{\Sigma}(\mathbf{\Psi} \otimes \mathbf{L})\Big\}.\end{aligned} \tag{3.4}$$

Letting $\mathbf{S}_{jk}$, $j, k \in \{1, 2\}$ be the $Y \times Y$ quadrants of the expression inside the $\Big\{ \cdots \Big\}$ above, and defining $s_{jk} = \mathbf{tr}(\mathbf{S}_{jk})$ and $\mathbf{S}^* = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}$, then

$$\mathbf{tr}(\frac{\partial \mathbf{\Omega}_i^{-1}}{\partial \theta}\mathbf{R}) = \mathbf{tr}(\mathbf{H}_\theta \mathbf{\Psi} \mathbf{S}^*).$$

Now to evaluate the contributions from each term of $\mathbf{S}$ to the respective quadrantwise traces $s_{11}, s_{12}, s_{21}, s_{22}$.

First determine the contributions of the terms of $\mathbf{S}$ inside the $\Big\{ \cdots \Big\}$ in (3.4).

The contributions of the $\mathbf{I}_Y$ term is simply $Y$ for each $s_{11}, s_{22}$ and 0 for $s_{12}, s_{21}$

Since every $\mathbf{\Sigma}$ is block diagonal $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_a & 0 \\ 0 & \mathbf{\Sigma}_b \end{pmatrix}$ where $\mathbf{\Sigma}_a, \mathbf{\Sigma}_b \in \{\mathbf{\Sigma}_s, \mathbf{\Sigma}_f\}$, depending on the outcome events, as defined in Section 3.1, then

$$\mathbf{\Sigma}(\mathbf{\Psi} \otimes \mathbf{L}) = \begin{pmatrix} \sigma_a^2 \mathbf{\Sigma}_a \mathbf{L} & \sigma_{ab} \mathbf{\Sigma}_a \mathbf{L} \\ \sigma_{ab} \mathbf{\Sigma}_b \mathbf{L} & \sigma_b^2 \mathbf{\Sigma}_b \mathbf{L} \end{pmatrix},$$

$$\mathbf{\Omega} = \mathbf{\Sigma}(\mathbf{\Psi} \otimes \mathbf{L})\mathbf{\Sigma}' = \begin{pmatrix} \sigma_a^2 \mathbf{\Sigma}_a \mathbf{L} \mathbf{\Sigma}_a' & \sigma_{ab} \mathbf{\Sigma}_a \mathbf{L} \mathbf{\Sigma}_b' \\ \sigma_{ab} \mathbf{\Sigma}_b \mathbf{L} \mathbf{\Sigma}_a' & \sigma_b^2 \mathbf{\Sigma}_b \mathbf{L} \mathbf{\Sigma}_b' \end{pmatrix} \equiv \begin{pmatrix} \mathbf{\Omega}_a & \mathbf{\Omega}_{ab} \\ \mathbf{\Omega}_{ab}' & \mathbf{\Omega}_b \end{pmatrix}.$$

The second and fourth terms of (3.4) are of the form $\mathbf{\Sigma}'\mathbf{D}\mathbf{\Sigma}(\mathbf{\Psi} \otimes \mathbf{L})$ where $\mathbf{D}$ is diagonal, i.e $\mathbf{D} = \begin{pmatrix} \mathbf{D}_a & 0 \\ 0 & \mathbf{D}_b \end{pmatrix}$, with $\mathbf{D}_a, \mathbf{D}_b$ both $Y \times Y$ diagonal.

Then

$$\mathbf{\Sigma}'\mathbf{D}\mathbf{\Sigma}(\mathbf{\Psi} \otimes \mathbf{L}) = \begin{pmatrix} \sigma_a^2 \mathbf{\Sigma}_a' \mathbf{D}_a \mathbf{\Sigma}_a \mathbf{L} & \sigma_{ab} \mathbf{\Sigma}_a' \mathbf{D}_a \mathbf{\Sigma}_a \mathbf{L} \\ \sigma_{ab} \mathbf{\Sigma}_b' \mathbf{D}_b \mathbf{\Sigma}_b \mathbf{L} & \sigma_b^2 \mathbf{\Sigma}_b' \mathbf{D}_b \mathbf{\Sigma}_b \mathbf{L} \end{pmatrix}.$$

In both cases the $\mathbf{D}$ is a product of $\mathbf{Q} = \text{diag}(\mathbf{\Omega})^{-1}$ and another diagonal matrix, $\mathbf{A}$ where $\mathbf{A} = \text{diag}(\mathbf{F\Omega})$ or $\mathbf{A} = \mathbf{K}$ as appropriate.

Since $\mathbf{\Omega}_a = \sigma_a^2 \mathbf{\Sigma}_a \mathbf{L} \mathbf{\Sigma}_a'$, $\mathbf{tr}(\mathbf{\Sigma}_a' \mathbf{D}_a \mathbf{\Sigma}_a \mathbf{L}) = \mathbf{tr}(\mathbf{D}_a \mathbf{\Sigma}_a \mathbf{L} \mathbf{\Sigma}_a') = \sigma_a^{-2} \mathbf{tr}(\mathbf{D}_a \mathbf{\Omega}_a)$ and likewise $\mathbf{tr}(\mathbf{\Sigma}_b' \mathbf{D}_b \mathbf{\Sigma}_b \mathbf{L}) = \sigma_b^{-2} \mathbf{tr}(\mathbf{D}_b \mathbf{\Omega}_b)$, then

$$\mathbf{tr}[\mathbf{\Sigma}' \mathbf{D} \mathbf{\Sigma} (\mathbf{\Psi} \otimes \mathbf{L})_{jk}] = \begin{cases} \mathbf{tr}(\mathbf{D}_a \mathbf{\Omega}_a) & jk = 11 \\ \frac{\sigma_{ab}}{\sigma_a^2} \mathbf{tr}(\mathbf{D}_a \mathbf{\Omega}_a) & jk = 12 \\ \frac{\sigma_{ab}}{\sigma_b^2} \mathbf{tr}(\mathbf{D}_b \mathbf{\Omega}_b) & jk = 21 \\ \mathbf{tr}(\mathbf{D}_b \mathbf{\Omega}_b) & jk = 22 \end{cases}.$$

But since, for $q \in \{a, b\}$, $\mathbf{D}_q = \mathbf{A}_q \mathbf{Q}_q$ and $\mathbf{Q}_{q,ii} = 1/(\mathbf{\Omega}_{q,ii})$, $\mathbf{tr}(\mathbf{D}_a \mathbf{\Omega}_a) = \mathbf{tr}(\mathbf{A}_a)$ and $\mathbf{tr}(\mathbf{D}_b \mathbf{\Omega}_b) = \mathbf{tr}(\mathbf{A}_b)$, so

$$\mathbf{tr}[\mathbf{\Sigma}' \mathbf{D} \mathbf{\Sigma} (\mathbf{\Psi} \otimes \mathbf{L})_{jk}] = \begin{cases} \mathbf{tr}(\mathbf{A}_a) & jk = 11 \\ \frac{\sigma_{ab}}{\sigma_a^2} \mathbf{tr}(\mathbf{A}_a) & jk = 12 \\ \frac{\sigma_{ab}}{\sigma_b^2} \mathbf{tr}(\mathbf{A}_b) & jk = 21 \\ \mathbf{tr}(\mathbf{A}_b) & jk = 22 \end{cases}.$$

For $\mathbf{A} = \mathbf{K}$, $\mathbf{tr}(\mathbf{A}_a) = \boldsymbol{\nu}_a \cdot \mathbf{f}_a, \quad \mathbf{tr}(\mathbf{A}_b) = \boldsymbol{\nu}_b \cdot \mathbf{f}_b$.

For $\mathbf{A} = \text{diag}(\mathbf{\Omega F})$:

$$\mathbf{tr}(\mathbf{A}_a) = \sigma_a^2 F_a + \sigma_{ab} F_{ab},$$
$$\mathbf{tr}(\mathbf{A}_b) = \sigma_b^2 F_b + \sigma_{ab} F_{ab},$$

where $F_a \equiv \mathbf{tr}(\mathbf{\Omega}_a \mathbf{F}_a)$, $F_b \equiv \mathbf{tr}(\mathbf{\Omega}_b \mathbf{F}_b)$, $F_{ab} \equiv \mathbf{tr}(\mathbf{\Sigma}_a' \mathbf{F}_{ab} \mathbf{\Sigma}_b \mathbf{L}) = \mathbf{tr}(\mathbf{\Sigma}_a \mathbf{F}_{ab}' \mathbf{\Sigma}_b' \mathbf{L})$.

Finally, the third term of (3.4) is

$$\boldsymbol{\Sigma}'\mathbf{F}\boldsymbol{\Sigma}(\boldsymbol{\Psi}\otimes\mathbf{L}) = \begin{pmatrix} \sigma_a^2\boldsymbol{\Sigma}_a'\mathbf{F}_a\boldsymbol{\Sigma}_a\mathbf{L} + \sigma_{ab}\boldsymbol{\Sigma}_a'\mathbf{F}_{ab}\boldsymbol{\Sigma}_b\mathbf{L} & \sigma_{ab}\boldsymbol{\Sigma}_a'\mathbf{F}_a\boldsymbol{\Sigma}_a\mathbf{L} + \sigma_b^2\boldsymbol{\Sigma}_a'\mathbf{F}_{ab}\boldsymbol{\Sigma}_b\mathbf{L} \\ \sigma_a^2\boldsymbol{\Sigma}_b'\mathbf{F}_{ab}'\boldsymbol{\Sigma}_a\mathbf{L} + \sigma_{ab}\boldsymbol{\Sigma}_b'\mathbf{F}_b\boldsymbol{\Sigma}_b\mathbf{L} & \sigma_{ab}\boldsymbol{\Sigma}_b'\mathbf{F}_{ab}'\boldsymbol{\Sigma}_a\mathbf{L} + \sigma_b^2\boldsymbol{\Sigma}_b'\mathbf{F}_b\boldsymbol{\Sigma}_b\mathbf{L} \end{pmatrix},$$

$$\mathbf{tr}[\boldsymbol{\Sigma}'\mathbf{F}\boldsymbol{\Sigma}(\boldsymbol{\Psi}\otimes\mathbf{L})_{jk}] = \begin{cases} \sigma_a^2F_a + \sigma_{ab}F_{ab} & jk = 11 \\ \sigma_{ab}F_a + \sigma_b^2F_{ab} & jk = 12 \\ \sigma_a^2F_{ab} + \sigma_{ab}F_b & jk = 21 \\ \sigma_{ab}F_{ab} + \sigma_b^2F_b & jk = 22 \end{cases}.$$

The difference of the third and fourth terms of (3.4) then simplifies to

$$\mathbf{tr}\Big((\boldsymbol{\Sigma}'[\mathbf{F} - \mathrm{diag}(\boldsymbol{\Omega}\mathbf{F})]\boldsymbol{\Sigma})(\boldsymbol{\Psi}\otimes\mathbf{L})_{jk}\Big) = \begin{cases} 0 & jk = 11, 22 \\ (1-\rho^2)\sigma_b^2F_{ab} & jk = 12 \\ (1-\rho^2)\sigma_a^2F_{ab} & jk = 21 \end{cases}.$$

Putting all of the above together in an expression for $\mathbf{S}^*$, where

$$\mathbf{S}^*_{jk} = \mathbf{tr}\Big([\mathbf{I}_Y - \boldsymbol{\Sigma}'\mathbf{K}\mathbf{Q}\boldsymbol{\Sigma}(\boldsymbol{\Psi}\otimes\mathbf{L}) + \boldsymbol{\Sigma}'\mathbf{F}\boldsymbol{\Sigma}(\boldsymbol{\Psi}\otimes\mathbf{L}) - \boldsymbol{\Sigma}'\mathrm{diag}(\mathbf{F}\boldsymbol{\Omega})\mathbf{Q}\boldsymbol{\Sigma}(\boldsymbol{\Psi}\otimes\mathbf{L})]_{jk}\Big),$$

$$\mathbf{S}^* = \begin{pmatrix} Y & 0 \\ 0 & Y \end{pmatrix} - \boldsymbol{\nu}_a\cdot\mathbf{f}_a\begin{pmatrix} 1 & \rho\frac{\sigma_b}{\sigma_a} \\ 0 & 0 \end{pmatrix} - \boldsymbol{\nu}_b\cdot\mathbf{f}_b\begin{pmatrix} 0 & 0 \\ \rho\frac{\sigma_a}{\sigma_b} & 1 \end{pmatrix} + (1-\rho^2)F_{ab}\begin{pmatrix} 0 & \sigma_b^2 \\ \sigma_a^2 & 0 \end{pmatrix}.$$

Next

$$\begin{aligned}
\mathbf{H}_{\sigma_a}\boldsymbol{\Psi} &= \frac{1}{1-\rho^2}\begin{pmatrix}(\rho^2-2)\sigma_a^{-1} & -\rho\sigma_a^{-2}\sigma_b \\ \rho\sigma_b^{-1} & \rho^2\sigma_a^{-1}\end{pmatrix}, \\
\mathbf{H}_{\sigma_b}\boldsymbol{\Psi} &= \frac{1}{1-\rho^2}\begin{pmatrix}\rho^2\sigma_b^{-1} & \rho\sigma_a^{-1} \\ -\rho\sigma_a\sigma_b^{-2} & (\rho^2-2)\sigma_b^{-1}\end{pmatrix}, \\
\mathbf{H}_{\rho}\boldsymbol{\Psi} &= \frac{1}{1-\rho^2}\begin{pmatrix}\rho & -\sigma_a^{-1}\sigma_b \\ -\sigma_a\sigma_b^{-1} & \rho\end{pmatrix}, \\
\mathbf{tr}(\mathbf{H}_{\sigma_a}\boldsymbol{\Psi}\mathbf{S}^*) &= -2\frac{Y}{\sigma_a} + 2\frac{\boldsymbol{\nu}_a\cdot\mathbf{f}_a}{\sigma_a}, \\
\mathbf{tr}(\mathbf{H}_{\sigma_b}\boldsymbol{\Psi}\mathbf{S}^*) &= -2\frac{Y}{\sigma_b} + 2\frac{\boldsymbol{\nu}_b\cdot\mathbf{f}_b}{\sigma_b}, \\
\mathbf{tr}(\mathbf{H}_{\rho}\boldsymbol{\Psi}\mathbf{S}^*) &= 2\frac{\rho Y}{1-\rho^2} - 2\sigma_a\sigma_b F_{ab}.
\end{aligned}$$

Combining these results with other terms of the score vector derived above and summarizing:

$$\begin{aligned}
\frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \mathbf{X}'\boldsymbol{\Sigma}'\mathbf{f}, \\
\frac{\partial \log L}{\partial \sigma_a} &= -\frac{\boldsymbol{\nu}_a\cdot\mathbf{f}_a}{\sigma_a}, \\
\frac{\partial \log L}{\partial \sigma_b} &= -\frac{\boldsymbol{\nu}_b\cdot\mathbf{f}_b}{\sigma_b}, \\
\frac{\partial \log L}{\partial \rho} &= \sigma_a\sigma_b\mathbf{tr}(\boldsymbol{\Sigma}_a'\mathbf{F}_{ab}\boldsymbol{\Sigma}_b\mathbf{L}).
\end{aligned} \tag{3.5}$$

The above expression is for a single observation. The score for the complete sample is then the sum of all per-subject scores.

Note that the score components for $\boldsymbol{\beta}$ and either of the $\sigma$ are identical to the corresponding components of the single event framework.

### *3.3 Expectation-Maximization Algorithm*

While we may use the score function derived above in a gradient-based optimization method, such as L-BFGS-B of Byrd et al. (1995), an expectation-maximization (EM) algorithm may be preferable in some situations. The derivation for the EM algorithm in the bivariate case

follows that for the single event case and also uses results from the derivation of the dual event score function.

### *3.3.1 Expectation Step*

Starting with Equation (2.5) for the expected log likelihood of complete information given observed information of a single subject vulnerable to a single event, and adapting it for the dual event case:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_n) &= E_{\theta_n}\{\log L(\boldsymbol{\theta}; \mathbf{z}, \mathbf{X}, Y, \Delta)\} \\ &= -Y\Big\{\log(2\pi) + \log\sigma_a + \log\sigma_b + \frac{1}{2}\log(1-\rho^2)\Big\} - \sum_{t=1}^{Y}\log\ell_t \\ &\qquad - \frac{1}{2}E_{\boldsymbol{\theta}_n}\Big\{(\mathbf{z}-\boldsymbol{\nu})'\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu})|\mathbf{z}>\mathbf{0}\Big\}. \end{aligned}$$

Writing $\mathbf{z}-\boldsymbol{\nu} = \mathbf{z}-\boldsymbol{\nu}_n + \boldsymbol{\nu}_n - \boldsymbol{\nu}$ and expanding, the expectation term is

$$\begin{aligned} &E_{\boldsymbol{\theta}_n}\Big\{(\mathbf{z}-\boldsymbol{\nu})'\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu})\Big\} \\ &= E_{\boldsymbol{\theta}_n}\Big\{(\mathbf{z}-\boldsymbol{\nu}_n)'\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu}_n) + (\boldsymbol{\nu}-\boldsymbol{\nu}_n)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\nu}-\boldsymbol{\nu}_n) + 2(\boldsymbol{\nu}_n-\boldsymbol{\nu})'\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu}_n)\Big\} \\ &= E_{\boldsymbol{\theta}_n}\Big\{(\mathbf{z}-\boldsymbol{\nu}_n)'\boldsymbol{\Omega}^{-1}(\mathbf{z}-\boldsymbol{\nu}_n)\Big\} + [\boldsymbol{\Sigma}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)]'\boldsymbol{\Omega}^{-1}[\boldsymbol{\Sigma}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)] \\ &\qquad - 2[\boldsymbol{\Sigma}\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)]'\boldsymbol{\Omega}^{-1}E_{\boldsymbol{\theta}_n}(\mathbf{z}-\boldsymbol{\nu}_n) \\ &= \mathbf{tr}\Big\{\boldsymbol{\Omega}^{-1}E_{\boldsymbol{\theta}_n}[(\mathbf{z}-\boldsymbol{\nu}_n)(\mathbf{z}-\boldsymbol{\nu}_n)']\Big\} + (\boldsymbol{\beta}-\boldsymbol{\beta}_n)'\mathbf{X}'[\boldsymbol{\Psi}^{-1}\otimes\mathbf{L}]\mathbf{X}(\boldsymbol{\beta}-\boldsymbol{\beta}_n) \\ &\qquad - 2(\boldsymbol{\beta}-\boldsymbol{\beta}_n)'\mathbf{X}'[\boldsymbol{\Psi}^{-1}\otimes\mathbf{L}^{-1}][\boldsymbol{\Psi}_n\otimes\mathbf{L}]\boldsymbol{\Sigma}'\mathbf{f}_n. \end{aligned}$$

Dropping the constant terms, simplifying and summing over all observations:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_n) = &-\Big(\log\sigma_a + \log\sigma_b + \frac{1}{2}\log(1-\rho^2)\Big)\sum_{i=1}^{N} Y_i \\ &- \frac{1}{2}\sum_{i=1}^{N}\mathbf{tr}\Big\{\boldsymbol{\Omega}_i^{-1}E_{\boldsymbol{\theta}_n}[(\mathbf{z}-\boldsymbol{\nu}_{i,n})(\mathbf{z}-\boldsymbol{\nu}_{i,n})']\Big\} \\ &- \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)'\Big(\sum_{i=1}^{N}\mathbf{X}_i'[\boldsymbol{\Psi}^{-1}\otimes\mathbf{L}_i]\mathbf{X}_i\Big)(\boldsymbol{\beta}-\boldsymbol{\beta}_n) \\ &+ (\boldsymbol{\beta}-\boldsymbol{\beta}_n)'\sum_{i=1}^{N}\mathbf{X}_i'[(\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}_n)\otimes\mathbf{I}_{Y_i}]\boldsymbol{\Sigma}_i'\mathbf{f}_{i,n}. \end{aligned} \tag{3.6}$$

### *3.3.2 Maximization Step*

To find the values of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_a, \sigma_b, \rho)$ which maximize the above expectation:

$$\begin{aligned} 0 = \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}_n)}{\partial\boldsymbol{\beta}} &= -\Big(\sum_{i=1}^{N}\mathbf{X}_i'[\boldsymbol{\Psi}^{-1}\otimes\mathbf{L}_i]\mathbf{X}_i\Big)(\boldsymbol{\beta}-\boldsymbol{\beta}_n) + \sum_{i=1}^{N}\mathbf{X}_i'[(\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}_n)\otimes\mathbf{I}_{Y_i}]\boldsymbol{\Sigma}_i'\mathbf{f}_{i,n}, \\ \boldsymbol{\beta}_{n+1} &= \boldsymbol{\beta}_n + \Big(\sum_{i=1}^{N}\mathbf{X}_i'[\boldsymbol{\Psi}_{n+1}^{-1}\otimes\mathbf{L}_i]\mathbf{X}_i\Big)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'[(\boldsymbol{\Psi}_{n+1}^{-1}\boldsymbol{\Psi}_n)\otimes\mathbf{I}_{Y_i}]\boldsymbol{\Sigma}_i'\mathbf{f}_{i,n}. \end{aligned} \tag{3.7}$$

For $\theta \in (\sigma_a, \sigma_b, \rho)$, and using the definition of $\mathbf{H}_\theta = \frac{\partial}{\partial\theta}\boldsymbol{\Psi}^{-1}$ from the previous section:

$$\begin{aligned} 0 = \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}_n)}{\partial\theta} = &-\frac{1}{2}\sum_{i=1}^{N}\mathbf{tr}\Big\{[\mathbf{H}_{\theta,n+1}\otimes\mathbf{L}_i^{-1}]\boldsymbol{\Sigma}_i^{-1}E_{\boldsymbol{\theta}_n}[(\mathbf{z}-\boldsymbol{\nu}_{i,n})(\mathbf{z}-\boldsymbol{\nu}_{i,n})']\boldsymbol{\Sigma}_i'^{-1}\Big\} \\ &- \frac{1}{2}(\boldsymbol{\beta}_{n+1}-\boldsymbol{\beta}_n)'\Big(\sum_{i=1}^{N}\mathbf{X}_i'[\mathbf{H}_{\theta,n+1}\otimes\mathbf{L}_i]\mathbf{X}_i\Big)(\boldsymbol{\beta}_{n+1}-\boldsymbol{\beta}_n) \\ &+ (\boldsymbol{\beta}_{n+1}-\boldsymbol{\beta}_n)'\sum_{i=1}^{N}\mathbf{X}_i'[(\mathbf{H}_{\theta,n+1}\boldsymbol{\Psi}_n)\otimes\mathbf{I}_{Y_i}]\boldsymbol{\Sigma}_i'\mathbf{f}_{i,n} \\ &+ \sum_{i=1}^{N}Y_i \times \begin{cases} -\frac{1}{\sigma_a} & \theta = \sigma_a \\ -\frac{1}{\sigma_b} & \theta = \sigma_b \\ \frac{\rho}{1-\rho^2} & \theta = \rho \end{cases}. \end{aligned} \tag{3.8}$$

The trace term in (3.8) is similar to the trace evaluated in the variance components of the score, with the difference that here the $\mathbf{S}^*$ matrix entails quantities estimated in the $n$th step

of the EM algorithm, while the $\mathbf{H}_\theta$ factor is with estimates from the $n+1$ step.

To compute the M-step, note that $\boldsymbol{\beta}_{n+1}$ enters Equation (3.8) only in the form of $\boldsymbol{\beta}_{n+1} - \boldsymbol{\beta}_n$, which in turn is expressed in (3.7) in terms of $\boldsymbol{\Psi}_{n+1}$. Thus we can eliminate $\boldsymbol{\beta}_{n+1}$ from (3.8) by substituting the right side of (3.7) in its place. We then have a system of three simultaneous non-linear equations in the three unknowns $\sigma_a, \sigma_b, \rho$. These equations may be solved numerically by commonly available software, such as the R `rootSolve` package of Soetaert (2009), and with suitable transformations of the variables to eliminate the need for bound constraints[1]. This produces $\boldsymbol{\Psi}_{n+1}$ which can be inserted into (3.7) to yield $\boldsymbol{\beta}_{n+1}$.

The computation time needed to numerically solve the M-step is a fraction of the time needed for the most expensive computational steps, i.e. computing the marginal densities $\mathbf{f}_n, \mathbf{F}_n$, which are done only once in the E-step. This extra cost may, in some situations, be a good trade-off for the other advantages of the EM algorithm relative to gradient-based optimization (monotone convergence and adaptibility to acceleration methods).

As in the single event case, whether the MLE is estimated by gradient-assisted optimization or an EM algorithm, standard errors are estimated using the sample estimate of the expected information, as in Section 2.2.3.

### *3.4 Alternative Outcome Scenarios*

Throughout this chapter we have allowed events $a$ and $b$ to occur simultaneously. However we might want to restrict the universe of possible outcomes in different ways. For example, we might designate event $b$ to be "dominant", such that if both processes fall below the threshold concurrently, only outcome $b$ will be observed (the unobserved event, "a", is correspondingly designated "recessive"). Alternatively, we might impose the rule that if both processes fall below the threshold concurrently, then only the outcome which is farthest below the threshold is observed. These scenarios impose different sets of linear constraints on the innovations

[1] $\sigma_q \to \log \sigma_q, \rho \to \rho/(1-\rho^2)$

than the ones considered in Section 3.1, but the derivations of the score functions and EM algorithm closely follow the derivations above.

In the simulations we consider data sets generated under the assumption that both outcome events may be observed simultaneously ("Joint" mode), and also data sets generated with the constraint that one event is "dominant" ("Dominant" mode). For the analysis of the mortgage data, where the default and prepayment events are necessarily mutually exclusive, we assume that default is dominant over prepayment, and also test the sensitivity of the results to alternative assumptions.

### *3.4.1 Dominant Mode*

Assume without loss of generality that the dominant event is $b$ and the final observation time is $Y$. If the subject survives both events at $Y$, the constraint matrix for survival of both events is the same as in the Joint model, $\begin{pmatrix} \mathbf{\Sigma}_s & \mathbf{0}_Y \\ \mathbf{0}_Y & \mathbf{\Sigma}_s \end{pmatrix}$. If the recessive event $a$ is observed, then the subject necessarily survives $b$, so the constraint matrix is also the same is in the Joint model, $\begin{pmatrix} \mathbf{\Sigma}_f & \mathbf{0}_Y \\ \mathbf{0}_Y & \mathbf{\Sigma}_s \end{pmatrix}$. In these cases the subject's covariance matrix $\mathbf{\Omega}$ and its contributions to the likelihood and score are the same as they would be in the Joint model.

On the other hand, if a failure by $b$ is observed, then the only information we have about process $a$ is that it remained above the threshold through time $Y-1$, so only $Y-1$ innovations for $a$ are constrained. The constraints on $b$ are the same as if its failure were observed at $Y$ in Joint mode. So in this case the multivariate normal of interest is of dimension $2Y-1$ and the constraint matrix is $\begin{pmatrix} \mathbf{\Sigma}_{s[(Y-1)\times(Y-1)} & \mathbf{0}_{(Y-1)\times Y} \\ \mathbf{0}_{Y\times(Y-1)} & \mathbf{\Sigma}_{f[Y\times Y]} \end{pmatrix}$.

The covariance matrix is then

$$\mathbf{\Omega} = \mathbf{\Sigma} \begin{pmatrix} (\mathbf{\Psi} \otimes \mathbf{L}_{Y-1}) & \mathbf{0}_{(Y-1)\times 1} \\ \mathbf{0}_{1\times(Y-1)} & \ell_Y \sigma_b^2 \end{pmatrix} \mathbf{\Sigma}'.$$

The subject's contribution to the score function is of the same form as in (3.5), but for differences in dimensions of some quantities, i.e. $\mathbf{f}_a, \boldsymbol{\nu}_a$ are of length $Y-1$, $\mathbf{f}_b, \boldsymbol{\nu}_b$ are of

length $Y$ and $\mathbf{F}_{ab}$ is $(Y-1) \times Y$.

### *3.4.2 Most Negative Mode*

In this mode, only the event corresponding to the most negative process is observed. We do not implement the estimation algorithms for this mode or flesh out the derivations in this dissertation. We present the applicable constraint matrices to illustrate the flexibility of the basic mechanism and to suggest future work and additional applications.

Assume without loss of generality that the observed event is $a$. Then the constraints are:

$$
\begin{aligned}
&0 \leq z_{a1} = W_{a0} + w_{a1}, \\
&0 \leq z_{a2} = W_{a0} + w_{a1} + w_{a2}, \\
&\quad \cdots \\
&0 < z_{aY} = -W_{a0} - w_{a1} - w_{a2} - \cdots - w_{aY}, \\
&0 \leq z_{b1} = W_{b0} + w_{b1}, \\
&0 \leq z_{b2} = W_{b0} + w_{b1} + w_{b2}, \\
&\quad \cdots \\
&0 \leq z_{b,Y-1} = W_{b0} + w_{b1} + w_{b2} + \cdots + w_{b,Y-1}, \\
&0 < z_{bY} = -W_{a0} - w_{a1} - w_{a2} - \cdots - w_{aY} + W_{b0} + w_{b1} + w_{b2} + \cdots + w_{bY},
\end{aligned}
$$

where the final constraint corresponds to $W_{aY} < W_{bY}$. Then:

$$\boldsymbol{\Sigma}_{1,Y} = \begin{pmatrix} \boldsymbol{\Sigma}_f & \mathbf{0}_{Y\times Y} \\ \mathbf{V} & \boldsymbol{\Sigma}_s \end{pmatrix}_{2Y\times 2Y},$$

$$\mathbf{V} = \begin{pmatrix} & \mathbf{0}_{Y-1\times Y} & \\ (-1 & \cdots & -1)_Y \end{pmatrix},$$

$$\mathbf{b}_{1,Y} = \begin{pmatrix} W_{a0}\cdot\mathbf{u}_{f,Y} \\ W_{b0}\cdot\mathbf{u}_{s,[Y-1]} \\ 0 \end{pmatrix}_{2Y\times 1}.$$

# Chapter 4

# COMPUTATION

> "Computer programs can also do useful work.", Donald E. Knuth, *Literate Programming*, 1992

Computation of parameter estimates and standard errors is perhaps the greatest challenge in this dissertation. There are two main areas of computational challenge. One is efficient and precise calculation of the likelihood of first hitting time along with certain marginal densities of the complete data distribution as needed in score functions and EM algorithms; the second is implementing an optimization algorithm that uses the latter results to converge to maximum likelihood estimates as quickly as practical.

## *4.1 Likelihood of First Hitting Time and Marginal Densities*

As shown in Section 2.1, the likelihood that a process of the type we consider fails (survives) at time $Y$ is equivalent to the orthant probability of a $Y$-dimensional multivariate normal. Hayter (2011) shows how a multi-dimensional integral of a multivariate normal, where the integration region is defined by linear inequalities, can be computed recursively as a series of one-dimensional integrals. However the number of one-dimensional integrals that would be required here is impractically large. Other methods for computing orthant probabilities have been proposed in the literature and implemented in R packages, e.g. Genz and Bretz (2009) and Genz et al. (2013)'s `mvtnorm` package; Craig (2008) and Craig (2013)'s `orthants` package. Wilhelm and Manjunath (2013)'s `tmvtnorm` package computes the univariate and bivariate marginal densities of truncated multivariate normals as described in Appendix A. We found however that our own implemention of a specialized algorithm for computing the specific probabilities of interest has several advantages over the aforementioned general

purpose methods; our algorithm is considerably faster and more scalable and allows us to compute the marginal densities in the same execution loop as the hitting time probabilities. A further advantage is that our algorithm is deterministic and more suitable for use with conventional optimization methods than the `mvtnorm` and `tmvtnorm` algorithms, which are stochastic.

Our method is to reformulate the problem as a relatively fast and simple recursion, iteratively evaluating $Y$ lower dimensional integrals. We illustrate the basic concept with the univariate case. The concept extends to the bivariate case in the natural way. We explain the critical details where the bivariate case differ from the univariate case in a later section.

Returning to the recursive definition of our latent process:

$$W_t = W_{t-1} + w_t \quad t \geq 1,$$
$$w_t \sim N(\mu_t, \sigma^2).$$

Implicit in the need to compute $W_t$ is the assumption that the subject has survived at $t-1$, so $W_{t-1} \geq 0$.

A high level description of each iteration of the recursive algorithm, indexed by $t$:

1. From the previous iteration we have the density $g_{t-1}()$ of $W_{t-1}$ discretized over a grid and truncated so that $\forall w < 0,\ g_{t-1}(w) = 0$. More precisely, $g_{t-1}$ is a subdensity, since it integrates to $\mathrm{P}(W_{t-1} \geq 0) \leq 1$. The base case is $g_0(W_0) = 1$.

2. Since $W_t$ is the sum of $W_{t-1}$ and $w_t$, which are independent, its subdensity $g_t()$ is the convolution of the subdensity $g_{t-1}()$ of $W_{t-1}$ with the density $\phi_t()$ of $w_t$.

3. We compute the discretized convolution $g_t = (g_{t-1} * \phi_t)$ using a discrete Fourier transform (DFT).

4. If we need to compute the marginal densities as described in Appendix A, we extract and save $g_t(0)$.

5. If $t < Y$ then $g_t()$ is truncated so that $g_t(w) = 0$ for all $w < 0$; then go on to the next iteration.

6. If $t = Y$ then the probability of survival at $Y$ is $\int_0^\infty g_Y(w)dw$; the probability of failure at $Y$ is $\int_{-\infty}^0 g_Y(w)dw$. These integrals are computed by summing over the elements of the grid of discretized densities.

The next section provides background on discrete convolutions which will help motivate the details of our algorithm and its implementation.

### *4.2 Discrete Convolutions*

In this section we provide background on computing the convolution of two continuous probability distributions using discrete approximations, and define terms and notation to be used throughout this chapter. This section covers convolutions of univariate functions only. Later in the chapter we show how the concepts extend to bivariate functions in a natural way. Numerous sources cover the topic of discrete convolutions using discrete Fourier transforms. This treatment borrows from Smith et al. (1997) and Ruckdeschel and Kohl (2014).

First we consider a discrete approximation for a continuous subdensity $g(\cdot), g : \mathbb{R} \to [0, \infty)$, with subdistribution $G(\cdot)$, i.e.

$$G(y) \equiv \int_{-\infty}^{y} g(z)dz,$$
$$G(\infty) \leq 1.$$

Define a (non-unique) discretization of $g(\cdot)$ to be a tuple

$$\mathcal{D}g = \{\mathbf{g}, n, L, U, \delta\},$$

where $(L, U)$ is a finite interval containing all, or "nearly all", of the subdistribution, $n$ is the number of gridpoints and $\delta$ represents the width of each domain interval corresponding

to a gridpoint, such that for acceptably small $\epsilon \geq 0$,

$$G(\infty) - \epsilon \leq G(U) - G(L) \leq G(\infty),$$
$$\delta = \frac{U - L}{n}.$$

Since $U$ is redundant given $L, n, \delta$, explicitly including it in the tuple is optional.

$\mathbf{g}$ is a vector $\mathbf{g}[0], \ldots, \mathbf{g}[n-1]$ such that

$$\mathbf{g}[i] = \begin{cases} G(L + (i+1)\delta) - G(L + i\delta) & i = 1, \ldots, n-2 \\ G(L + \delta) & i = 0 \\ G(\infty) - G(L + (n-1)\delta) & i = n-1 \end{cases}. \tag{4.1}$$

Thus $\sum_{i=0}^{n-1} \mathbf{g}[i] = G(\infty)$.

To simplify notation, we allow $\mathbf{g}$ to be infinitely "zero-padded", i.e. for any integer $i < 0$ or $i > n-1$, we define $\mathbf{g}[i] = 0$. We may describe $\mathbf{g}$ as having support $\mathbf{g}[0], \ldots, \mathbf{g}[n-1]$, or simply $(0, n-1)$, or equivalently having support of length $n$.

Note that $\mathbf{g}$ represents only the shape of the distribution, while the scale and location properties are defined by $L, n, \delta$.

$\mathcal{D}g$ essentially defines a lattice distribution approximating the continuous $g$.

We can define a discrete *linear* convolution[1] of subdensities $g, h$, given discretizations $\mathcal{D}g = \{\mathbf{g}, n_g, L_g, \delta\}, \mathcal{D}h = \{\mathbf{h}, n_h, L_h, \delta\}$, where $\delta$ is necessarily common to both discretizations, as

$\mathcal{D}g * \mathcal{D}h = \{\mathbf{g} * \mathbf{h}, n_{g*h}, \delta, L_{g*h}\}$.

$\mathcal{D}g * \mathcal{D}h$ also defines a lattice distribution which approximates the continuous $g * h$. Approximation error can be made arbitrarily small by choice of sufficiently large $n$ and sufficiently small $\epsilon$.

The linear convolution of the vector components as such, which is one aspect of, but does

[1]There are other types of convolutions. Unless otherwise stated, the word "convolution" without a modifier refers to a linear convolution.

not fully describe $\mathcal{D}g * \mathcal{D}h$, is

$$(\mathbf{g} * \mathbf{h})[k] = \sum_{j=-\infty}^{\infty} \mathbf{g}[k-j]\mathbf{h}[j] = \sum_{j=0}^{k} \mathbf{g}[k-j]\mathbf{h}[j] \tag{4.2}$$

To fully describe $\mathcal{D}g * \mathcal{D}h$, specifically its location, we must also define the resultant support and lower bound.

The support for $(\mathbf{g} * \mathbf{h})$ is the range of $k$ such that $k - j$ is in the support of $\mathbf{g}$ and $j$ is in the support of $\mathbf{h}$, which is easily seen to be $0, \ldots, n_g + n_h - 2$, therefore $n_{g*h} = n_g + n_h - 1$.

A convolution of continous functions $g, h$ with bounded support $(L_g, U_g)$ and $(L_h, U_h)$ respectively would have support $(L_g + L_h, U_g + U_h)$. But to represent this entire range using a grid spaced by the same $\delta$ as the convolvants would require a vector of length $n_g + n_h$, which is longer by 1 than vector $\mathbf{g} * \mathbf{h}$. Any discrete representation of a continuous function necessarily requires imprecision and trade-offs. We could either preserve the bounds and endow the $\mathcal{D}g * \mathcal{D}h$ with a coarser grid, $\delta^* = \delta \times \frac{n_g+n_h}{n_g+n_h-1}$, or preseve the grid width and shrink the bounds evenly at both tails. In our application, the bounds are arbitrary truncation points at the negligible tails of a Gaussian. Furthermore, there are computational advantages to reuse the same $\delta$ across sequential convolutions. Therefore we take the latter approach and define $L_{g*h} = L_g + L_h + \delta/2$, consequently $U_{g*h} = U_g + U_h - \delta/2$.

To summarize:

$$\mathcal{D}g * \mathcal{D}h = \{\mathbf{g} * \mathbf{h}, n_g + n_h - 1, \delta, L_g + L_h + \delta/2\}$$

It is easy to see that the discrete convolution is commutative as is the continuous convolution.

The discrete convolution has the following useful properties which we exploit for more efficient computation

1. **It is translation-invariant up to convolution of the vector components.**

   For function $g(\cdot)$ define $g^{+c}(\cdot)$ as the translation of $g$ to right by $c$, i.e. $g^{+c}(x+c) = g(x)$, alternatively $g^{+c}(x) = g(x - c)$.

It is easy to see that if $\mathcal{D}g = \{\mathbf{g}, n, L, \delta\}$ discretizes $g$, then $\{\mathbf{g}, n, L{+}c, \delta\}$ discretizes $g^{+c}$. i.e. the vector component is location-invariant, and a discretization can be translated simply by translating the lower bound $L$. It follows, given discretization $\mathcal{D}h$, that $\mathcal{D}g * \mathcal{D}h$ can be translated to represent $\mathcal{D}g^{+c} * \mathcal{D}h$ and $\mathcal{D}g * \mathcal{D}h^{+c}$ without any changes to the vector component, and only by setting $L_{g^{+c}*h} = L_{g*h^{+c}} = L_{g*h} + c$.

2. **It is scale-invariant up to convolution of the vector components**

   For function $g(\cdot)$ and $c > 0$ define $g^{\times c}(\cdot)$ as a scaling of $g$ such that $cg^{\times c}(cx) = g(x)$, alternatively $g^{\times c}(x) = \frac{1}{c}g(x/c)$.

   It follows that $G^{\times c}(cx) = G(x)$ and therefore $G^{\times c}(cx{+}c\delta){-}G^{\times c}(cx) = G(x{+}\delta){-}G(x)$. Accordingly, if $\mathcal{D}g = \{\mathbf{g}, n, L, \delta\}$ discretizes $g$ then $\{\mathbf{g}, n, cL, c\delta\}$ discretizes $g^{\times c}$. In other words the vector component is scale invariant in the sense that a discretization for $g$ can be scaled to a discretization for $g^{\times c}$ by simply multiplying the lower bound and gridwidth by $c$ while preserving $\mathbf{g}$.

   It follows further that a discrete convolution of $g$ and $h$ can be similarly scaled to represent a convolution of identically scaled $g^{\times c}$ and $h^{\times c}$, preserving the vector component. i.e. if $\mathcal{D}g * \mathcal{D}h = \{\mathbf{g} * \mathbf{h}, n_{g*h}, L_{g*h}, \delta\}$ Then $\mathcal{D}g^{\times c} * \mathcal{D}h^{\times c} = \{\mathbf{g} * \mathbf{h}, n_{g*h}, cL_{g*h}, c\delta\}$.

3. The implication of the above properties for our purposes is that the vector component of a discretized standard Gaussian can be used as the vector component to discretize a Gaussian with arbitrary mean $\mu$ and variance $\sigma^2$

   i.e. if $\mathcal{D}N(0,1) = \{\mathbf{g}, n, L, \delta\}$, then $\{\mathbf{g}, n, \sigma L + \mu, \sigma\delta\}$ discretizes $\mathcal{D}N(\mu, \sigma^2)$.

   Thus the discrete vector for a standard Gaussian with gridwidth $\delta$ can be used in any discrete convolution where one convolvant is arbitrary Gaussian, provided that the other convolvant is similarly scaled, i.e its gridwidth is equal to the same $\sigma\delta$ as the Gaussian. Later in this chapter we explain the computational efficiencies this observation facilitates.

Given discretized functions, the computationally demanding aspect of a convolution is the vector operation (4.2), while computing the resultant location and support length is trivial. A literal implementation of a convolution from definition (4.2) requires $O(n^2)$ operations if the two vectors are of the same length. The same resulting vector $\mathbf{g} * \mathbf{h}$ can be computed much faster instead by way of a Fast Fourier Transform (FFT) in only $O(n \log n)$ steps.

### *4.2.1 Fast computation of discrete convolutions using Fourier transform*

Fast algorithms for computing discrete Fourier transforms, and applications to convolutions is a rich topic, and largely beyond the scope of this thesis. A comprehensive treatment may be found in Nussbaumer (1982). We summarize only a few salient concepts which motivate some of the design choices in our original software and which we describe later in this chapter.

A Discrete Fourier Transform (DFT) of a vector $\mathbf{g} \in \mathbb{C}^n$, indexed $0, \ldots, n-1$, is a vector, $\mathcal{F}\mathbf{g} \in \mathbb{C}^n$ defined as

$$(\mathcal{F}\mathbf{g})[k] = \sum_{j=0}^{n-1} \mathbf{g}[j] e^{-i\frac{2\pi kj}{n}}. \tag{4.3}$$

When the $\mathbf{g}$ is strictly real as in our case, this becomes

$$(\mathcal{F}\mathbf{g})[k] = \sum_{j=0}^{n-1} \mathbf{g}[j] \cos\Big(\frac{2\pi kj}{n}\Big) - i \sum_{j=0}^{n-1} \mathbf{g}[j] \sin\Big(\frac{2\pi kj}{n}\Big).$$

Following the conventions of signal processing, the space of $\mathbf{g}$ is called the *time domain* and the space of its transform $\mathcal{F}\mathbf{g}$ is called the *frequency domain*. The Inverse Discrete Fourier Transform (IDFT), $\mathcal{F}^{-1}$, of $\mathbf{f} \in \mathbb{C}^n$ is defined as:

$$(\mathcal{F}^{-1}\mathbf{f})[j] = \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{f}[k] e^{i\frac{2\pi kj}{n}}. \tag{4.4}$$

This is equivalent to $1/n$ times the DFT of the reverse of $\mathbf{f}$, so DFT and IDFT have the same computational complexity.

DFT has the property that it maps a discrete *circular* convolution in the time domain to pointwise multiplication in the frequency domain. A circular convolution (as distinct from

the *linear* convolution discussed in the previous section), of vectors $\mathbf{g}, \mathbf{h}$, of equal length $n$, written $\mathbf{g} \circ \mathbf{h}$, yields a vector of length $n$. It is defined as

$$(\mathbf{g} \circ \mathbf{h})[k] = \sum_{j=0}^{n-1} \mathbf{g}[k-j]\mathbf{h}[j], \quad k = 0, \ldots, n-1.$$

It follows from repeated application of (4.3) that

$$\mathcal{F}(\mathbf{g} \circ \mathbf{h})[k] = (\mathcal{F}\mathbf{g})[k] \cdot (\mathcal{F}\mathbf{h})[k].$$

Since $\mathcal{F}\mathbf{g}, \mathcal{F}\mathbf{h}$ are complex, the multiplication on the right hand side above is complex, i.e.

$$\begin{aligned} \operatorname{Re} \mathcal{F}(\mathbf{g} \circ \mathbf{h})[k] &= \operatorname{Re} \mathcal{F}\mathbf{g}[k] \operatorname{Re} \mathcal{F}\mathbf{h}[k] - \operatorname{Im} \mathcal{F}\mathbf{g}[k] \operatorname{Im} \mathcal{F}\mathbf{h}[k], \\ \operatorname{Im} \mathcal{F}(\mathbf{g} \circ \mathbf{h})[k] &= \operatorname{Re} \mathcal{F}\mathbf{g}[k] \operatorname{Im} \mathcal{F}\mathbf{h}[k] + \operatorname{Im} \mathcal{F}\mathbf{g}[k] \operatorname{Re} \mathcal{F}\mathbf{h}[k]. \end{aligned} \tag{4.5}$$

If we define the $\odot$ operation on equal length vectors $\mathbf{g}, \mathbf{h}$ as $(\mathbf{g} \odot \mathbf{h})[k] = \mathbf{g}[k] \cdot \mathbf{h}[k]$, then

$$\mathbf{g} \circ \mathbf{h} = \mathcal{F}^{-1}(\mathcal{F}\mathbf{g} \odot \mathcal{F}\mathbf{h}). \tag{4.6}$$

To use this result to compute a linear convolution of vectors $\mathbf{g}, \mathbf{h}$ with arbitrary lengths $n_g, n_h$, note that such a linear convolution can be formulated as a circular convolution of the two vectors when both are zero-padded up to length $n_g + n_h - 1$.

Accordingly, a convolution by way of DFT/IDFT requires three transformations. One DFT to transform each time domain input into the frequency domain, followed by one IDFT to transform the frequency domain product into the time domain.

Computing DFT and IDFT by literal implementation of (4.3) and (4.4) requires $O(n^2)$ arithmetic operations. So (4.6) by itself is not an inherently faster way to compute a discrete convolution than the literal implementation of (4.2). However, a variety of Fast Fourier Transform (FFT) algorithms have been discovered which compute a DFT in $O(n \log n)$. The most commonly used such algorithms, dating to Cooley and Tukey (1965), operate on the input vector recursively, computing the DFT of smaller subvectors and combining the results. This approach is most efficient when the length of the input vector is highly composite, as a multiple of small primes, to allow multiple levels of recursion.

When based on transformations computed in $O(n \log n)$, (4.6) offers a dramatic improvement in speed relative to a literal implementation of (4.2). For the vector lengths we consider in our calculations, the former algorithm could be hundreds or thousands of times faster than the latter in some cases.

We do not implement any FFT algorithms as such and rely on readily available libraries (The GNU Scientific Library (GSL), and Frigo and Johnson (2014)'s Fastest Fourier Transform in the West (FFTW3) library). We will not go into further detail on FFT algorithms here. However our original software is designed taking into account the basic principles described in this section.

## *4.3 Algorithm for First Hitting Time Likelihood and Truncated Gaussian Marginal Densities*

A few additional remarks before presenting the algorithm pseudo-code for the single-event version of the algorithm, and in light of the background on discrete convolutions using FFT.

1. Although computing a discrete convolution by way of FFT is considerably faster than the brute force approach, FFT is still a computationally costly algorithm. As noted, each convolution step requires 3 FFT operations – a forward FFT for each of the subdensity of interest, $g_{t-1}$ and the Gaussian innovation $\phi_t$, and then a reverse FFT to obtain the new subdensity $g_t$. A key driver of minimizing overall execution time is minimizing both the number of FFT operations and the length of the convolved vectors.

2. Representing a continuous distribution over the entire real line using a finite grid is necessarily imprecise. Certain operations tax precision, in particular, truncation of the subdistribution at the threshold and calcuation of derivatives at boundaries that lie between gridpoints. Care must be taken to approximate these values within acceptable tolerance.

3. With each successive convolution the untruncated support of the subdensity above the threshold grows wider, with an increasing number of gridpoints. Keeping the grid to a manageable size requires us to trim the most extreme gridpoints where the density is negligible.

4. All of these issues are that much more salient for 2-dimensional convolutions than for the 1-dimensional case, as the number of gridpoints needed for acceptable precision is considerably larger in the former than in the latter (more than an order of magnitude in our implementation).

5. As noted in 4.2 (3), a convolution of function $g(\cdot)$ with an arbitrary Gaussian is equivalent to convolution with a standard Gaussian, up to appropriate scaling and translation, provided that $\mathcal{D}g$ is correspondingly scaled. Exploiting this property, we standardize every problem so that the only Gaussian convolvant ever required is a standard Gaussian. By doing so, we only need to perform a DFT of the discretized standard normal density once at the beginning of a program. We can then reuse the frequency domain vector in every subsequent application of (4.6). This reduces the number of FFT operations per convolution from 3 to 2.

   In order to achieve this, to calculate the likelihood of an outcome under parameters $(W_0, \boldsymbol{\mu}, \sigma)$, we rescale the parameters to $(W_0/\sigma, \boldsymbol{\mu}/\sigma, 1)$. This transformation is only for the purpose of calculating the likelihood of a process for a single set of values, not for estimating the unknown parameters of the problem as a whole. Thus instead of convolving at the $t$ step with with a one-off vector for $\mathcal{D}N(\mu_t, \sigma^2)$, we convolve with the $\mathcal{D}N(0, 1)$ and translate the lower bound by $\mu_t/\sigma$.

   Henceforth we refer to the standard normal density used in every convolution operation as the convolution *kernel*.

6. As noted in Section 4.2.1, convolution by FFT requires that the input vectors be

of identical length and zero-padded to the length of the linear convolution. As the length of $\mathbf{g}_t$ is increased with each successive convolution, we need to produce multiple convolution kernels, zero-padded to match the length of $\mathbf{g}_t$. On the other hand, $\mathbf{g}_t$ is also reduced in some steps due to truncation at the threshold, perhaps more than it is increased by convolution. Given the imperative of minimizing vector length in order to minimize execution time, the convolution kernel used in one iteration may be zero-padded to a shorter length than the one used in the prior iteration.

Furthermore, the FFT algorithm is fastest when the length of the input vectors is a product of small primes, ideally a power of 2. When the length is a large prime number, execution time can be relatively slow. A distribution with support length, say 119, can be efficiently transformed by zero-padding it to 128. A distribution of length 129 would possibly be transformed more efficiently by padding it to 256 than by simply transforming it as a vector of length 129. However, if padded to length $192 = 3 \times 2^6$ it might be transformed more quickly than if padded to 256.

The implication here is that an important task for minimizing execution time is maintaining a cache of transformed standard normal convolution kernels from vectors zero-padded to varying lengths, where the lengths are products of small primes. The transformation need be done only the first time a given length is called for, and the frequency domain vector saved for reuse for the remainder of the computation process.

7. In many instances convolution operations that are required in theory may be avoided altogether in practice. The *Shortcut Theorem* given in Appendix B presents conditions where the outcome probabilities can be satisfactorily approximated by easily computed probabilities of a univariate Gaussian, in particular, certain outcomes that are either highly improbable ($\psi(Y, \Delta) < \epsilon$) or nearly certain ($\psi(Y, \Delta) > 1 - \epsilon$) for acceptably small $\epsilon$. In these cases the likelihood algorithm can return either a minimal value or 1, respectively. In less extreme cases, it may possible to bypass some number of initial convolutions. These conditions can be tested quickly with a standard normal c.d.f.,

possibly obviating a number of time-consuming convolutions.

Specifically, if we wish to calculate an outcome probability for our standardized process, which stops upon $W_t < 0$, where:

$$W_t = W_{t-1} + w_t \quad w_t \sim N(\mu_t, 1) \quad \forall t > 0.$$

We also define a dual process, $X$, governed by the same $W_0, \boldsymbol{\mu}$ and generating mechanism, but which continues indefinitely without stopping. Then

$$X_t \sim N(M_t, t), \qquad \text{where } M_t = W_0 + \sum_{i=1}^{t} \mu_t,$$
$$P(X_t < 0) = \boldsymbol{\Phi}(-M_t/\sqrt{t}),$$
$$P(X_t \geq 0) = 1 - \boldsymbol{\Phi}(-M_t/\sqrt{t}),$$

where $\boldsymbol{\Phi}$ is the standard normal c.d.f. The Shortcut Theorem states:

(a) If $P(X_Y < 0) < \epsilon$ then $\psi(1, Y) < \epsilon$.

(b) If $P(X_t \geq 0) < \epsilon$ for some $t < Y$, then $\psi(0, Y) < \epsilon$ and $\psi(1, Y) < \epsilon$.

(c) If $P(X_t < 0) < \epsilon/Y$ for every $t \leq Y$ then $\psi(0, Y) > 1 - \epsilon$.

Furthermore, as a corollary, let $t_0$ be the maximum time $0 \leq t_0 < Y$ such that for all $t \leq t_0$, $P(X_t < 0) < \epsilon/Y$. Then we can avoid all convolutions before step $t_0 + 2$ and still approximate the outcome probability within $\epsilon$.

The details and proof of the Theorem and this corollary are given in Appendix B.

The following algorithm to calculate outcome likelihoods fills in some of the key non-trivial implementation details for the high-level description in Section 4.1. When a subroutine is indicated in the pseudo-code as "Call `[subroutine name]()`", additional detail on the named subroutine is provided below. Following the notation from Section 4.2, Let $\mathcal{D}\phi = \{\boldsymbol{\phi}, n_\phi, L_\phi, U_\phi, \delta\}$ be the standard normal convolution kernel, and $\{\mathbf{g}_t, n_t, L_t, U_t, \delta\}$

represent the discrete approximation for $g_t$. $n_t, n_\phi$ represent the actual support, exclusive of any zero-padding.

Subroutine `Likelihood(`$Y, \Delta, W_0, \boldsymbol{\mu}$`)` returns $P(Y, \Delta)$ and vector of threshold densities $\mathbf{d}[]$.

$W_0, \boldsymbol{\mu}$ are pre-standarized to unit variance.

1. Apply the tests from the "Shortcut Theorem" described above to determine whether any convolutions can be avoided; e.g. is the outcome probability "nearly zero" or "nearly one"? If so, return accordingly, or start the loop with $t = t_0 > 1$ as appropriate.

2. Call `Initialize()` to create a standard normal convolution kernel, or reuse a previously created one.

3. Loop over $t = 1..Y$

   (a) If $t = 1$ then set $g_1$ to be the kernel, but centered at $W_0$, i.e.

   $$\mathbf{g}_1 \leftarrow \boldsymbol{\phi},$$
   $$L_1 \leftarrow L_\phi + W_0.$$

   (b) If $t > 1$ then Call `Convolve()` to set $\mathcal{D}g_t \leftarrow \mathcal{D}g_{t-1} * \mathcal{D}\phi$.

   (c) Translate $\mathcal{D}g_t$ by setting $L_t \leftarrow L_t + \mu_t, \quad U_t \leftarrow U_t + \mu_t$.

   (d) If $L_t > 0$, then essentially all of the mass is above the threshold, so we can skip to the next the step.

   If $U_t < 0$, then essentially all of the mass is below the threshold, so we know that failure probability at this $t$ equals total mass, survival probability is essentially 0, and probability of any event occurring at any subsequent $t'$, $t < t' \leq Y$ is essentially 0. So we can break out of the iteration loop and return from the function.

Otherwise $L_t \leq 0 \leq U_t$, in which case there is some gridpoint, the domain of which includes the threshold 0. Call `SetThreshold()` to determine this gridpoint and the fractions of its mass which lie above and below the threshold, call them respectively $\mathbf{g}^*_+, \mathbf{g}^*_-$.

(e) Call `ComputeThresholdDensity()` to set $\mathbf{d}[t] \leftarrow g_t(0)$.

(f) If $t = Y$ then sum the mass of the gridpoints strictly above the threshold for survival probability; sum the mass of the gridpoints strictly below the threshold for failure probability. Add the appropriate fractional mass of the gridpoint containing the threshold, $\mathbf{g}^*_+$ or $\mathbf{g}^*_-$ depending on whether we want likelihood of survival or failure. Return this result as $P(Y, \Delta)$, along with the vector of threshold densities $\mathbf{d}[t]$.

(g) If $t < Y$ and $L_t < 0$, then Call `Truncate()` to truncate $\mathbf{g}_t$, leaving only the probability mass above the threshold. Set $L_t \leftarrow 0$.

(h) Call `Trim()` to consolidate tail elements of $\mathbf{g}_t$ with negligible mass, resulting in a vector with equal total mass but shorter support.

(i) Select from the cache of kernels, or create a new kernel if needed, the zero-padded kernel of shortest length such that the total vector length is at least $n_\phi + n_t - 1$. Then zero-pad $\mathbf{g}_t$ as needed to match the length of the zero-padded kernel.

To compute the vector of univariate marginal densities $\mathbf{f}[]$ as defined in Chapter 2:

Subroutine `MarginalDensity`$(Y, \Delta, \mathbf{d}[], P(Y, \Delta))$, returns $\mathbf{f}[]$. $P(Y, \Delta), \mathbf{d}[]$ are the values returned from `Likelihood()`.

1. If $P(Y, \Delta) \approx 0$ or $P(Y, \Delta) \approx 1$, then $\mathbf{f}[] \leftarrow \mathbf{0}$; return.

2. Loop over $t = 1..Y - 1$

   (a) Set $\mathbf{f}[t] \leftarrow$ `Likelihood`$(Y - t, \Delta, W_0 = 0, \boldsymbol{\mu}_{t+1..Y}) \times \mathbf{d}[t]/P(Y, \Delta)$.

3. $\mathbf{f}[Y] \leftarrow \mathbf{d}[Y]/P(Y, \Delta)$; return $\mathbf{f}[]$.

Further detail on subroutines invoked above:

- `Initialize()`: In the easy case, all time intervals between observations of the given subject are equal length, in which case a single kernel (up to zero-padding) can be used at every iteration.

  The inputs are a grid length $n$, and tolerance parameter $\epsilon$. (Our defaults are $n = 512$, $\epsilon = 1 \times 10^{-9}$). Set the lower bound $L_\phi \leftarrow \Phi^{-1}(\epsilon/2)$, where $\Phi$ is the standard normal c.d.f. Then upper bound $U_\phi \leftarrow -L_\phi$, gridwidth $\delta = (U_\phi - L_\phi)/n$. Assign values to $\boldsymbol{\phi}$, $\boldsymbol{\phi}[k] = \Phi(L_\phi + (k+1)\delta) - \Phi(L_\phi + k\delta)$, etc. per (4.1). Compute $\mathcal{F}\boldsymbol{\phi}$.

  If there multiple observation intervals for the subject, then each distinct observation length $\ell$ will require a distinct kernel, appropriately scaled by $\sqrt{\ell}$, but calibrated so that all kernels share a common $\delta$.

- `Convolve()`:

  To obtain $\mathbf{g}_t \leftarrow \mathbf{g}_{t-1} * \boldsymbol{\phi}$, compute $\mathcal{F}\mathbf{g}_{t-1}$, then $\mathbf{g}_t \leftarrow \mathcal{F}^{-1}(\mathcal{F}\mathbf{g}_{t-1} \odot \mathcal{F}\boldsymbol{\phi})$.

  $n_t \leftarrow n_{t-1} + n_\phi - 1$,
  $L_t \leftarrow L_{t-1} + L_\phi + \delta/2$,
  $U_t \leftarrow U_{t-1} + U_\phi - \delta/2$.

- `SetThreshold()`:

  This covers the case where the threshold lies on some gridpoint, i.e there is some $k_0$, with $0 \le k_0 \le n-1$ where $L_t + k_0 \cdot \delta \le 0 < L_t + (k_0+1) \cdot \delta$. Use 3-point polynomial interpolation to approximate the fraction of the probability mass above the threshold; i.e. we assume that the subdensity is approximated near 0 by a quadratic polynomial $h_0(x) = a_0 + a_1 x + a_2 x^2$, and therefore the nearby gridpoints are approximated by

$H_0 = \int h_0(x)dx = a_0x + \frac{1}{2}a_1x^2 + \frac{1}{3}a_2x^3 + C$ over corresponding intervals. Then solve for $a_0, a_1, a_2$ by using 3 gridpoints in the neighborhood of 0, i.e.

$$\begin{aligned}
\mathbf{g}_t[k_0] &= H_0(L_t + (k_0+1)\delta) - H_0(L_t + k_0\delta) \\
\mathbf{g}_t[k_0-1] &= H_0(L_t + (k_0)\delta) - H_0(L_t + (k_0-1)\delta) \\
\mathbf{g}_t[k_0+1] &= H_0(L_t + (k_0+2)\delta) - H_0(L_t + (k_0+1)\delta)
\end{aligned}$$

(We use instead $\mathbf{g}_t[k_0], \mathbf{g}_t[k_0+1], \mathbf{g}_t[k_0+2]$ if $k_0 = 0$, and use $\mathbf{g}_t[k_0], \mathbf{g}_t[k_0-1], \mathbf{g}_t[k_0-2]$ if $k_0 = n_t - 1$). Then the fractional mass of $\mathbf{g}_t[k_0]$ above the threshold is
$\mathbf{g}^*_+ = H_0(L_t + (k_0+1)\cdot\delta)$ and $\mathbf{g}^*_- = \mathbf{g}_t[k_0] - \mathbf{g}^*_+$.

- `ComputeThresholdDensity()`:

  If $L_t > 0$ or $U_t < 0$ then $g_t(0) \approx 0$ so $\mathbf{d}[t] \leftarrow 0$. Otherwise the threshold is on a gridpoint. Using the same polynomial interpolation as in `SetThreshold()`, the marginal subdensity at the threshold is $\mathbf{d}[t] \leftarrow h_0(0) = a_0$.

- `Truncate()`:

  Copy into a new vector $\mathbf{g}^*_t$ only that portion of $g_t$ which lies above the threshold. i.e $\mathbf{g}^*_t[0] \leftarrow \mathbf{g}^*_+$ as computed in `SetThreshold()`, then for $k = k_0+1, \ldots, n_t - 1$, $\mathbf{g}^*_t[k-k_0] \leftarrow \mathbf{g}_t[k]$.

  Ł$_t \leftarrow L_t + k_0\cdot\delta$,
  $n_t \leftarrow n_t - k_0 + 1$.

- `Trim()`:

  Adjust the tails of the grid to shorten the gridlength if a tail has negligible mass. i.e. for small $\epsilon$ find the least $M$ such that $\tau = \sum_{j>M} \mathbf{g}^*_t[j] < \epsilon$. Then $\mathbf{g}^*_t[M] \leftarrow \mathbf{g}^*_t[M] + \tau$. Similarly, find the largest $m$ such that $\tau = \sum_{j<m} \mathbf{g}^*_t[j] < \epsilon$. Set $\mathbf{g}^*_t[m] \leftarrow \mathbf{g}^*_t[m] + \tau$.

Finally, $n_t \leftarrow M - m + 1$,
$\mathbf{g}_t[0] \leftarrow \mathbf{g}_t^*[m], ..., \mathbf{g}_t[n_t - 1] \leftarrow \mathbf{g}_t^*[M]$,
$L_t \leftarrow L_t + m \cdot \delta$, $U_t \leftarrow L_t + n_t \cdot \delta$.

The algorithm is implemented in C, using the fast Fourier transform functions from the GNU Scientific Library of Galassi et al. (2009).

### *4.3.1 Bivariate*

The algorithm for computing the outcome likelihood and marginal densities in the bivariate case is structurally similar to the univariate case. In addition to depending on 2-dimensional grids, convolutions and Fourier transforms, the key differences are:

- While we can still standardize every problem and convolution kernel as unit variance, the shape of the distribution depends on the correlation $\rho$. A new kernel needs to be constructed for every value of $\rho$.

  Our algorithm to construct the kernel grid is based on Genz (2013)'s TVPACK Fortran routines, described in Genz (2004). His function `BVND(`$h, k, \rho$`)` computes, for standard bivariate normal with correlation $\rho$, $P(x_1 > h, x_2 > k)$

  Similar to the univariate case, we define a grid of dimensions $n \times n$ on $[L_\phi, -L_\phi] \times [L_\phi, -L_\phi]$ where $L_\phi = \Phi^{-1}(\epsilon)$. By default we use $n = 96, \epsilon = 1 \times 10^{-5}$.

- The `SetThreshold()` and `Truncate()` subroutines are similar to the univariate case, except that the fractional mass has to be calculated at every gridpoint crossed by the bidimensional threshold. i.e. if the threshold crosses the grid in the middle of the $k$th row point and the $j$th column point, then the fractional mass is computed for every boundary element $(k, j), (k, j + 1), \ldots (k, n)$ and $(k + 1, j), \ldots (n, j)$.

- The biggest difference is in computing the marginal densities, which are done in both dimensions.

Assume that the rows of the grid matrix correspond to the $a$ axis and the columns to the $b$ axis. To compute the univariate marginal density that process $a$ hits the threshold at time $t$, suppose the thresholds cross the intervals at row $k$ and column $j$. Then applying the same polynomial interpolation approach as above, find the threshold density along the $a$ axis at every gridpoint $(k, j), (k, j+1), \ldots, (k, n)$. Call the vector where these densities are stored the "density stripe".

The sum of the elements in the "density stripe" vector is the marginal density of hitting the $a$ threshold at time $t$ conditioned only on information through time $t$. In order to get the corresponding density conditioned on the observed outcome at time $Y$, we next need to compute the probability of the observed outcome at $Y$ conditioned on starting with this "density stripe" at time $t$. In the univariate case we would invoke the "Likelihood calculation routine" with $W_0 = 0$ and observation time $Y - t$. In the bivariate case the initial data grid is set to all "0"s but for the "density stripe".

We also need the bivariate marginal densities, i.e. for every time $s$, $t < s \leq Y$, the probability that the process *also* hits the $b$ threshold at time $s$ after hitting the $a$ threshold at $t$ and on the way to reaching the observed outcome state at $Y$. Therefore during the likelihood calculation routine which started at $t$ with the last density stripe, for each such $s$, compute a "density stripe" for hitting the $b$ threshold at $s$. Then on yet another loop, compute the likelihood of the observed outcome conditioned on starting with the "density stripe" at $s$.

- The 2-dimensional convolutions are based on FFT routines from Frigo and Johnson (2014)'s FFTW3 software package, described by Frigo and Johnson (2005).

### *4.4 Computation Times*

In this section we illustrate the execution times of certain key operations. The costs of these operations motivate much of the work to devise practical estimation approaches.

To measure execution times we use 537 subjects from the mortgage data set which we describe in detail in Chapter 5. All of the tested subjects terminate with prepayment and with final observation times ranging from 2 to 118 months. For each subject we measure the elapsed time to compute the following operations: both likelihood and gradient in the 1-dimensional case; and in the 2-dimensional case the likelihood, partial gradient (with just univariate marginal densities) and full gradient, with both univariate and bivariate marginal densities. For each scenario we plot execution time (in seconds) vs. final observation time $Y$ in time periods. (In all cases we use the same parameter to determine a subject's time-varying drift vector as a function of its covariates).

The measurements were performed in the fastest environment available to us, Amazon EC2 c3-large (dual core equivalent) instance running 64-bit Linux, and with no other user jobs running concurrently.

The plots are in Figures 4.1 and 4.2. To summarize the worst case times ($Y = 118$), the 1-dimensional likelihood is 0.034 sec., the 1-dimensional gradient is 1.36 seconds. The 2-dimensional likelihood is 2.15 seconds. The 2-dimensional partial gradient is 304.5 seconds and the full gradient (for $Y = 91$) is 3,003 seconds.

## *4.5 Optimization Algorithms for Finding MLEs*

### *4.5.1 Overview of optimization methods*

We tested a number of different optimization strategies, bearing in mind that computing gradients is significantly more computationally expensive than the objective (log-likelihood) function itself. The cost is more pronounced for bivariate models than for univariate models, and those differences increase polynomially with observation time.

The main optimization strategies we considered are:

1. Powell (2009)'s BOBYQA derivative-free optimization method has the advantage that it does not require any gradient computations. However in our simulations it converged quite slowly relative to other approaches, and with weak starting values it sometimes

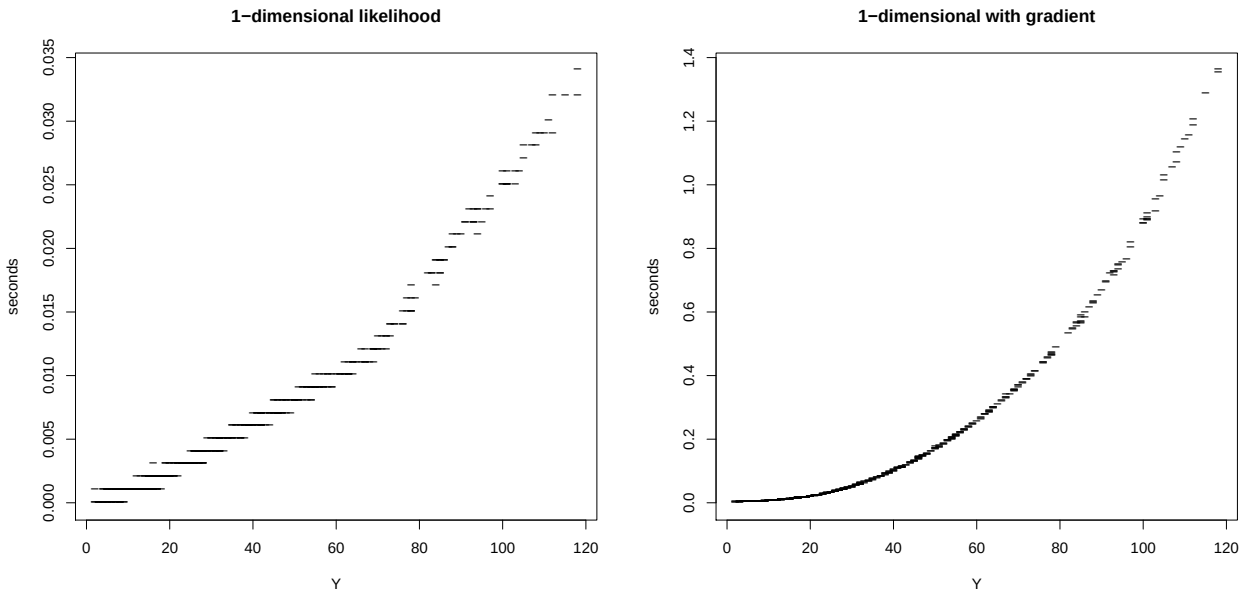


Figure 4.1: Computation times for 1-dimensional likelihood and gradient operations

quit after a few iterations, barely budging from the starting position. This behavior may be attributable to the fact that our likelihood calculation algorithm may return a single low value to indicate a highly implausible likelihood below a tolerance level. Thus BOBYQA can not necessarily discern between different states with extremely low likelihood. Even with an informed starting value as described below, BOBYQA was sufficiently slow relative to other methods that we discontinued experimenting with it.

2. In the single-event case and starting with a naïve initial value we had the best results with the EM algorithm described in Chapter 2, accelerated by Varadhan (2012)'s `SQUAREM` package. However, the the estimates were comparable to those with a gradient descent algorithm, and the latter seemed to converge faster when started with an informed initial value.

3. Our most successful experiments in the dual-event case were using Byrd et al. (1995)'s L-BFGS-B quasi-Newton algorithm with box constraints, as implemented in the R

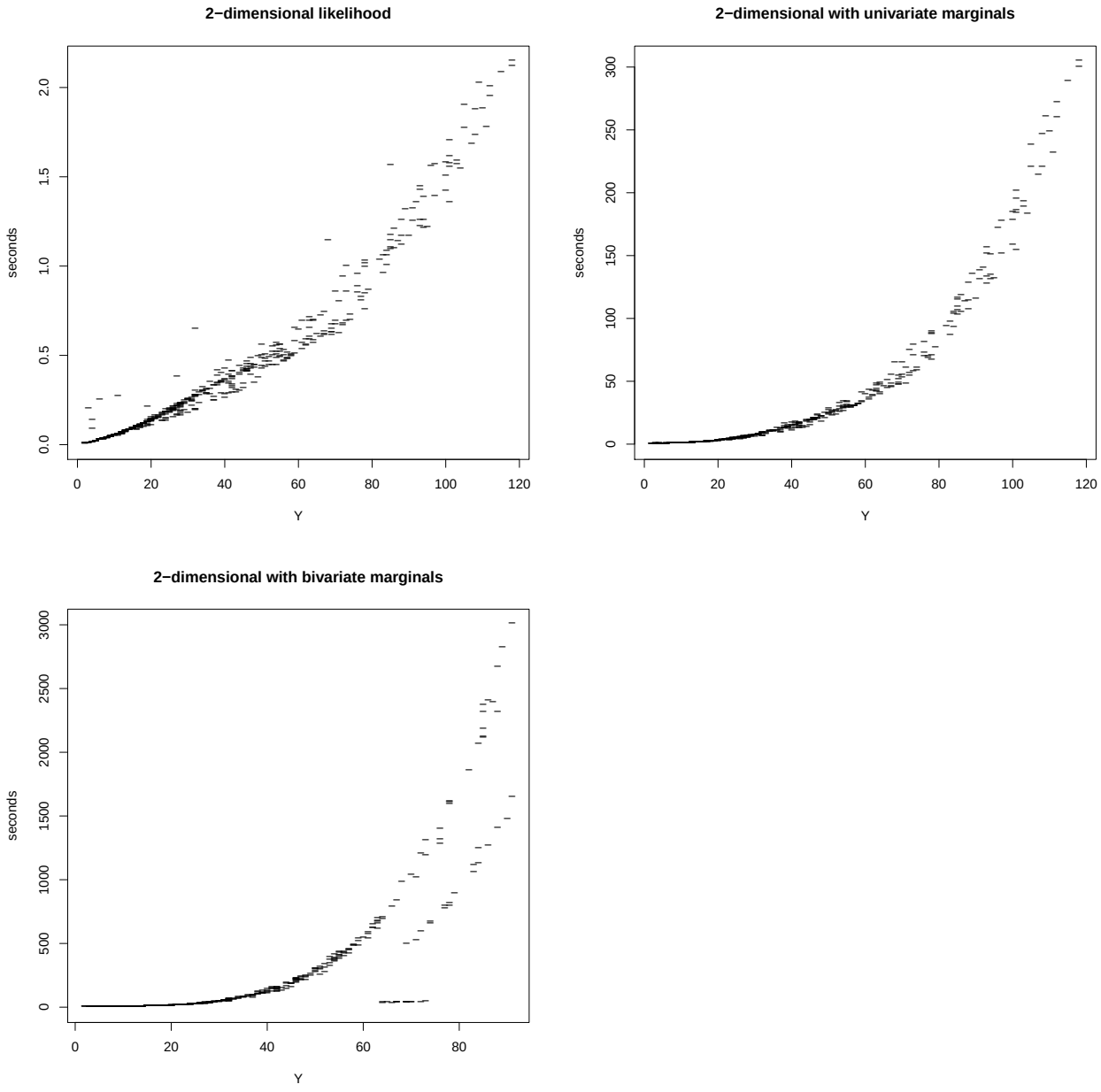


Figure 4.2: Computation times for 2-dimensional likelihood and gradient operations

`optim()` function, with our algorithm providing the explicit gradient.

We also implemented the dual-event EM algorithm given in Chapter 3, but it was relatively slow even accelerated by `SQUAREM`.

We also explore the following techniques to reduce computation time that may work with various optimization algorithms. (1) and (3) apply only to the dual-event case, and (2), (4) and (5) apply to both single and dual event models:

1. Numerically differentiate $\rho$ instead of computing its gradient component analytically.

2. Find an informative starting value.

3. Estimate the $\boldsymbol{\beta}, \sigma$ parameters for each event process as if independent single-event processes, then estimate $\rho$ conditioned on $\boldsymbol{\beta}$ and $\sigma$.

4. "Compress" time, ie. combine observations from multiple consecutive time intervals into a single summary observation.

5. Exploit inherent parallelism.

Each of these techniques is discussed in more detail in a separate subsection below.

### *4.5.2 Numerical differentiation of $\rho$*

As illustrated in Section 4.4, the most time-consuming operation by far is calculation of the full bivariate gradient, including the matrix of bivariate marginal densities, $\mathbf{F}$. The latter requires $O(Y^4)$ convolutions, but is needed only for the $\rho$ component of the gradient. The gradient components for all other parameters require only the univariate marginal densities, $\mathbf{f}$, computable with $O(Y^2)$ convolutions. We obviate the need for computing $\mathbf{F}$ by numerically estimating $\frac{\partial L}{\partial \rho}$ using 3-point quadratic approximation. This requires, for each subject, only two additional likelihood computations of $O(Y)$ convolutions.

In our trials we found that the time savings was considerable, For example, the time to compute a full-sample gradient for our 1,000 subject mortgage data set ($\max Y = 118$, but compressed by a factor of 6), was roughly 2,000 seconds when computing $\mathbf{F}$ for each subject, but dropped to less than 800 seconds to when approximating the $\rho$ derivative, with the discrepancy in values on the order of $1 \times 10^{-5}$.

Numerical differentiation for all other gradient components would also be asymptotically $O(Y)$ in the number of convolutions, but total computation time in practice also depends on the length of $\boldsymbol{\beta}$, and on characteristics of the data. For many of the subjects tested it is faster to compute the univariate marginal vector $\mathbf{f}$ than to numerically differentiate the other parameters. We suggest, but do not implement, an adaptive algorithm that could judiciously compute some gradients analytically and others numerically.

### *4.5.3 Heuristic method for finding an informative starting value*

Speed of convergence for an optimization method can depend widely on the choice of starting value.

We found that the following approach provided a reasonably informative starting value in our experiments for the bivariate case, and which dramatically reduced convergence time relative to starting with an uninformative value. In our trials this heuristic required on the scale of 1/5 of the execution time of a single iteration of computing the gradient of the entire sample, and obviated the need for perhaps dozens of iterations from an uninformative starting point.

1. First, use a Minorize-maximization (MM) algorithm to find an initial value for $\boldsymbol{\beta}$. Specifically, find the value of $\boldsymbol{\beta}$ which maximizes the likelihood across the modal values of the complete data for each subject. This does not necessarily find the MLE for $\boldsymbol{\beta}$, but in all trials to date it has been a useful approximation.

   Starting with a naïve $\boldsymbol{\beta}_0 = \mathbf{0}$, iterate over $n$ until either $\|\boldsymbol{\beta}_{n+1} - \boldsymbol{\beta}_n\| < \epsilon$ or there is only a negligible increase in log likelihood.

(a) Find the modal value of complete data for each subject $i$, by solving for each subject the quadratic program

$$\mathbf{z}_{i,n} = \underset{\mathbf{z}}{\text{argmax}} \ -\frac{1}{2}(\mathbf{z} - \boldsymbol{\nu}_i(\boldsymbol{\beta}_n))'\boldsymbol{\Omega}_i^{-1}(\mathbf{z} - \boldsymbol{\nu}_i(\boldsymbol{\beta}_n)),$$

subject to $\mathbf{z} \geq 0$, where $\boldsymbol{\nu}_i(\boldsymbol{\beta}) = \boldsymbol{\Sigma}_i\mathbf{X}_i\boldsymbol{\beta} + \mathbf{b}_i$, with $\boldsymbol{\Sigma}_i, \mathbf{b}_i$ depending on outcomes $Y_i, \Delta_i$ as described in Chapter 3, and where $\boldsymbol{\Omega}_i$ is a function of $Y_i, \Delta_i$ with the simplifying assumption of $\sigma_a = \sigma_b = 1, \rho = 0$.

(b) Find $\boldsymbol{\beta}_{n+1}$ which maximizes the log likelihood of the set $\{\mathbf{z}_{i,n}; i = 1, \ldots, N\}$. i.e.

$$\boldsymbol{\beta}_{n+1} = \underset{\boldsymbol{\beta}}{\text{argmax}}\ Q(\boldsymbol{\beta}) = \sum_{i=1}^{N} -\frac{1}{2}(\mathbf{z}_{i,n} - \boldsymbol{\nu}_i(\boldsymbol{\beta}))'\boldsymbol{\Omega}_i^{-1}(\mathbf{z}_{i,n} - \boldsymbol{\nu}_i(\boldsymbol{\beta})),$$

$$\boldsymbol{\beta}_{n+1} = \Big(\sum_{i=1}^{N} \mathbf{X}'\mathbf{X}\Big)^{-1} \sum_{i=1}^{N} \mathbf{X}'\boldsymbol{\Sigma}^{-1}(\mathbf{z}_{i,n} - \mathbf{b}_i).$$

2. Next, given the initial value for $\boldsymbol{\beta}$ found above, use numerical maximum likelihood estimation for $\sigma_a, \sigma_b, \rho$ under the relaxed assumption that each subject's process is an unconstrained multivariate normal, i.e. where the threshold may be crossed before the observation time.

   Then the value of the unconstrained process at observation time $Y_i$, $\mathbf{W}_{i,Y_i}$, is a bivariate normal, $\mathbf{W}_{i,Y_i} \sim N(\mathbf{W}_0 + \sum_{t=1}^{Y_i} \boldsymbol{\mu}_{i,t}, \sum_{t=1}^{Y_i} \ell_{i,t}\boldsymbol{\Psi})$. The $\boldsymbol{\mu}_i$ are defined by the $\boldsymbol{\beta}$ from the previous step. The likelihood of the observed outcome is then the probability that $\mathbf{W}_{i,Y_i}$ lies in the quarter-plane defined by the $x$ and $y$ axes corresponding to the outcome $\Delta_i$.

   It is a straightforward optimization problem to find the $\sigma_a, \sigma_b, \rho$ which maximize the log likelihood. We solve it using R `optim()`'s L-BFGS-B, and where the quarter-plane probabilities are given by the `pmvnorm()` function.

   Again, this approach does not pretend to find the MLE of the constrained problem, but in our experience does find a sufficiently informative starting value to dramatically reduce time of convergence to the desired estimates.

### *4.5.4 Independent Estimation*

We attempt to estimate the parameters of each event process as a single-event process, as if the two events are independent. First estimate the parameters $\boldsymbol{\beta}_a, \sigma_a$, where outcome $a$ is the event of interest and where an observation of event $b$ is handled as independent random censoring, and *vice versa*. Then $\rho$ is estimated over the full dual-event sample, conditional on the other parameters.

If this approach provides sufficiently unbiased and efficient estimates, then it could save considerable time relative to joint estimation of correlated processes, depending on the data characteristics. However, consistent estimates are guaranteed only if the event processes are independent.

As we discuss in more detail in Chapter 5, in certain simulation scenarios, this approach offered reasonable estimates and standard errors. However for the mortgage data it produced estimates that were quite different in some cases from the joint estimates.

### *4.5.5 Time compression*

An additional approach that we use for reducing convergence time is to administratively compress observations from multiple consecutive time periods into a single observation. If the trade-offs of losing some information, and possibly introducing some bias are acceptable in a given situation, time compression can save considerable computation time where subjects have longer observational series. As illustrated in Section 4.4, our mortgage data includes subjects with up to 118 monthly observations. Even when differentiating $\rho$ numerically, it took over 300 seconds to compute the gradient for a single subject of that observation length.

As shown in Chapter 3, non-informative observation times can vary both for a given subject and across subjects. For practical implementation reasons we choose a uniform *compression factor* which we apply to every subject, and which denotes the integer number of consecutive actual observations which are combined into a single aggregate observation for analysis. Since the parameters are estimated for their effects per unit time, the compression

factor need not necessarily correspond to any intuitive time period in the application domain.

For the purposes of this section, we assume that the actual observations are at unit intervals. The basic principles extend to data with arbitrary observation times.

Along with shrinking the final observation time, we also aggregate covariate values for each compressed time period. The elements of $\boldsymbol{\beta}$ represent the expected *change* in a subject's distance from the threshold over a unit time interval that is associated with a unit of the covariate. Therefore the value of a subject's covariate for a compressed period is the sum of that covariate's values for each unit observation within the period.

Say a subject has final observation time $Y$ and the compression factor is $m$. We produce a new observation time $Y'$, covariate matrix $\mathbf{X}'$ and interval length vector $\boldsymbol{\ell}$ as follows:

1. Let $y = \lfloor Y/m \rfloor, \quad r = Y \mod m$, i.e. $Y = m \cdot y + r, \quad 0 \leq r < m$.

2. To represent the initial $y$ aggregated intervals of length $m$ unit intervals: for $s = 1, \ldots, y$, the covariate vector $\mathbf{x}'_s$ for the aggregated interval $s$ is the sum of the covariate vectors for the actual intervals to be aggregated:

$$\begin{aligned} \ell_s &= m, \\ \mathbf{x}'_s &= \sum_{t=m\cdot(s-1)+1}^{m\cdot s} \mathbf{x}_t. \end{aligned} \tag{4.7}$$

   If $r = 0$ then $Y' = y$ and we're done. Otherwise, to ensure that the compressed observation times are entirely non-informative, the final aggregated interval depends on the subject's outcome and how the data is observed, as follows:

   (a) If the subject's final observation period coincides with a non-informative administrative observation, such as the end of the research study:

$$\begin{aligned} Y' &= y + 1, \\ \ell_{Y'} &= r, \\ \mathbf{x}'_{Y'} &= \sum_{t=m\cdot y+1}^{Y} \mathbf{x}_t. \end{aligned}$$

(b) If the subject is censored due to random loss to follow up before the end of the research study:

The only information we have on the subject is that it had survived up to $m \cdot y$, therefore $Y' = y$, and $\boldsymbol{\ell}, \mathbf{X}'$ are as defined in (4.7).

(c) If the subject is shown to have experienced an event after $m \cdot y$ and the observation at $Y$ does not coincide with the end of the research study:

Since we are analyzing the subject as if it is observed only at non-stochastic times, i.e. every $m$ time units or at the end of the research study, we make the following adjustment. Let $R$ be the number of time intervals from the subject's actual $m \cdot y$ observation until the end of the research study. Let $r' = m \wedge R$. i.e. the number of actual intervals to the next non-stochastic observation. Then

$$\begin{aligned} Y' &= y + 1, \\ \ell_{Y'} &= r', \\ \mathbf{x}'_{Y'} &= \sum_{t=m\cdot y+1}^{m\cdot y+r'} \mathbf{x}_t. \end{aligned}$$

(This embeds the assumption that the time dependent covariates are external to the subject and may be observed up to the end of the research study, even after the subject experiences a terminal event).

Computation of the likelihoods and marginal densities follows exactly as per the analytic formulas in the previous chapters and the algorithms in this chapter above. The only modification to the algorithm is in the construction of convolution kernels. Each kernel corresponding to an innovation of different length $\ell$ time units must correspond to $N(0, \ell)$. Since discrete convolutions assume that the two vectors represent identically spaced grids, every kernel applied to a given subject must assume the same $\delta$. The vectors themselves, however, will differ.

### *4.5.6 Exploit inherent parallelism*

Although we did not implement this approach, we observe that the most time-consuming aspects of parameter estimation lend themselves to Single Program, Multiple Data (SPMD) parallelization in a straightforward way. In each iteration of a likelihood maximization algorithm, given a candidate value of $\theta$, the likelihood and/or gradient for each subject are calculated independently, and then the results are summed and returned to the maximizer. On a single processor the values for each subject are calculated in sequence. The data could be shared among arbitrarily many processors, even separate workstations, each of which calculates the likelihood and gradients for a designated subset of the subjects each iteration, then combining the partial results to return the aggregate result to the maximizer. Various R packages provide the tools to implement this approach, including Chen et al. (2014)'s `pbdMPI` and Tierney et al. (2013)'s `snow`.

# Chapter 5

# NUMERICAL RESULTS

> "At this juncture, however, the impact on the broader economy and financial markets of the problems in the subprime market seems likely to be contained. In particular, mortgages to prime borrowers and fixed-rate mortgages to all classes of borrowers continue to perform well, with low rates of delinquency.", Federal Reserve Chairman Ben Bernanke, Testimony to U.S. Congress Joint Economic Committee, March 28, 2007.

## *5.1 Simulations*

### *5.1.1 Simulations for a single terminating event*

We apply our estimation methods to simulated data using our implementation of the expectation-maximization algorithm with some code in R calling our C code for computing event time likelihoods and univariate marginals. We also use Varadhan (2012)'s `SQUAREM` R package to accelerate convergence time.

*Data Generation*

Each simulated data set is generated as follows. For each of $n$ subjects, a single baseline covariate $x_B$ is drawn from $N(\mu = 12, \sigma^2 = 4)$. Each of the $n$ subjects also has a single time-varying covariate, $x_V$, for $t = 1, \ldots, 20$. This sequence is drawn from a generated AR(1) series, with $\rho = 0.9$, intercept $a = 1$ and standard normal innovations. The subsequence which we use for the covariate is drawn starting at a random location in the generated series. In some simulation scenarios all $n$ subjects share the same values of $x_V$, i.e. this simulates an exogenous macroeconomic effect (such scenarios are denoted in the results tables

as "common"). In other scenarios each subject has distinct values for $x_V$. They are drawn from a common realization of an AR(1), but each the values for each subject commence at a random starting point. Such scenarios are denoted as "individual".

Each simulation scenario, consisting of $N$ independent data sets has common values of $\sigma^2$ and of slope coefficients $(\beta_B, \beta_V)$. For each subject, a random walk of up to 20 steps is generated where $W_0 = 100$ and the time-varying drift is $\mu_t = x_B\beta_B + x_{V,t}\beta_V$. If the subject fails for some $t \in (1, \ldots, 20)$ then its failure time $T = t$ is so recorded. Otherwise it is administratively censored at $C = 20$. In some scenarios, subjects in each data set are also censored at random as follows. First, a specified fraction of the subjects are randomly selected for censoring. For each selected subject a censoring time $C$, is also drawn uniformly from $(2, \ldots, 20)$. If $C \leq T$, where $T$ is the event time determined in the previous step, then the new recorded event time is $C$ and the subject is recorded as censored. If $C > T$ then the original event and time are kept.

*Results*

We ran simulations for a variety of scenarios with different sample sizes, different values of $\boldsymbol{\theta} = (\sigma^2, \beta_B, \beta_V)'$ and different combinations of random censoring and individual vs. common time-varying covariates. Here we report results for six scenarios all with $\beta_B = -2, \beta_V = 1.6$, a sample size of $n = 50$ and $N = 1,000$ data sets per scenario. For each value of $\sigma^2 \in (6, 40, 200)$ there are two scenarios, in each case one scenario with individual time-varying covariates and no random censoring; and one scenario with either common time-varying covariates or random censoring. We also report one scenario $\beta_V = 0$ to gauge performance where the time-varying effect is not actually significant.

Using the expectation-maximization algorithm we estimate all three elements of $\boldsymbol{\theta}$. For camparison, we also attempt to estimate $(\sigma^2, \beta_B)$ using the form for threshold regression for baseline covariates only, based on the inverse Gaussian distribution similar to that of Lee and Whitmore (2006), but here with known initial position and unknown variance.

Table 5.1 summarizes the simulation scenarios. Tables 5.2 and 5.3 present the results for

each scenario. For each scenario and for each parameter we report the average estimates, the mean of the estimated standard errors (SEE), the standard deviation of estimates (SSE) and the coverage probability (CP), i.e. the fraction of the estimates for which the true value is within the normal 95% confidence intervals as implied by estimated value and standard errors. With our proposed method, we estimate standard errors as in Section 2.2.3. The standard errors for the inverse Gaussian method are from the numerical Hessian produced by the quasi-Newton optimization algorithm.

We see that in scenarios A–F under our proposed method, the mean estimates of the slope coefficients $\beta_B, \beta_V$ are quite close to the true values, that the mean estimated standard errors are close to the sample standard deviations and that the coverage probabilities are close to 0.95. We note that the slope estimates appear to diverge slightly as $\sigma^2$ increases. Taking into account several other scenarios not reported here, there does not appear to be a consistent pattern in the direction of the divergence.

The estimates of $\sigma^2$ are not as close. In all scenarios reported here (but not in all unreported scenarios) the average estimate of $\sigma^2$ is $2-5\%$ lower then the true value, the standard errors are roughly 10% wider than the sample standard deviations and the coverage probability is roughly 0.92. In other scenarios not reported here the average estimates of $\sigma^2$ were higher than the true values and/or the coverage probabilities were slightly higher than 95%.

The mean parameter estimates for Scenario G, whose data was generated without a time-varying effect were close to the true values, but the estimates had greater variance than predicted by the estimated standard errors.

We observe other patterns in the results which are reassuringly unsurprising – standard errors increase with $\sigma^2$; standard errors increase slightly with random censoring. Standard errors are lower when each subject has its own unique time-varying covariate series than when there is a single time-varying covariate series common to all subjects in the run.

By comparison, the inverse Gaussian method performed poorly, substantially underestimating the baseline coefficient in the presence of a positive time-varying effect, and over-

Table 5.1: Simulation scenarios for a single event process.

| Name | Sample Size | $\mathbf{X}_V$ | Censoring | $\sigma^2$ | $\beta_B$ | $\beta_V$ |
|---|---|---|---|---|---|---|
| A | 50 | *Individual* | *none* | 6 | -2 | 1.6 |
| B | 50 | *Individual* | 0.3 | 6 | -2 | 1.6 |
| C | 50 | *Individual* | *none* | 40 | -2 | 1.6 |
| D | 50 | *Common* | *none* | 40 | -2 | 1.6 |
| E | 50 | *Individual* | *none* | 200 | -2 | 1.6 |
| F | 50 | *Individual* | 0.3 | 200 | -2 | 1.6 |
| G | 200 | *Common* | *none* | 6 | -10 | 0 |

estimating the variance. Even in Scenario G, without a time-varying effect, the baseline coefficient was significantly underestimated. This may be related to the mismatch between the continuous inverse Gaussian distribution and the discrete event times in our data.

### *5.1.2 Simulations for dual competing risks*

The data generating process is essentially the same as for the univariate case described above, with the obvious differences that the innovations to each process are bivariate, and there are four possible outcomes for each subject, where none, one or another, or both of the competing events might be observed. All subjects are censored administratively at a fixed time which varies with the scenario. We do not run any scenarios with random censoring. Every run of every scenario consists of 200 subjects.

The simulated data sets for competing risks are different from the single event data (i.e. Scenario 'A' in one section has no relationship to scenario 'A' in the other section, etc.). A summary of simulation scenarios with the true parameters is given in Table 5.4.

Figure 5.1 gives visual descriptive statistics of the various scenarios, showing the fractions

Table 5.2: Simulation outcomes for a single event process (1).

| | | Proposed Method | | | | Inverse Gaussian | | | |
|---|---|---|---|---|---|---|---|---|---|
| Name | Parameter | Estimate | SEE | SSE | CP | Estimate | SEE | SSE | CP |
| | $\sigma^2$ | 5.65 | 1.84 | 1.61 | 0.92 | 109.1 | 25.1 | 22.1 | 0 |
| A | $\beta_B$ | -2.00 | 0.07 | 0.07 | 0.94 | -0.60 | 0.04 | 0.05 | 0 |
| | $\beta_V$ | 1.60 | 0.09 | 0.08 | 0.94 | - | - | - | - |
| | $\sigma^2$ | 5.85 | 2.08 | 1.87 | 0.92 | 105.4 | 25.7 | 22.1 | 0 |
| B | $\beta_B$ | -2.00 | 0.08 | 0.07 | 0.94 | -0.61 | 0.04 | 0.05 | 0 |
| | $\beta_V$ | 1.60 | 0.09 | 0.09 | 0.95 | - | - | - | - |
| | $\sigma^2$ | 38.61 | 10.83 | 9.52 | 0.92 | 125.9 | 28.2 | 24.1 | 0 |
| C | $\beta_B$ | -2.00 | 0.14 | 0.14 | 0.94 | -0.60 | 0.04 | 0.05 | 0 |
| | $\beta_V$ | 1.60 | 0.18 | .17 | 0.94 | - | - | - | - |
| | $\sigma^2$ | 38.46 | 11.82 | 10.67 | 0.91 | 68.9 | 16.3 | 33.1 | 0.57 |
| D | $\beta_B$ | -2.01 | 0.24 | 0.22 | 0.96 | -0.65 | 0.03 | 0.19 | 0 |
| | $\beta_V$ | 1.61 | 0.30 | 0.28 | 0.96 | | | | |

Table 5.3: Simulation outcomes for single event process (2).

| | | Proposed Method | | | | Inverse Gaussian | | | |
|---|---|---|---|---|---|---|---|---|---|
| Name | Parameter | Estimate | SEE | SSE | CP | Estimate | SEE | SSE | CP |
| | $\sigma^2$ | 192.8 | 53.6 | 49.5 | 0.91 | 188.5 | 38.2 | 38.6 | 0.95 |
| E | $\beta_B$ | -2.02 | 0.28 | 0.27 | 0.95 | -0.60 | 0.05 | 0.06 | 0 |
| | $\beta_V$ | 1.62 | 0.34 | 0.33 | 0.95 | - | - | - | - |
| | $\sigma^2$ | 195.2 | 58.5 | 54.1 | 0.92 | 200.0 | 44.6 | 50.7 | 0.94 |
| F | $\beta_B$ | -2.02 | 0.30 | 0.28 | 0.95 | -0.60 | 0.06 | 0.07 | 0 |
| | $\beta_V$ | 1.62 | 0.37 | 0.35 | 0.95 | - | - | - | - |
| | $\sigma^2$ | 6.04 | 1.53 | 1.80 | 0.89 | 39.1 | 5.8 | 10.9 | 0 |
| G | $\beta_B$ | -10.29 | 0.12 | 0.20 | 0.39 | -8.38 | 0.12 | 0.18 | 0 |
| | $\beta_V$ | 0.028 | 0.023 | 0.029 | 0.74 | - | - | - | - |

of original subjects that terminate with each possible outcome at each time. Data in Scenarios $A, B, C$ are generated with the same parameters, but for correlation, and produce nearly identical outcome plots. We show the plot for $A$ only.

We estimate parameters for the various scenarios using multiple approaches (compressed vs. uncompressed; under the assumption of independent vs. dependent processes), and compare results. For each scenario we generated and analyzed 100 data sets. (exceptions were scenario Q, 200 data sets, and the scenario with uncompressed data for S, 54 data sets due to the computational burden). Results are shown graphically in Figures 5.2, 5.3, 5.4, 5.5. Coverage probabilities for the standard errors are in Table 5.5.

Scenarios where parameters are estimated based on uncompressed data, and under the assumption of dependent risks, appear to provide unbiased estimates of all parameters, with reasonable coverage probabilities. Coverage probabilities for $\sigma_a, \sigma_b$ are not as good as for the $\boldsymbol{\beta}$ parameters. Both the dispersion and standard errors of the $\rho$ estimates are quite wide.

Estimates made from compressed data are somewhat biased, in the cases we have examined by about 2%, up to 6%. Estimates made under the incorrect assumption that the event processes are independent also introduce bias. Bias introduced by compression may affect different parameters than bias introduced by estimates made under the incorrect assumption of independent events. The standard errors for compressed data estimates have poor coverage, attributable to bias, but the estimated standard errors are consistent with the standard deviation of the parameter estimates. Standard errors estimated under the independent risks assumption have both poor coverage and are also substantially smaller than the standard deviations of the parameter estimates – where the latter are fairly close to the corresponding quantities for the proper estimates.

We make special note of the pitfalls of model misspecification when the "Joint" mode assumption is applied in scenarios where outcomes are mutually exclusive. Scenario N data was generated such that event $b$ dominates event $a$. When the correct model is applied the estimates appear unbiased and generally on-target coverage probabilities. But when incorrectly analyzed under the assumption of Joint mode, the bias is severe and coverage

Table 5.4: Simulation scenarios for dual competing risks. "Max. Y" indicates the administrative censoring time. "Mode" indicates the outcome mode as defined in Section 3.4. "Covariates" indicates whether a subject has the same time-varying covariate series for process 'a' as for 'b' or whether they're different.

| Name | $\boldsymbol{\beta}_a$ | $\boldsymbol{\beta}_b$ | $\sigma$ | $\rho$ | Max. Y | Mode | Covariates |
|---|---|---|---|---|---|---|---|
| A | (2,-1.5) | (-2,-2) | (1,4) | 0.5 | 20 | Joint | Different |
| B | (2,-1.5) | (-2,-2) | (1,4) | -0.5 | 20 | Joint | Different |
| C | (2,-1.5) | (-2,-2) | (1,4) | 0.85 | 20 | Joint | Different |
| H | (1,-0.5) | (-4,0.5) | (1,2) | 0.5 | 60 | Joint | Different |
| N | (2,-1.5) | (-2,-2) | (1,4) | 0.5 | 20 | Dominant | Same |
| Q | (2,-1.5) | (-2,-2) | (1,4) | -0.85 | 20 | Joint | Same |
| S | (-0.5,-1.2) | (-1.5,-0.7) | (15,10) | -0.6 | 40 | Joint | Same |

probabilities poor.

## *5.2 Mortgage Data*

We apply our methods to a data sample from the Freddie Mac Single Family Loan-Level Dataset [1] , representing 30-year fixed-rate mortgages purchased or guaranteed by Freddie Mac from 1999 through 2013. The dataset contains both loan origination characteristics as well as longitudinal monthly payment events for each loan through 2013, including missed payments and default and/or full prepayment events.

### *5.2.1 Single event process – mortgage prepayment*

Our analysis data set consists of 398 mortgages with the following characteristics:

- Originated between February 1999 and January 2012; observed through December 2012.

[1] On the Internet at `http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html`

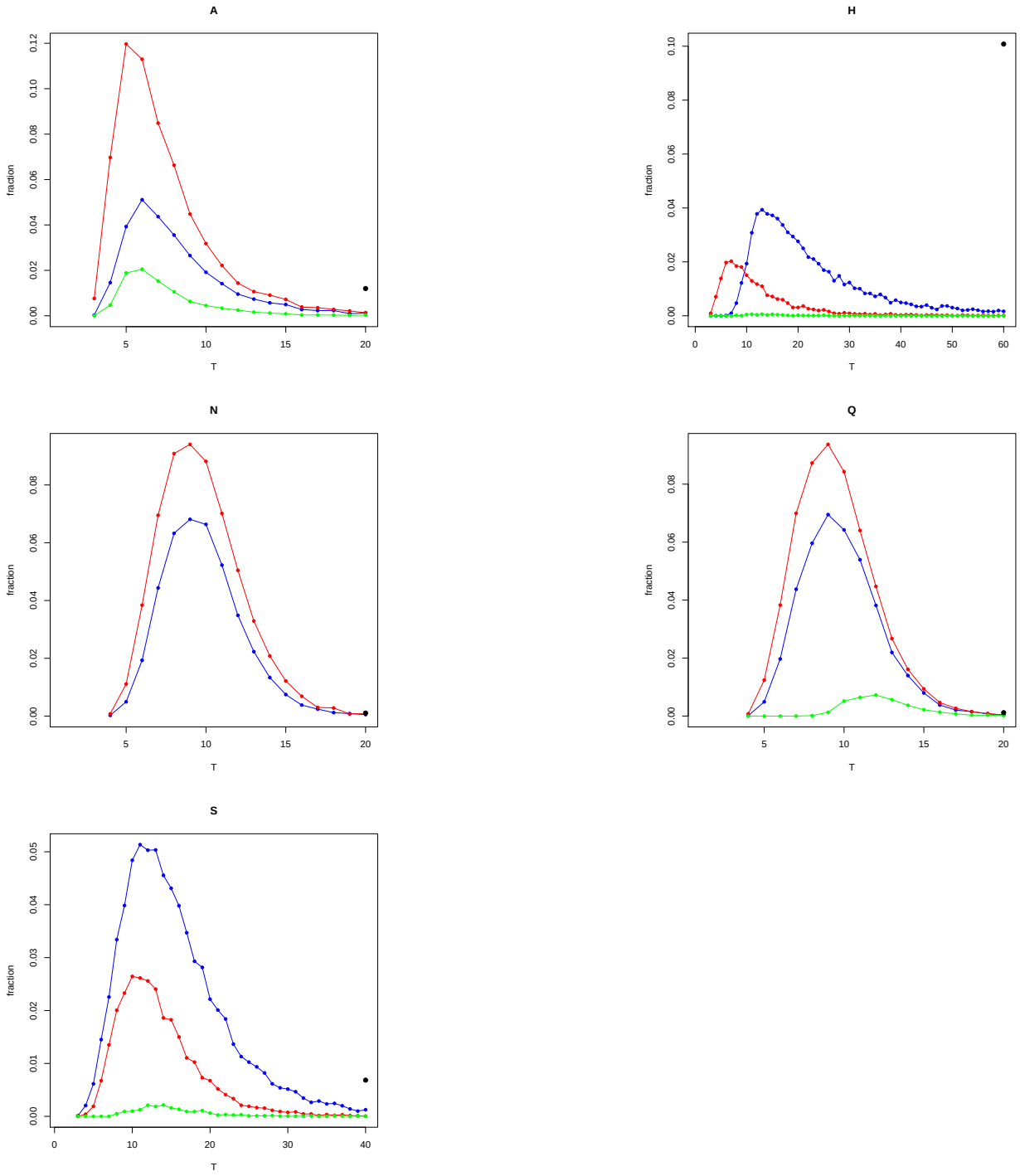


Figure 5.1: Descriptive outcome plots for dual event simulations, illustrating the mean fraction of subjects that experience a given event at a given time. Blue corresponds to event 1, Red to event 2, Green to simultaneous events, and the Black dot at the final, administrative censoring, time represents the fraction of subjects left surviving at the end.

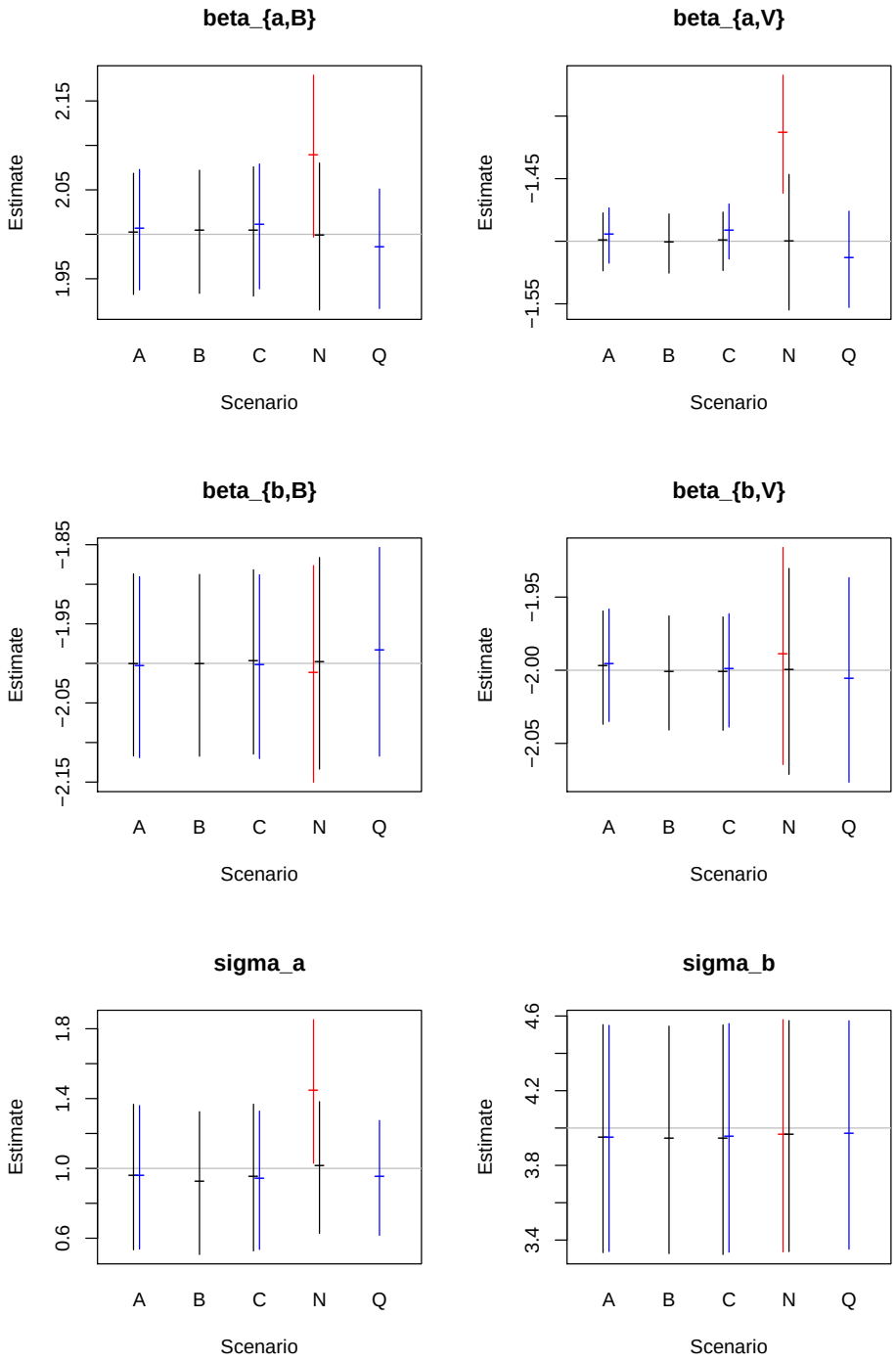


Figure 5.2: Average point estimates and average confidence intervals for dual event simulations. Grey horizontal lines represent true values. Estimates made under correct model assuming dependent risks and correct mode are in black; assuming independent processes and uncompressed data in blue. In scenario N data was generated with process b is Dominant. Red indicates estimates made under incorrect assumption of Joint mode.

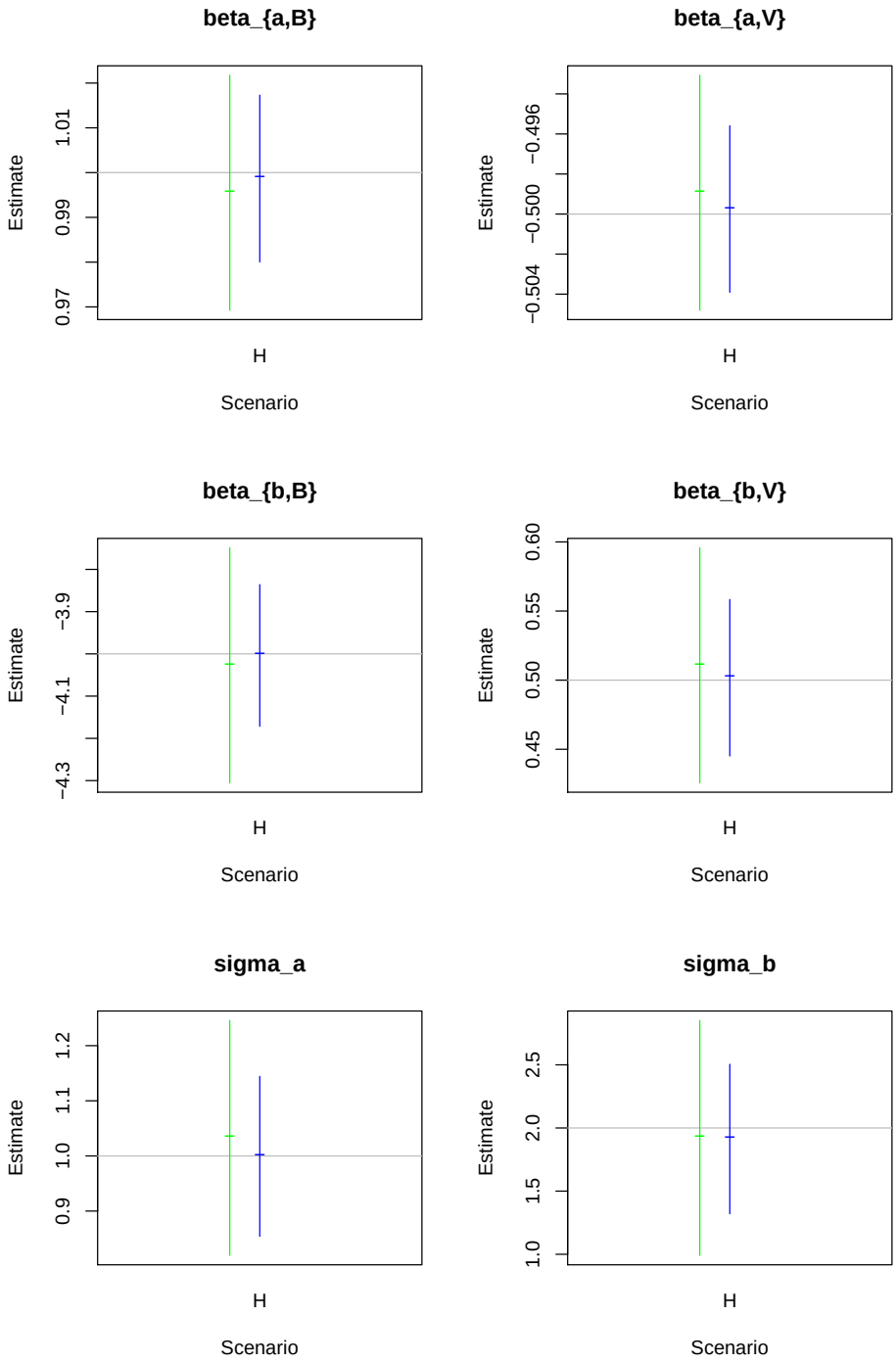


Figure 5.3: Average point estimates and average confidence intervals for dual event simulations. Grey horizontal lines represent true values. Estimates made assuming dependent risks and compression factor 3 in green; assuming independent processes and uncompressed data in blue.

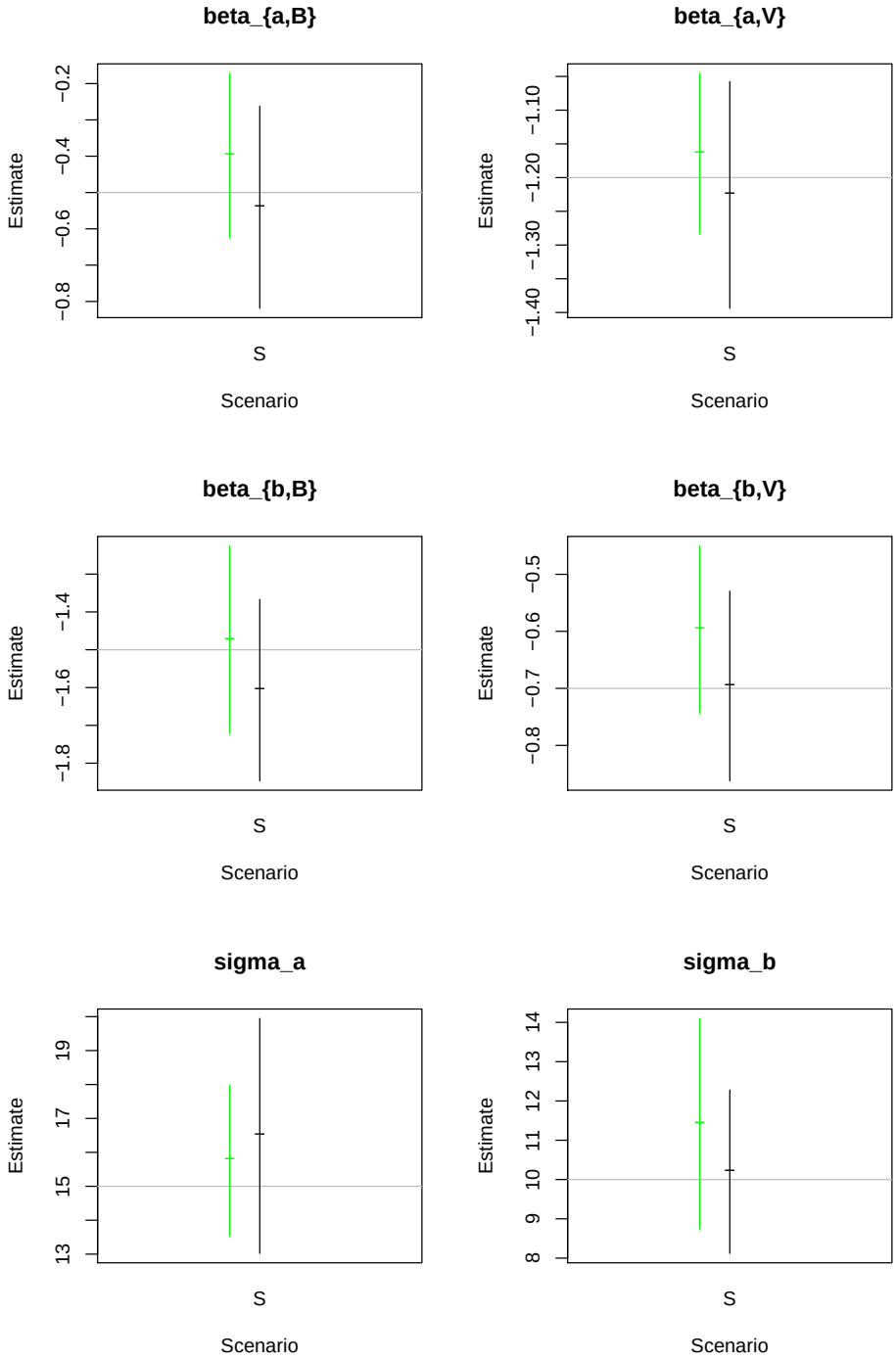


Figure 5.4: Average point estimates and average confidence intervals for dual event simulations. Grey horizontal lines represent true values. Estimates made assuming dependent risks, and uncompressed data are in black; assuming dependent risks and compression factor 3 in green.

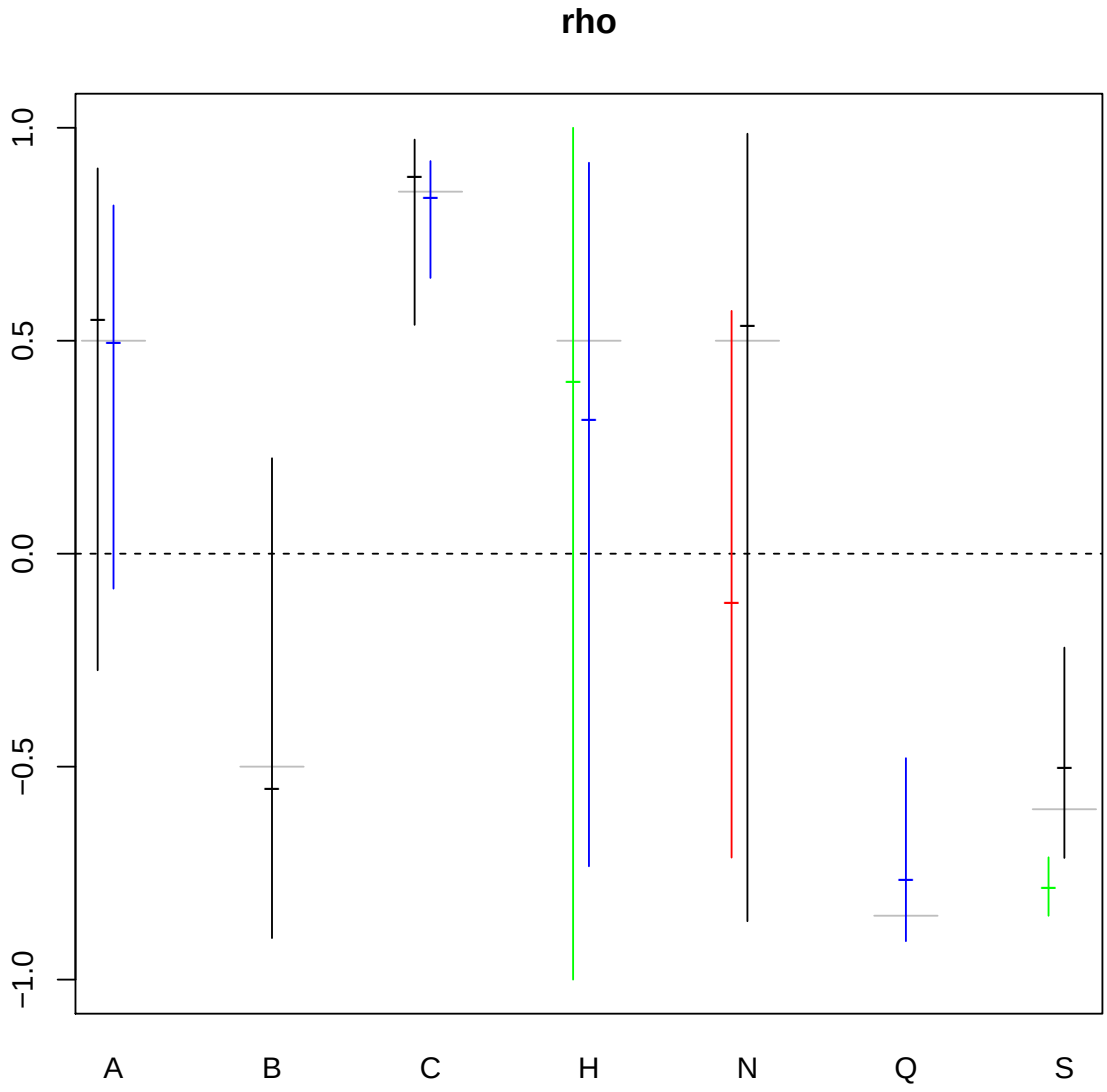


Figure 5.5: Average point estimates and average confidence intervals of $\rho$ for all dual event simulation scenarios. Grey horizontal lines indicate true values. Estimates made under correct model assuming dependent risks, uncompressed data and correct mode are in Black; assuming dependent risks and compression factor 3 in Green; assuming independent processes and uncompressed data in Blue. In scenario N data was generated with process b is Dominant. Red indicates estimates made under incorrect assumption of Joint mode.

| | $\beta_{a,B}$ | $\beta_{a,V}$ | $\beta_{b,B}$ | $\beta_{b,V}$ | $\sigma_a$ | $\sigma_b$ | $\rho$ |
|---|---|---|---|---|---|---|---|
| A | 0.96 | 0.96 | 0.97 | 0.96 | 0.96 | 0.92 | 0.94 |
| A $\times 1$ | 0.97 | 0.94 | 0.97 | 0.96 | 0.95 | 0.92 | 0.87 |
| B | 0.94 | 0.94 | 0.95 | 0.98 | 0.96 | 0.95 | 0.95 |
| C | 0.97 | 0.98 | 0.94 | 0.97 | 0.94 | 0.95 | 1.00 |
| C $\times 1$ | 0.95 | 0.94 | 0.94 | 0.97 | 0.96 | 0.93 | 1.00 |
| H3 | 0.93 | 0.92 | 0.95 | 0.97 | 0.88 | 0.94 | 0.99 |
| H $\times 1$ | 0.95 | 0.94 | 0.96 | 0.98 | 0.93 | 0.92 | 0.79 |
| N Joint | 0.50 | 0.12 | 0.93 | 0.97 | 0.42 | 0.99 | 0.60 |
| N Dominant | 0.98 | 0.95 | 0.90 | 0.98 | 0.96 | 0.98 | 0.97 |
| Q $\times 1$ | 0.92 | 0.90 | 0.96 | 0.94 | 0.94 | 0.95 | 0.98 |
| S3 | 0.86 | 0.87 | 0.98 | 0.76 | 0.95 | 0.94 | 0.43 |
| S | 0.96 | 0.98 | 0.91 | 0.87 | 0.89 | 0.91 | 0.91 |

Table 5.5: Coverage probabilities for dual-event simulated data.

- Single-family, owner-occupied residence
- Loan purpose is for new purchase (as opposed to refinance)
- No prepayment penalty
- Borrower is not a first-time home buyer
- The mortgaged property is located in one of the 19 metropolitan areas for which there was a S&P Case–Shiller Home Price Index since January 1999.

In addition to the Freddie Mac mortgage dataset, we include two macroeconomic time series, the S&P Case–Shiller Home Price Indices[2], and mortgage interest rates [The "30-Year Fixed Rate Mortgage Average in the United States" (MORTGAGE30US) series from the Federal Reserve Bank of St. Louis [3]]

Table 5.6 summarizes descriptive statistics for prepayment, default and adminstratively censored events.

We model associations between time to prepayment and certain origination and time-varying covariates. For the purpose of this analysis the relatively small number of default events are treated as non-informative censoring. The baseline (origination) covariates taken from the Freddie Mac data set are "FICO", the borrower's credit score (in the range 300-850); the loan interest rate; the base 10 logarithm of the initial loan amount. The time-varying covariates are the change since origination in the US average 30-year fixed mortgage interest rate; the logarithmic change since origination in the S&P Case–Shiller metro housing price index for the property's metropolitan area.

We present our estimates under two first hitting time models discussed in Chapter 2. In the first model, the drift is a linear function of both baseline and time-varying covariates

[2]On the Internet at `http://www.spindices.com/index-family/real-estate/sp-case-shiller` Accessed July 5, 2013

[3]On the Internet at `http://research.stlouisfed.org/fred2/series/MORTGAGE30US` Accessed July 5, 2013

and the variance is an unknown parameter common to all subjects. As noted in Section 2.1, the initial position $W_0$ must then be a fixed *a priori* constant. We choose the arbitrary but convenient value of 100 as it allows the parameters to be interpreted in terms of percentage change per unit time from the initial state toward the terminal event. In the second model, the variance is fixed at 1, the mean initial position is a linear function of the baseline covariates and the drift is a linear function of the two time-varying covariates. Alongside the first hitting time models are corresponding estimates from an analogous Cox proportional hazards regression. The results are shown in Table 5.7.

We interpret the estimates as follows. In the primary model (Model 1) the coefficients of baseline covariates (FICO, loan interest rate, log of loan amount) have similar interpretations as that of loan interest rate ($-1.20$): A positive unit difference in loan interest rate is associated with an expected increased drift toward prepayment in each month equivalent to 1.20% of the original distance from prepayment. The coefficients of time-varying covariates (change in new mortgage interest rate, log change in housing price index ) have similar interpretations as that of log change in housing price index ($-3.48$): A positive unit difference in change in log housing price index since origination at a given month is associated with an expected increased drift toward prepayment in that month equivalent to 3.48% of the original distance from prepayment.

In Model 2, the interpretation of loan interest rate ($-0.744$) and coefficients for other baseline covariates is somewhat different: A positive unit difference in loan interest rate is associated with an expected negative difference of $-0.744$ abstract unit initial distance from prepayment. The interpretation of log change in housing price index ($-0.234$) and coefficients of other time-varying covariates is slightly different from Model 1: A positive unit difference in log change in housing price index since origination to a given month is associated with an expected increased drift toward prepayment in that month equivalent to 0.234 abstract units of distance from prepayment.

We see that the coefficient signs are consistent across all 3 models, in the sense that the signs for the first hittting time models are exactly opposite the corresponding signs under

Table 5.6: Summary statistics of mortgage event times for single event process.

| Outcome | Number | Mean Time to Event (months) | Interquartile Range of Times |
|---|---|---|---|
| Prepaid | 312 | 32.9 | (14,46.5) |
| Defaulted | 12 | 49.2 | (34.8,63.0) |
| Censored | 74 | 50.5 | (22.5,81.2) |

the Cox model. This makes sense given that a negative sign in the threshold model and a positive sign in the proportional hazards model both imply shorter time to event. The coefficients are also consistent in significance across models. All coefficients appear to be significant in all models, with the exception of FICO, which is not significant in any model.

The significant estimates are sign-wise consistent with economic intuition, that homeowners should tend to refinance or buy new homes as interest rates decline; are likely to be more inclined to refinance when they have a large principal balance and/or a high rate to begin with; and may be more inclined to sell their current home the more it has appreciated in value.

The two first hitting time models are also consistent in another sense. As noted, at least one of the starting position and variance must be a fixed constant. The invariant is the ratio of starting position over standard deviation, $W_0/\sigma$. Under the first model, the ratio of fixed initial value to estimated $\hat{\sigma}$ is 4.76. Under the second model with unit variance, the mean fitted value is a close 4.53.

### *5.2.2 Dual competing risks – mortgage default and prepayment*

The data for the mortgage competing risks study is from the same Freddie Mac set as the prepayment only data. However it is a different sample and consists of 1,000 mortgages with the following characteristics:

Table 5.7: Parameter estimates for mortgage data (prepayment event only).

| | Model 1 | SE | Model 2 | SE | CoxPH Coef. | SE |
|---|---|---|---|---|---|---|
| (Intercept) | 27.3 | 6.0 | 28.0 | 5.5 | - | - |
| FICO | -0.0052 | 0.0034 | -0.0055 | 0.0036 | 0.0015 | 0.0011 |
| Loan interest rate | -1.20 | 0.26 | -0.744 | 0.15 | 0.34 | 0.08 |
| log Initial loan amount | -3.17 | 0.99 | -3.47 | 1.00 | 1.06 | .33 |
| $\Delta$ new mortage interest rate | 2.96 | 0.38 | 0.177 | 0.0084 | -0.85 | 0.09 |
| $\Delta$ log Case-Shiller index | -3.48 | 0.85 | -0.234 | 0.031 | 1.05 | .25 |
| $\sigma$ | 21.0 | 0.78 | - | - | - | - |

- Originated between February 2003 and January 2010; observed through December 2012.
- Single-family, owner-occupied residence
- Loan purpose is for new purchase (as opposed to refinance)
- No prepayment penalty
- Borrower is not a first-time home buyer
- The mortgaged property is located in the Las Vegas metropolitan area

The metropolitan area and time period were chosen retrospectively to have relatively high rates of both defaults and prepayments in order to illustrate the statistical methods. The sample was chosen randomly from that universe.

The model includes all covariates used in the prepayment-only model with some slight changes, along with some additional covariates. In the prepayment-only model, the time-varying covariates for changes in log housing price index and in new mortgage interest rate since origination are continuous variables indicating the actual change in those quantities. In the competing risk model these are binary variables, with 1 indicating that the change is negative. We add two new baseline covariates, Loan-to-value ratio (e.g. a value of 80 indicates that the borrower made a 20% down payment and financed 80% of the home value), and Debt-to-Income Ratio, the ratio of the borrower's total monthly debt payments divided by total monthly income (as a percentage) and as reported to the lender prior to origination. In addition, we add a new time-varying covariate indicating the monthly unemployment rate (as a percentage) for the Las Vegas metro area, as reported by the U.S. Bureau of Labor Statistics[4].

Descriptive statistics of baseline covariates are in Table 5.8. Descriptive statistics of mortgage outcomes are shown graphically in Figure 5.6.

The sample includes mortgages that were observed for up to 118 months. Computation for the full sample under the dependent risk assumption and without time compression would be impractically time-consuming. We estimate parameters using a few approximate approaches and compare the results. Finally, we validate these estimates by computing the likelihood of the full uncompressed sample under the dependent risk model with each set of the estimated parameters.

Figures 5.7, 5.8, 5.9, 5.10, 5.11, 5.12 illustrate the estimates and confidence intervals graphically. Table 5.9 provides the point estimates from each approach, and also reports the log likelihood of the full uncompressed sample under each set of estimates.

As we see from the plots, there is little difference between estimates made with a compression factor of 3 and a compression factor of 6, and only slight differences, all else equal, between estimates in Default-dominant vs. Prepay-dominant mode. Most of the parameters

[4]On the Internet at `http://data.bls.gov/cgi-bin/dsrv?la` Accessed December 22, 2014

yield qualitatively similar estimates across approaches. The consistent outlier is the estimate on the far left of every plot and the table, which was made under the assumption of independent processes in Joint mode. The approach which produced the estimates yielding the highest likelihood for the uncompressed sample under the correct model, was from independent processes in Default-dominant mode. These estimates prevailed both for the training sample and for a hold-out sample from the same population. This outcome might appear to throw cold water on the benefits of time compression. On the other hand, we also determined, by varying $\sigma_{Prepay}$ and $\rho$ that the time-compressed estimates as reported, are local, not global maxima. It appears that the gradient descent algorithm got stuck near the initial value $\sigma_{Prepay} = 50$, when in fact $\sigma_{Prepay} \approx 30$ yields a considerably higher likelihood.

We do notice that some parameters appear to be consistently associated with outcome events across estimation approaches, with high or modest significance, and generally consistent with economic intuition. The interpretation of the parameters is as in the single risk model, where a negative coefficient represents a positive association between the covariate and event risk (equivalently, relatively shorter time to event).

In this sample there are no parameters which are consistently and significantly associated with time to prepayment. However there appears to be some evidence that Loan-to-value ratio and Loan interest rate are positively associated with longer time to prepay, and decline in interest rates is associated with shorter time to prepay.

A few covariates are significantly associated with default risk. Consistent with economic intuition and broadly with the Merton model, default risk increases with loan-to-value ratio, and housing price declines. Homeowners who have high debt relative to their home equity are reasonably more likely to default on their loans than other borrowers. Also, as one would expect, default risk is negatively associated with FICO credit score, i.e. higher FICO scores are associated with lower risk of (longer time to) default.

Table 5.8: Descriptive statistics for the dual-event mortgage data baseline covariates. Mean (standard deviation) for the entire sample and means by respective outcome.

| | Population | Surviving | Prepay | Default |
|---|---|---|---|---|
| Loan amount ($\log_{10}$) | 5.28 (0.19) | 5.24 | 5.27 | 5.37 |
| Loan-to-value ratio (%) | 77.14 (15.95) | 76.46 | 75.46 | 82.23 |
| Debt-to-income ratio (%) | 37.39 (11.74) | 37.27 | 36.23 | 40.50 |
| Loan interest rate (%) | 5.98 (0.49) | 5.80 | 6.00 | 6.16 |
| FICO score | 728.78 (54.06) | 734.65 | 729.92 | 718.85 |

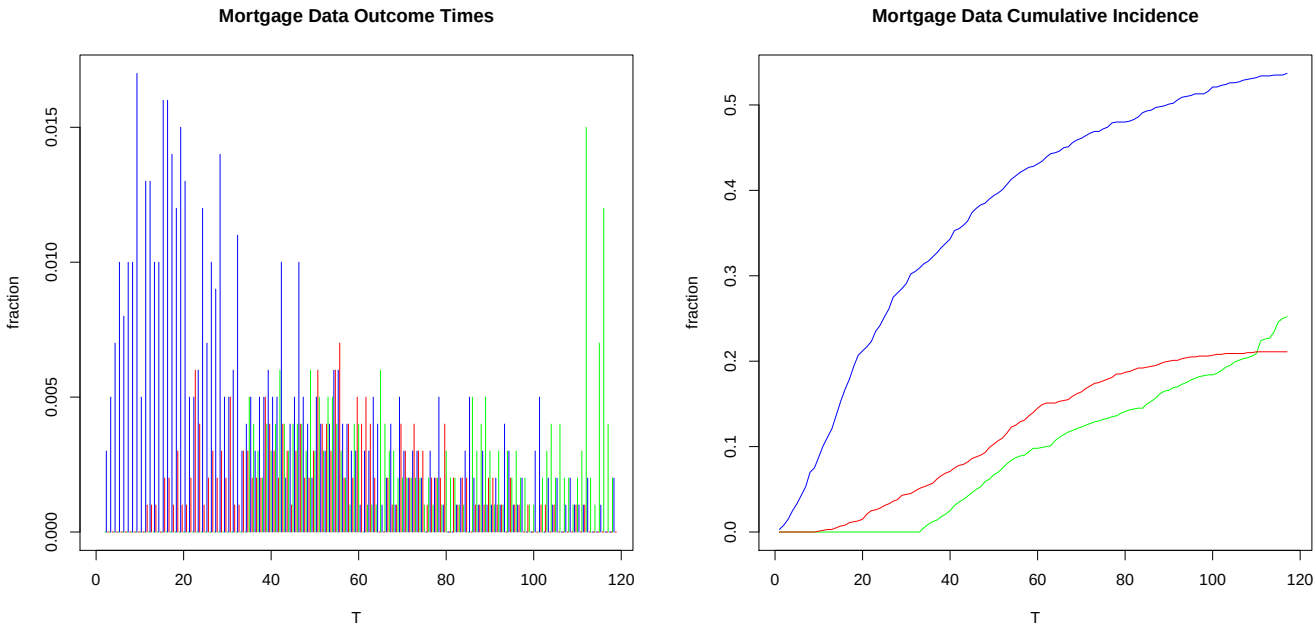


Figure 5.6: Dual-event mortgage outcome times as a histogram (left), indicating fraction of subjects that experience given outcome at each T; and as a cumulative incidence plot (right). Prepayments shown in blue, defaults in red, survival in green.

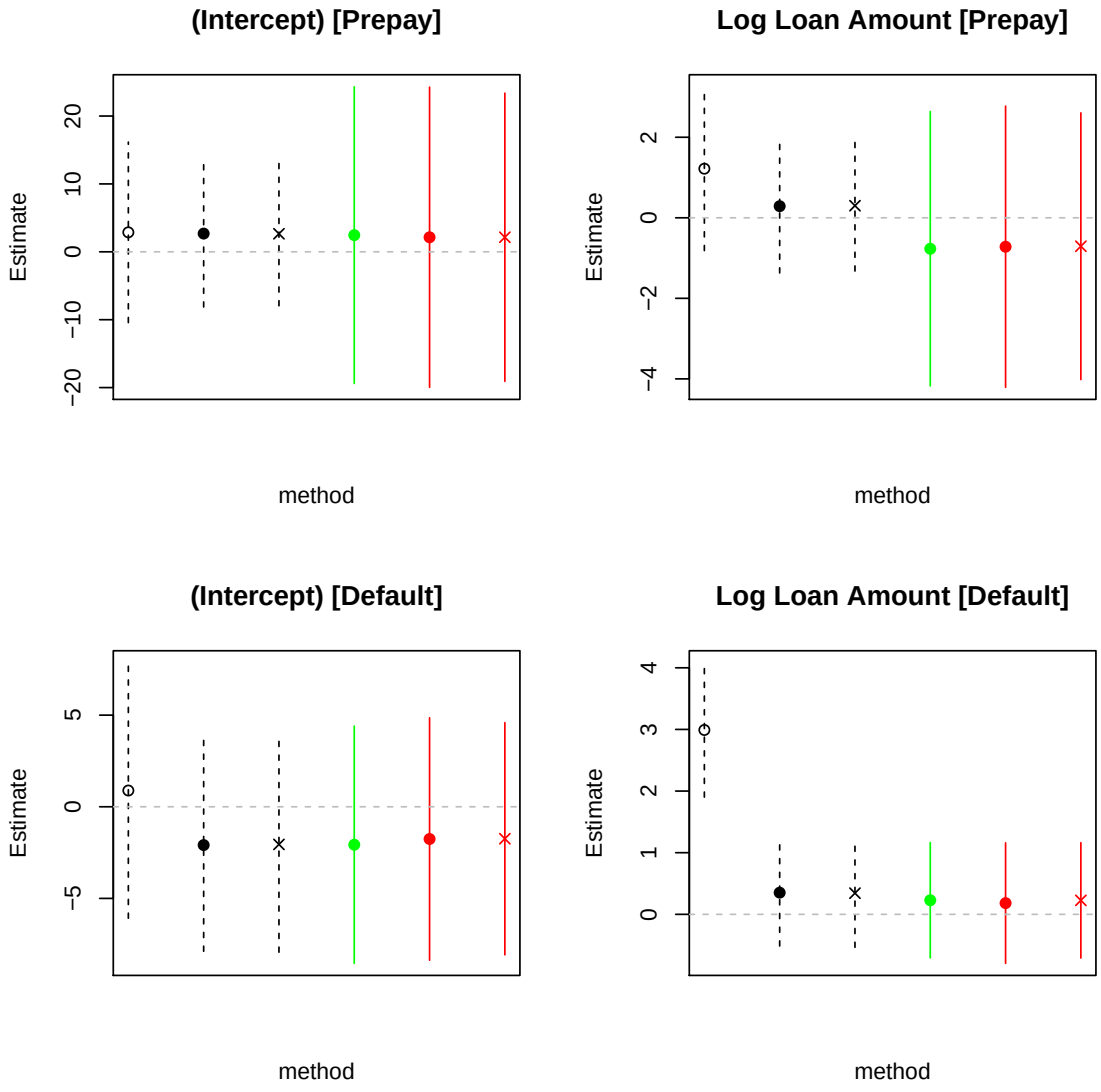


Figure 5.7: This plot, and the following similarly formatted plots compare the parameter estimates and confidence intervals for the dual-event mortgage data, estimated under various assumptions. In each plot, the left-most estimates, plotted in black, were obtained under the assumption of independent processes, with $\rho$ then estimated conditional upon the per-process estimates. The rightmost estimates, in red, were obtained with a compression factor of 6. The middle estimate, in green was obtained with a compression factor of 3. In all cases, the glyph representing the point estimate indicates the assumed mode. An unfilled circle indicates "Joint" mode, a filled circle indicates "Default-dominant", and an X indicates "Prepay-dominant".

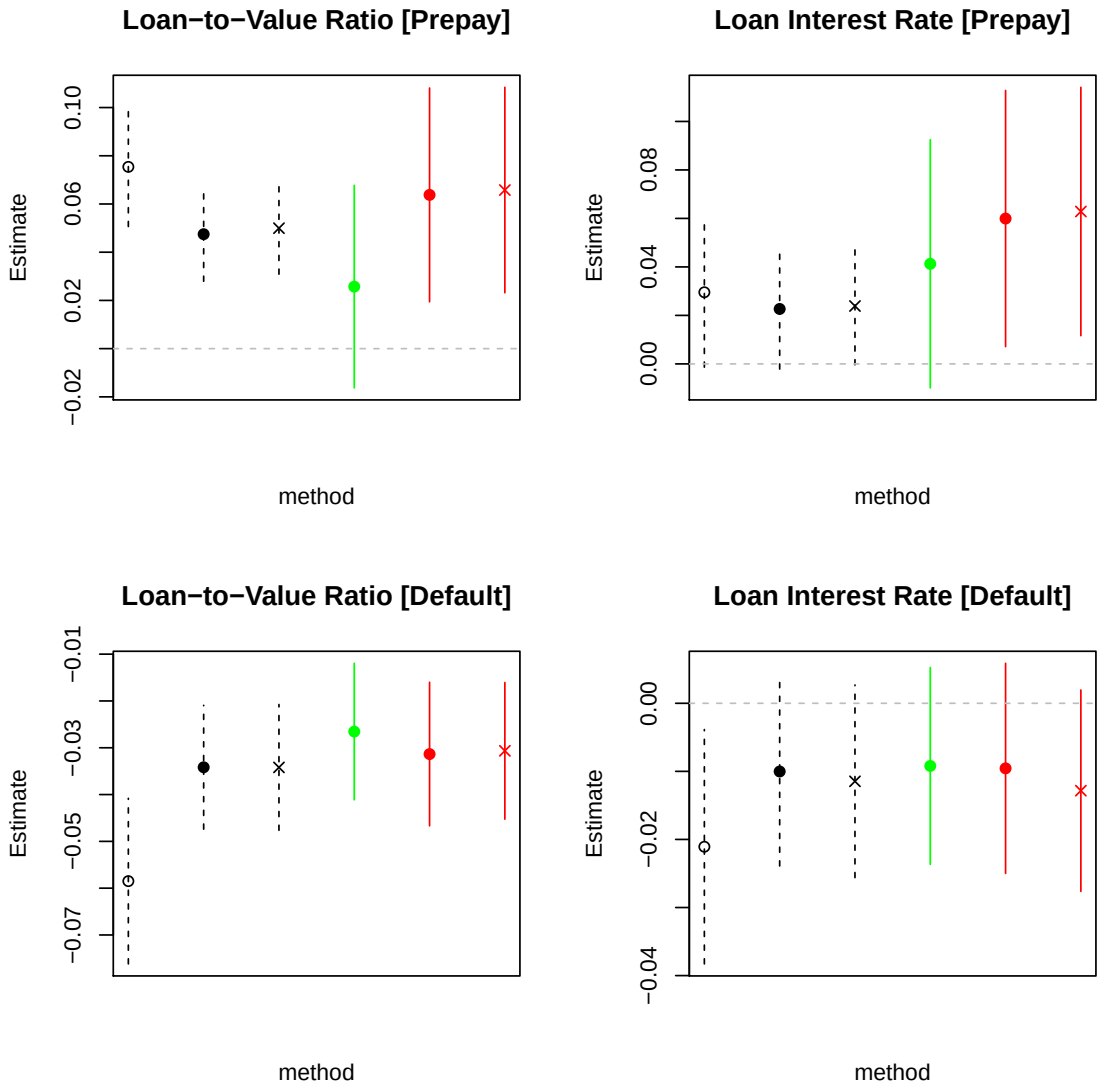


Figure 5.8: Dual-event mortgage estimates for Loan-to-Value Ratio and Loan Interest Rate. The left-most estimates, plotted in black, were obtained under the assumption of independent processes, with $\rho$ then estimated conditional upon the per-process estimates. The rightmost estimates, in red, were obtained with a compression factor of 6. The middle estimate, in green was obtained with a compression factor of 3. In all cases, the glyph representing the point estimate indicates the assumed mode. An unfilled circle indicates "Joint" mode, a filled circle indicates "Default-dominant", and an X indicates "Prepay-dominant".

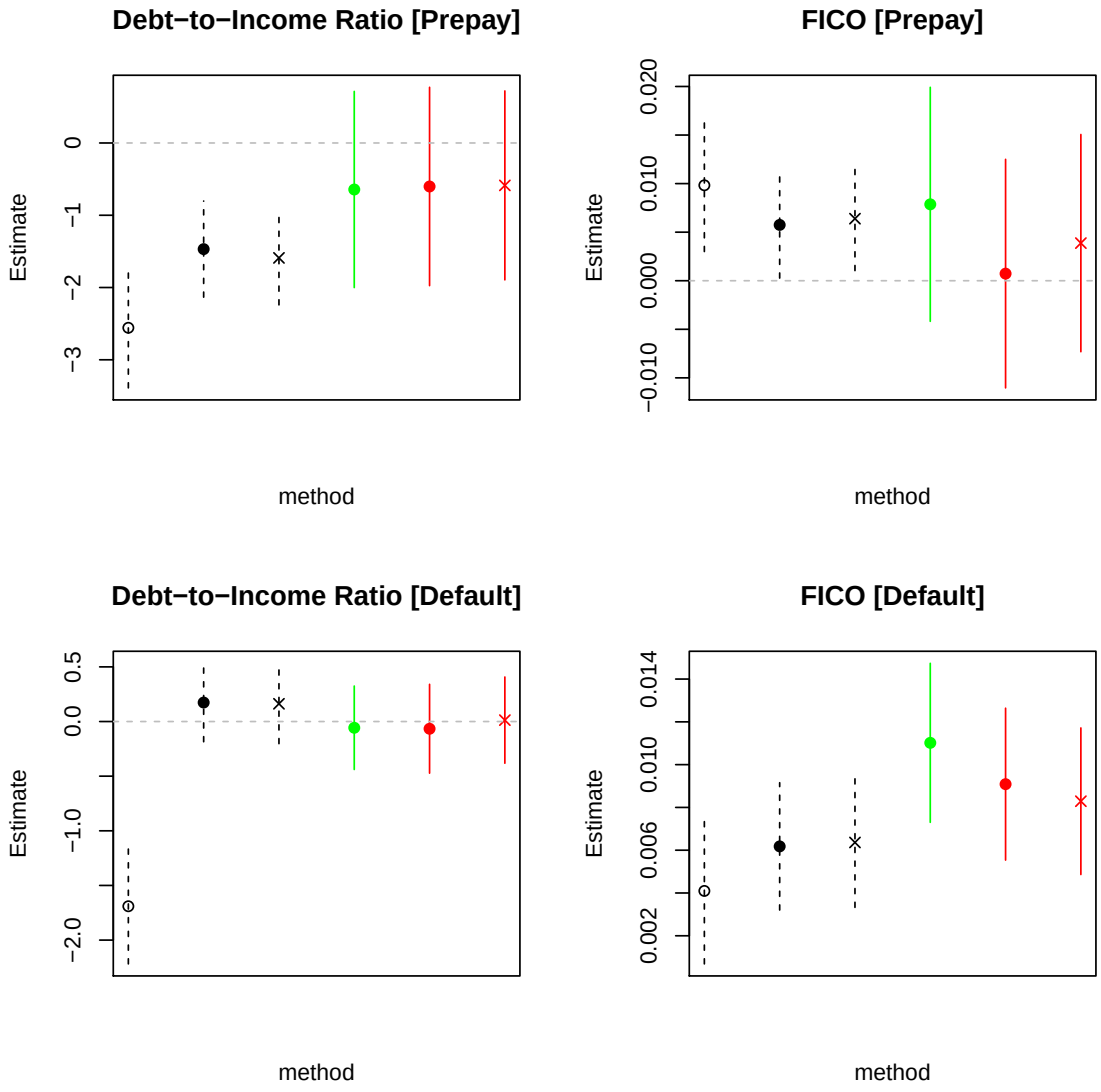


Figure 5.9: Dual-event mortgage estimates for Debt-to-Income Ratio and FICO score. The left-most estimates, plotted in black, were obtained under the assumption of independent processes, with $\rho$ then estimated conditional upon the per-process estimates. The rightmost estimates, in red, were obtained with a compression factor of 6. The middle estimate, in green was obtained with a compression factor of 3. In all cases, the glyph representing the point estimate indicates the assumed mode. An unfilled circle indicates "Joint" mode, a filled circle indicates "Default-dominant", and an X indicates "Prepay-dominant".

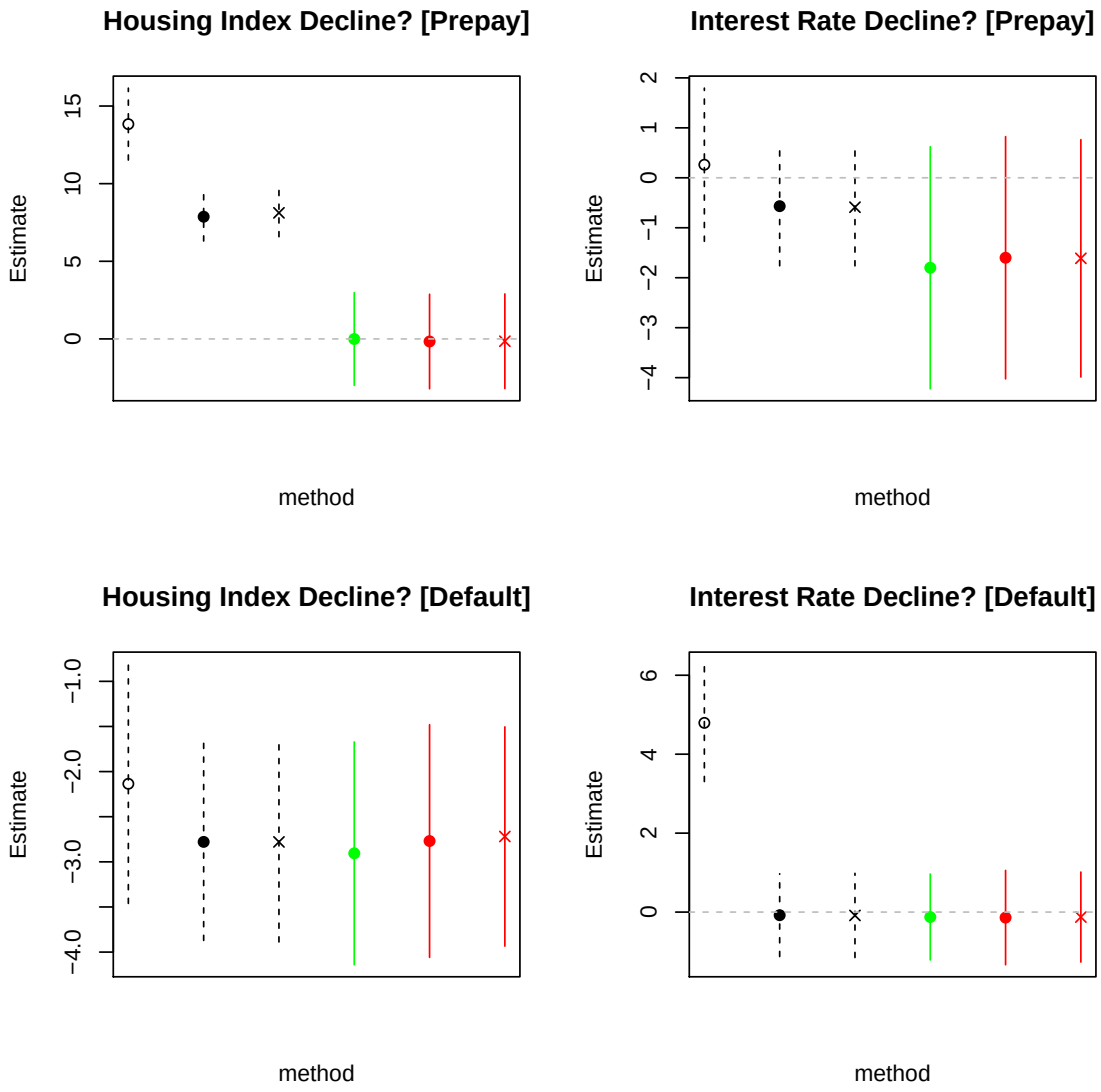


Figure 5.10: Dual-event mortgage estimates for Housing Index Decline and Interest Rate Decline. The left-most estimates, plotted in black, were obtained under the assumption of independent processes, with $\rho$ then estimated conditional upon the per-process estimates. The rightmost estimates, in red, were obtained with a compression factor of 6. The middle estimate, in green was obtained with a compression factor of 3. In all cases, the glyph representing the point estimate indicates the assumed mode. An unfilled circle indicates "Joint" mode, a filled circle indicates "Default-dominant", and an X indicates "Prepay-dominant".

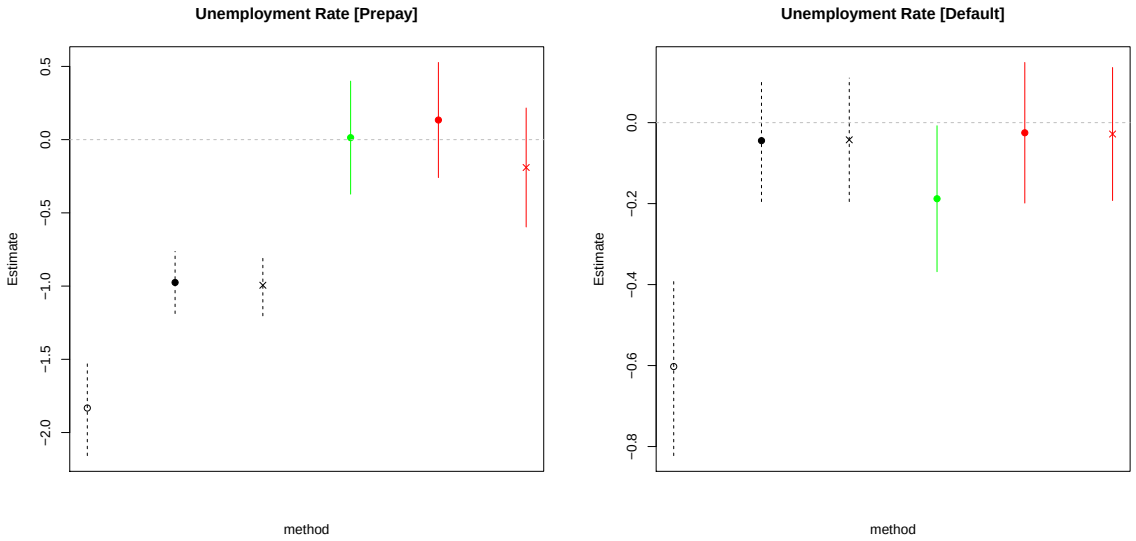


Figure 5.11: Dual-event mortgage estimates for Unemployment Rate. The left-most estimates, plotted in black, were obtained under the assumption of independent processes, with $\rho$ then estimated conditional upon the per-process estimates. The rightmost estimates, in red, were obtained with a compression factor of 6. The middle estimate, in green was obtained with a compression factor of 3. In all cases, the glyph representing the point estimate indicates the assumed mode. An unfilled circle indicates "Joint" mode, a filled circle indicates "Default-dominant", and an X indicates "Prepay-dominant".

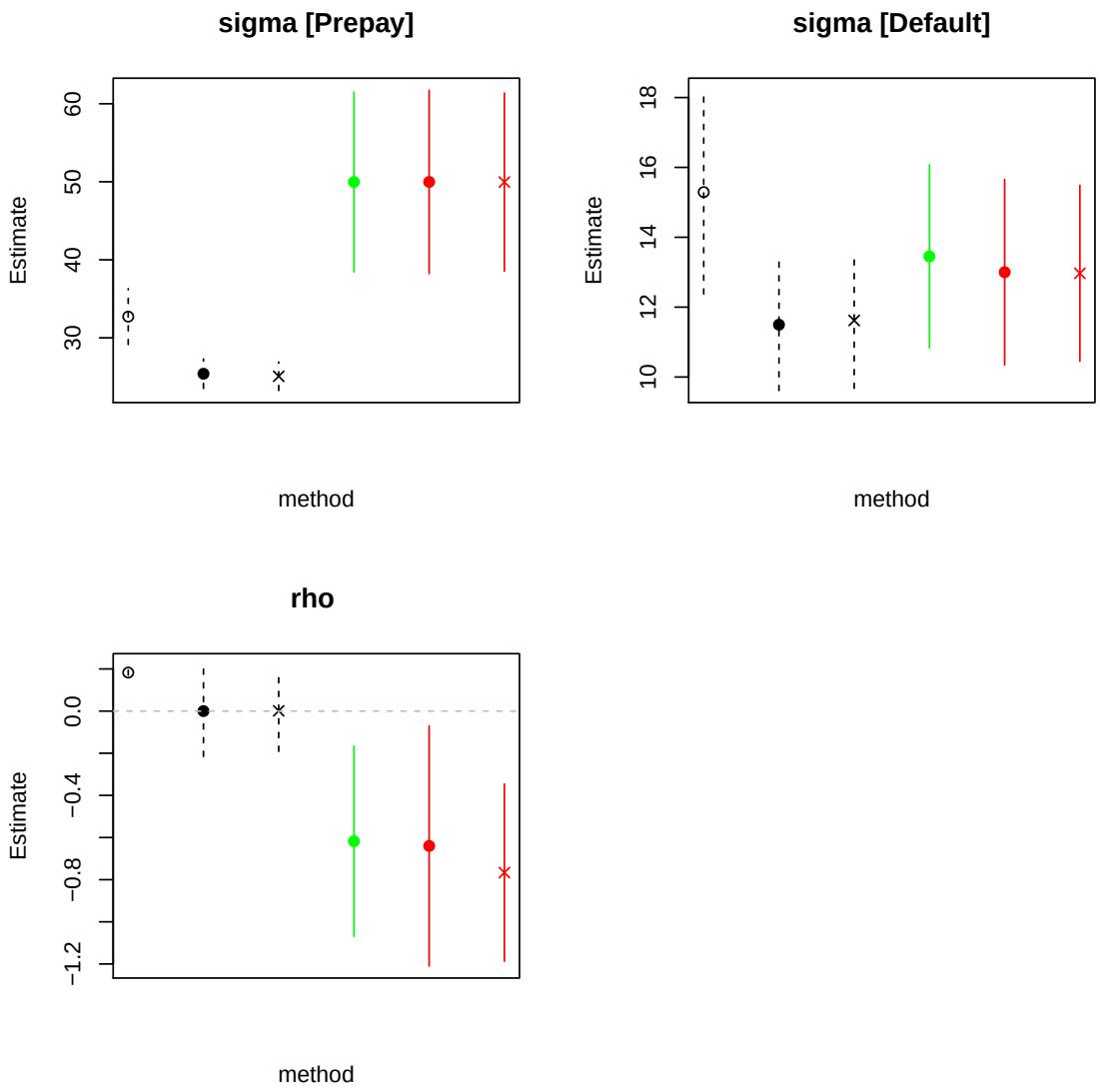


Figure 5.12: Dual-event mortgage estimates for $\sigma$ and $\rho$. The left-most estimates, plotted in black, were obtained under the assumption of independent processes, with $\rho$ then estimated conditional upon the per-process estimates. The rightmost estimates, in red, were obtained with a compression factor of 6. The middle estimate, in green was obtained with a compression factor of 3. In all cases, the glyph representing the point estimate indicates the assumed mode. An unfilled circle indicates "Joint" mode, a filled circle indicates "Default-dominant", and an X indicates "Prepay-dominant".

Table 5.9: Parameter estimates for the dual-event mortgage data, estimated under various approaches. The estimation approach is described by the 3-character code at the column heading. The first letter specifies assumption of Independent vs. Dependent processes. The second letter indicates the mode (Joint, Default-dominant or Prepay-dominant). The numeral indicates the compression factor (or 1 if no compression). The left-right order of columns corresponds to the order of plotted estimates in Figures 5.7–5.12. The bottom rows show the log likelihood of a full uncompressed sample under the dependent risk model with the corresponding set of estimated parameters, both in-sample and for a same-sized hold out sample from the same population.

| | IJ1 | ID1 | IP1 | DD3 | DD6 | DP6 |
|---|---|---|---|---|---|---|
| Prepayment | | | | | | |
| (Intercept) | 2.871 | 2.689 | 2.657 | 2.457 | 2.144 | 2.146 |
| Log Loan Amount | 1.220 | 0.289 | 0.301 | -0.769 | -0.719 | -0.706 |
| Loan-to-Value Ratio | 0.075 | 0.047 | 0.050 | 0.026 | 0.064 | 0.066 |
| Loan Interest Rate | 0.030 | 0.023 | 0.024 | 0.041 | 0.060 | 0.063 |
| Debt-to-Income Ratio | -2.557 | -1.468 | -1.590 | -0.644 | -0.602 | -0.587 |
| FICO | 0.010 | 0.006 | 0.006 | 0.008 | 0.001 | 0.004 |
| Housing Index Decline? | 13.836 | 7.871 | 8.121 | -0.006 | -0.166 | -0.150 |
| Interest Rate Decline? | 0.262 | -0.567 | -0.589 | -1.800 | -1.600 | -1.612 |
| Unemployment Rate | -1.833 | -0.975 | -0.994 | 0.014 | 0.134 | -0.190 |
| Default | | | | | | |
| (Intercept) | 0.887 | -2.088 | -2.054 | -2.070 | -1.757 | -1.744 |
| Log Loan Amount | 2.991 | 0.352 | 0.343 | 0.228 | 0.182 | 0.227 |
| Loan-to-Value Ratio | -0.058 | -0.034 | -0.034 | -0.027 | -0.031 | -0.031 |
| Loan Interest Rate | -0.021 | -0.010 | -0.011 | -0.009 | -0.010 | -0.013 |
| Debt-to-Income Ratio | -1.690 | 0.174 | 0.162 | -0.057 | -0.066 | 0.013 |
| FICO | 0.004 | 0.006 | 0.006 | 0.011 | 0.009 | 0.008 |
| Housing Index Decline? | -2.135 | -2.779 | -2.779 | -2.906 | -2.769 | -2.719 |
| Interest Rate Decline? | 4.794 | -0.079 | -0.085 | -0.124 | -0.140 | -0.127 |
| Unemployment Rate | -0.602 | -0.044 | -0.042 | -0.188 | -0.025 | -0.028 |
| $\sigma_{Prepay}$ | 32.720 | 25.399 | 25.069 | 49.973 | 49.983 | 49.959 |
| $\sigma_{Default}$ | 15.293 | 11.498 | 11.619 | 13.454 | 13.000 | 12.968 |
| $\rho$ | 0.183 | 0.000 | 0.001 | -0.618 | -0.640 | -0.767 |
| In-sample | -4355.734 | -4169.682 | -4170.163 | -4526.357 | -4528.402 | -4507.509 |
| Hold out sample | -4325.844 | -4170.617 | -4171.315 | -4521.203 | -4532.034 | -4500.282 |

# Chapter 6

# CONCLUSIONS

"... work is never completed except by some accident such as weariness, satisfaction, the need to deliver, or death.", Paul Valery, "Recollection", *Collected Works, Volume 1.*

Threshold regression is a relatively new field of exploration with many opportunities for extension, and open questions for further investigation. Earlier work suggests that the model is particularly applicable to studies of event times where the proportional hazards assumption does not hold and/or where the degradation process itself is of interest.

In this dissertation we have both developed the theory and demonstrated implementation of important extensions of the threshold model. We offer a general framework for discrete-time threshold regression where the underyling process is a Gaussian random walk, allowing time-dependent covariates and dual correlated competing risks. In Chapters 2 and 3 we provide analytic expressions for outcome likelihoods, score functions and expectation-maximization algorithms for estimating model parameters. In Chapter 4 we sketch algorithms for computing likelihoods and scores, which we implemented in software; and we discuss various practical approaches for obtaining parameter estimates, which we also implemented in software. Chapter 5 presents results of applying the methods to both simulated data and to the Freddie Mac residential mortgage loan-level dataset.

The salient lessons are that the method performs quite well on the simulated data sets that we tested. Parameter estimates appear to be consistent and with negligible bias. Standard error coverage probabilities are very close to the desired values. The main limitation with the current implementation is computation time, particularly for models with dependent competing risks. Our attempts to estimate parameters for a competing risk model of

mortgage default and prepayment were somewhat successful, and we established that estimates for a number of parameters are qualitatively consistent with economic intuition and with estimates made by other methods. Yet some important open questions remain.

Computation time given the long observation times in the mortgage data set render it impractical to estimate parameters under the dependent risk model without making compromises. The two basic approaches for accelerating computation – (1) estimating the two event processes as independent and then estimating the correlation coefficient $\rho$, and (2) compressing the observation length by aggregating observations into longer intervals – do speed up time to convergence. But the two methods produce sufficiently different estimates for some parameters that our ability to draw conclusions from the analysis is limited.

The results with simulated data, with full (uncompressed) time series and applying assumptions of dependent vs independent processes as appropriate to the data, were sufficiently successful that we are encouraged to continue investigating these new methods. The extent to which these methods may be adopted will be constrained by computational requirements. The current implementation of these methods is practical for problems of modest size, observation length and complexity. A few examples of data sets and total estimation time might be informative, where execution times were on our fastest Amazon EC2 host. Simulated data sets with 200 subjects, a maximum observation time of 20 time periods and 2 linear covariates per dimension plus variance parameters, required less than 30 seconds when analyzed as a pair of single event processes, and about 20 minutes when analyzed under the dependent dual risk assumption. By contrast, the 1000 subject mortgage sample, with maximum observation time of 118 time points and 9 linear covariates in each dimension, took about 18 hours to estimate its parameters as a pair of single event processes. Analyzing the same data, but compressed by a factor of 6 and under the dependent competing risk assumption took 67 hours. We believe our methods are useful to data sets nearer the lower end of this range of execution times.

We suggest the following topics as next steps to build on the foundations we have established in this work and to develop discrete-time threshold regression into a practical and

broadly applicable tool.

1. **Reduction of computation time.** Exploiting parallelism as suggested in Section 4.5.6 should reduce elapsed time for obtaining estimates, albeit without reduction in required computing resources. Reducing the number of necessary CPU cycles would probably call for more sophisticated mathematics and clever hacks to reduce the number, data length and burden of the convolutions. Some preliminary ideas worth exploring:

   (a) Generalize the Shortcut Theorem of Appendix B to encompass more cases, such as where some explicit truncated probabilities have been calculated, but the probability of the final outcome is sufficiently well approximated by a function of the interim probability mass and a univariate normal to eliminate the need for additional convolutions.

   (b) There is a substantial body of research on special purpose optimizations of fast Fourier transforms in convolutions, particularly in digital signal processing. A deeper study of this literature might yield ways to simplify or eliminate calculations when it is known that the end result of a convolution will be truncated.

   In the meantime, it would be useful to gain a clearer and more formalized understanding of conditions which influence bias and efficiency loss when making parameter estimates from compressed data and when estimating dependent risk processes as if they were independent; and understanding when the bias and inefficiency costs are acceptably bounded for a given application.

2. **Measuring goodness of fit.** There is no precise equivalent of $R^2$ for censored survival data generally. Stare et al. (2011) propose a measure of explained variation, which they call $R_E$, and have implemented for Cox Proportional Hazard models and other survival methods. It would be interesting to apply $R_E$ to results from our methods.

3. **Evaluating predictive performance.** The fitted values from our threshold regression are equivalent to probabilistic forecasts conditioned on a realization of covariates. It would be interesting to evaluate these forecasts for sharpness and calibration, as per Gneiting and Katzfuss (2014). Alternatively, an extension of receiver operating characteristics (ROC) curves for competing risks, such as that of Saha and Heagerty (2010), would be another tool for evaluating predictive accuracy.

4. **Additional analysis of mortgage data.** We conclude this dissertation as we opened it, by invoking our interest in studying mortgage risk. While computational constraints have limited our results with mortgage data thus far, we have made sufficient progress to see the natural fit and applicability of these methods to analyzing mortgage risk. With improvements in computational speed, starting with parallelization, and with addtional techniques for measuring goodness of fit and predictive performance, we anticipate a practical and effective tool for analyzing mortgage behavior, and to consider richer models with additional covariates and across time periods and geographic areas.

# BIBLIOGRAPHY

Aalen, O. O. and Gjessing, H. K. (2001). Understanding the shape of the hazard rate: A process point of view (with comments and a rejoinder by the authors). *Statistical Science*, 16(1):1–22.

Bajari, P., Chu, C. S., Nekipelov, D., and Park, M. (2013). A dynamic model of subprime mortgage default: Estimation and policy implications. Technical report, National Bureau of Economic Research.

Black, F. and Cox, J. C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. *The Journal of Finance*, 31(2):351–367.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Chen, B. and Hong, Y. (2011). Testing for the Markov property in time series. *Econometric Theory*, 28(1):130.

Chen, W.-C., Ostrouchov, G., Schmidt, D., Patel, P., and Yu, H. (2014). *A Quick Guide for the pbdMPI Package (Ver. 0.2-2).* R Vignette, URL http://cran.r-project.org/package=pbdMPI.

Clapp, J. M., Deng, Y., and An, X. (2006). Unobserved heterogeneity in models of competing mortgage termination risks. *Real Estate Economics*, 34(2):243–273.

Clapp, J. M., Goldberg, G. M., Harding, J. P., and LaCour-Little, M. (2001). Movers and shuckers: interdependent prepayment decisions. *Real Estate Economics*, 29(3):411–450.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Cox, D. (1999). Some remarks on failure-times, surrogate markers, degradation, wear, and the quality of life. *Lifetime Data Analysis*, 5(4):307–314.

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

Craig, P. (2008). A new reconstruction of multivariate normal orthant probabilities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):227–243.

Craig, P. (2013). orthants (R-package). http://www.maths.dur.ac.uk/%7Edma0psc/orthants/.

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

Demyanyk, Y. and Van Hemert, O. (2011). Understanding the subprime mortgage crisis. *Review of Financial Studies*, 24(6):1848–1880.

Deng, Y., Quigley, J. M., and Van Order, R. (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica*, 68(2):275–307.

Doksum, K. A. and Hóyland, A. (1992). Models for variable-stress accelerated life testing experiments based on Wiener processes and the inverse Gaussian distribution. *Technometrics*, 34(1):pp. 74–82.

Duffie, D. (2011). *Measuring corporate default risk*. OUP Oxford.

Eaton, W. W. and Whitmore, G. (1977). Length of stay as a stochastic process: A general approach and application to hospitalization for schizophrenia. *Journal of Mathematical Sociology*, 5(2):273–292.

Frigo, M. and Johnson, S. G. (2005). The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231. Special issue on "Program Generation, Optimization, and Platform Adaptation".

Frigo, M. and Johnson, S. G. (2014). FFTW (Fastest Fourier Transform in the West) v3.3.4 software library. http://www.fftw.org/.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., and Rossi, F. (2009). *GNU Scientific Library reference manual-(3rd Edition, v1. 12)*, volume 83. Network Theory Ltd.

Genz, A. (2004). Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing*, 14(3):251–260.

Genz, A. (2013). TVPACK (Fortran code). http://www.math.wsu.edu/faculty/genz/homepage.

Genz, A. and Bretz, F. (2009). *Computation of multivariate normal and t probabilities*, volume 195. Springer.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2013). *mvtnorm: Multivariate Normal and t Distributions*. R package version 0.9-9995.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Green, J. and Shoven, J. B. (1986). The effects of interest rates on mortgage prepayments. *Journal of Money, Credit, and Banking*, 18(1):41–59.

Han, A. and Hausman, J. A. (1990). Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics*, 5(1):1–28.

Hayter, A. (2011). Recursive integration methodologies with applications to the evaluation of multivariate normal probabilities. *Journal of Statistical Theory and Practice*, 5(4):563–589.

Horrocks, J. and Thompson, M. E. (2004). Modeling event times with multiple outcomes using the Wiener process with drift. *Lifetime Data Analysis*, 10(1):29–49.

LaCour-Little, M. (2008). Mortgage termination risk: A review of the recent literature. *Journal of Real Estate Literature*, 16(3):295–326.

Lancaster, T. (1972). A stochastic model for the duration of a strike. *Journal of the Royal Statistical Society. Series A (General)*, pages 257–271.

Lee, M.-L. T., DeGruttola, V., and Schoenfeld, D. (2000). A model for markers and latent health status. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):747–762.

Lee, M.-L. T. and Whitmore, G. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*, pages 501–513.

Lee, M.-L. T. and Whitmore, G. (2010). Proportional hazards and threshold regression: their theoretical and practical connections. *Lifetime Data Analysis*, 16(2):196–214.

Lee, M.-L. T., Whitmore, G., Laden, F., Hart, J. E., and Garshick, E. (2009). A case–control study relating railroad worker mortality to diesel exhaust exposure using a threshold regression model. *Journal of Statistical Planning and Inference*, 139(5):1633–1642.

Lee, M.-L. T., Whitmore, G., and Rosner, B. (2010a). Benefits of threshold regression: a case-study comparison with Cox proportional hazards regression. In *Mathematical and Statistical Models and Methods in Reliability*, pages 359–370. Springer.

Lee, M.-L. T., Whitmore, G., and Rosner, B. A. (2010b). Threshold regression for survival data with time-varying covariates. *Statistics in Medicine*, 29(7-8):896–905.

Li, J. and Lee, M.-L. T. (2011). Analysis of failure time using threshold regression with semi-parametric varying coefficients. *Statistica Neerlandica*, 65(2):164–182.

Li, T. and Anderson, J. J. (2009). The vitality model: A way to understand population survival and demographic heterogeneity. *Theoretical Population Biology*, 76(2):118–131.

Lindqvist, B. H. and Skogsrud, G. (2008). Modeling of dependent competing risks by first passage times of Wiener processes. *IIE Transactions*, 41(1):72–80.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

McNeil, A. J., Frey, R., and Embrechts, P. (2010). *Quantitative risk management: concepts, techniques, and tools*. Princeton university press.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates*. *The Journal of Finance*, 29(2):449–470.

Metzler, A. (2008). *Multivariate first-passage models in credit risk*. PhD thesis, University of Waterloo.

Metzler, A. (2010). On the first passage problem for correlated Brownian motion. *Statistics & Probability Letters*, 80(5):277–284.

Nussbaumer, H. J. (1982). *Fast Fourier transform and convolution algorithms*. Springer.

Oakes, D. (1999). Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):479–482.

Pennell, M. L., Whitmore, G., and Lee, M.-L. T. (2010). Bayesian random-effects threshold regression with application to survival data with nonproportional hazards. *Biostatistics*, 11(1):111–126.

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*.

Ruckdeschel, P. and Kohl, M. (2014). General purpose convolution algorithm in S4 classes by means of FFT. *Journal of Statistical Software*, 59(4):1–25.

Saha, P. and Heagerty, P. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*, 66(4):999–1011.

Schaffer, M. S. (2011). Modeling mortgage market behavior: Forecasting default and voluntary prepayment from a loan-level credit perspective. *Unpublished draft, Western Michigan University Dept. of Economics*.

Smith, S. W. et al. (1997). *The scientist and engineer's guide to digital signal processing*. California Technical Pub. San Diego.

Soetaert, K. (2009). *rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations*. R package 1.6.

Stare, J., Perme, M. P., and Henderson, R. (2011). A measure of explained variation for event history data. *Biometrics*, 67(3):750–759.

Stogiannis, D. and Caroni, C. (2013). Issues in fitting inverse Gaussian first hitting time regression models for lifetime data. *Communications in Statistics-Simulation and Computation*, 42(9).

Sueyoshi, G. T. (1992). Semiparametric proportional hazards estimation of competing risks models with time-varying covariates. *Journal of Econometrics*, 51(1):25–58.

Tierney, L., Rossini, A. J., Li, N., and Sevcikova, H. (2013). *snow: Simple Network of Workstations*. R package version 0.3-13.

Varadhan, R. (2012). Squarem: Squared extrapolation methods for accelerating fixed-point iterations. R package 2012.7-1.

Watkins, J. G., Vasnev, A. L., and Gerlach, R. (2013). Multiple event incidence and duration analysis for credit data incorporating non-stochastic loan maturity. *Journal of Applied Econometrics*.

Whitmore, G. (1979). An inverse Gaussian model for labour turnover. *Journal of the Royal Statistical Society. Series A (General)*, pages 468–478.

Whitmore, G. (1983). A regression method for censored inverse-Gaussian data. *Canadian Journal of Statistics*, 11(4):305–315.

Whitmore, G. (1986). First-passage-time models for duration data: regression structures and competing risks. *The Statistician*, pages 207–219.

Whitmore, G., Crowder, M., and Lawless, J. (1998). Failure inference from a marker process based on a bivariate Wiener model. *Lifetime Data Analysis*, 4(3):229–251.

Wilhelm, S. and Manjunath, B. G. (2012). Moments calculation for the doubly truncated multivariate normal density. *arXiv preprint arXiv:1206.5387*.

Wilhelm, S. and Manjunath, B. G. (2013). *tmvtnorm: Truncated Multivariate Normal and Student t Distribution*. R package version 1.4-8.

Xu, R., McNicholas, P. D., Desmond, A. F., and Darlington, G. A. (2011). A first passage time model for long-term survivors with competing risks. *The International Journal of Biostatistics*, 7(1):1–15.

Yu, Z., Tu, W., and Lee, M.-L. T. (2009). A semi-parametric threshold regression analysis of sexually transmitted infections in adolescent women. *Statistics in Medicine*, 28(24):3029–3042.

Zhang, A. (2009). *Statistical Methods in Credit Risk Modeling*. PhD thesis, The University of Michigan.

# Appendix A

# MOMENTS OF THE LATENT PROCESS

Wilhelm and Manjunath (2012) provide formulas for computing the mean and second moments of a truncated multivariate normal, which they implement in the R `tmvtnorm` package (Wilhelm and Manjunath, 2013). We apply versions of their formulas to derive our score functions and expectation-maximization algorithms in Chapters 2 and 3. In these contexts, we wish to compute $E[\mathbf{z} - \boldsymbol{\nu}]$ and $E[(\mathbf{z} - \boldsymbol{\nu})(\mathbf{z} - \boldsymbol{\nu})']$ truncated to the region $0 < \mathbf{z} < \infty$, where $\mathbf{z} \sim MVN_Y(\boldsymbol{\nu}, \boldsymbol{\Omega})$.

## *A.1 First moments*

Wilhelm and Manjunath (2012)'s general formula for the mean of $\mathbf{z} \sim MVN_d(\boldsymbol{\nu}, \boldsymbol{\Omega})$ truncated to $\mathbf{a} \leq \mathbf{z} \leq \mathbf{b}$ immediately simplifies, in our case, to:

$$E[\mathbf{z}] = \boldsymbol{\nu} + \boldsymbol{\Omega}\mathbf{f}, \tag{A.1}$$

where $\mathbf{f}$ is the vector of univariate marginal densities of the truncated $\mathbf{z}$ at $\mathbf{0}$; i.e $f_j$ is the marginal density at 0 of the $j$th component.

$$f_j = \int_{z_Y=0}^{\infty} \cdots \int_{z_{j+1}=0}^{\infty} \int_{z_{j-1}=0}^{\infty} \cdots \int_{z_1=0}^{\infty} \phi(z_1, \ldots, z_{j-1}, 0, z_{j+1}, \ldots, z_Y) dz_1 \ldots dz_{j-1} dz_{j+1} \ldots dz_Y, \tag{A.2}$$

where $\phi(\mathbf{z})$ is the density for $\mathbf{z} \sim MVN_Y(\boldsymbol{\nu}, \boldsymbol{\Omega})$ and normalized so that $\int_{\mathbf{0}}^{\infty} \frac{1}{\psi(Y,\Delta)} \phi(\mathbf{z}) d\mathbf{z} = 1$, where $\psi(Y, \Delta)$ is the probability of observed final outcome $\Delta$ at time $Y$.

There are a few differences in applying (A.1) and computing $\mathbf{f}$ in the single event case vs. the dual event case.

### *A.1.1 Single event process*

The marginal density $f_j$ factors into independent probabilities we know how to calculate, specifically

$$\begin{aligned} f_j &= \mathrm{P}(z_j = 0 \mid W_0, \mathbf{X}, \boldsymbol{\theta}, Y, \Delta) = \frac{\mathrm{P}(Y, \Delta, z_j = 0 \mid W_0, \mathbf{X}, \boldsymbol{\theta})}{\mathrm{P}(Y, \Delta \mid W_0, \mathbf{X}, \boldsymbol{\theta})} \\ &= \frac{\mathrm{P}(z_j = 0 \mid W_0, \mathbf{X}, \boldsymbol{\theta}) \cdot \mathrm{P}(Y, \Delta \mid z_j = 0, \mathbf{X}, \boldsymbol{\theta})}{\mathrm{P}(Y, \Delta \mid W_0, \mathbf{X}, \boldsymbol{\theta})} \\ &= \frac{\mathrm{P}(z_j = 0 \mid W_0, \mathbf{X}, \boldsymbol{\theta}) \cdot \mathrm{P}\{Y - j, \Delta \mid W_0 = 0, (\nu_{j+1}, \ldots, \nu_Y)', \boldsymbol{\theta}\}}{\mathrm{P}(Y, \Delta \mid W_0, \mathbf{X}, \boldsymbol{\theta})}. \end{aligned} \tag{A.3}$$

The denominator is simply $\psi(Y, \Delta)$. We show in Section 4.3 how to obtain the left-hand factor of the numerator as a side-effect of our algorithm for computing the likelihood of event time in the denominator. The right-hand factor of the numerator is calculated for each $j$ by running the Section 4.3 algorithm to compute $\psi(Y - j, \Delta)$ where the initial position is $W_0 = 0$ and where the mean vector is the right-most $Y - j$ values of $\boldsymbol{\nu}$, i.e. $(\nu_{j+1}, \ldots, \nu_Y)'$.

Our method of computing the expectation for this specific truncated multivariate requires $O(Y^2)$ convolutions in the worst case. However it is orders of magnitude faster for this application than the more general algorithm in the current implementation of `tmvtnorm`.

### *A.1.2 Dependent competing processes*

When the random walk is bivariate Gaussian, we are interested in the univariate marginal densities of both processes, i.e the densities of the events where process $a$ touches 0 at each time $1 \leq j \leq Y$ (and where $b$ has arbitrary value) as well as the densities of the events where process $b$ touches 0 at all times $1 \leq j \leq Y$ and $a$ has arbitrary value. So for each process

$q \in (a, b)$, and where $r \in (a, b)$, $r \neq q$, we generalize (A.3), as follows:

$$f_{q;j} = \mathrm{P}(z_{q;j} = 0 \mid \mathbf{W}_0, \mathbf{X}, \boldsymbol{\theta}, Y, \Delta) = \frac{\int_{z_{r;j} \in \mathcal{U}} \mathrm{P}(Y, \Delta, z_{q;j} = 0, z_{r;j} \mid \mathbf{W}_0, \mathbf{X}, \boldsymbol{\theta}) dz_{r;j}}{\mathrm{P}(Y, \Delta \mid \mathbf{W}_0, \mathbf{X}, \boldsymbol{\theta})}$$
$$= \frac{\int_{z_{r;j} \in \mathcal{U}} \mathrm{P}(z_{q;j} = 0, z_{r;j} \mid \mathbf{W}_0, \mathbf{X}, \boldsymbol{\theta}) \cdot \mathrm{P}\{Y - j, \Delta \mid z_{q;j} = 0, z_{r;j}, (\nu_{j+1}, \ldots, \nu_Y)', \theta\} dz_{r;j}}{\mathrm{P}(Y, \Delta \mid \mathbf{W}_0, \mathbf{X}, \boldsymbol{\theta})}. \tag{A.4}$$

We compute the integral in the denominator using "density stripes", as described in Section 4.3.1, where the integration region $\mathcal{U}$ depends on $j$ and $\Delta$. Specifically, if $j < Y$, then the integral is over $z_{r;j} > 0$. If $j = Y$ then it depends on whether process $r$ survives or fails at $Y$.

### *A.2 Second moments*

Wilhelm and Manjunath (2012)'s formula for the second moment of a truncated multivariate normal entails both the univariate marginal density as above and the bivariate marginal density. Their general formula, for $\mathbf{x} \sim MVN_d(\mathbf{0}, \boldsymbol{\Omega})$, $\quad \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$, is

$$\begin{aligned} E(x_i x_j) = \omega_{ij} &+ \sum_{q=1}^{d} \omega_{iq} \frac{\omega_{jq}\{a_q g_q(a_q) - b_q g_q(b_q)\}}{\omega_{qq}} \\ &+ \sum_{q=1}^{d} \omega_{iq} \sum_{r \neq q} \left(\omega_{jr} - \frac{\omega_{qr}\omega_{jq}}{\omega_{qq}}\right) \Big[ \Big\{ g_{qr}(a_q, a_r) - g_{qr}(a_q, b_r) \Big\} \\ &- \Big\{ g_{qr}(b_q, a_r) - g_{qr}(b_q, b_r) \Big\} \Big], \end{aligned} \tag{A.5}$$

where $g_{qr}(x_q, x_r)$ is the bivariate marginal density of elements $q, r$ evaluated at the given values and $\omega_{ij}$ is the $i, j$th element of $\boldsymbol{\Omega}$.

In our expectation-maximization algorithms and score functions, we consider $E\{(\mathbf{z} - \boldsymbol{\nu})(\mathbf{z} - \boldsymbol{\nu})'\}$. Substituting $\mathbf{z} - \boldsymbol{\nu}$ for $\mathbf{x}$ in (A.5), the constraints are $\mathbf{a} = -\boldsymbol{\nu} < \mathbf{z} - \boldsymbol{\nu} < \infty = \mathbf{b}$. Also the densities $g_q(-\boldsymbol{\nu}), g_{qr}(-\boldsymbol{\nu}, -\boldsymbol{\nu})$ for $\mathbf{z} - \boldsymbol{\nu}$ are equal to the marginal densities of $\mathbf{z}$ at 0 and $(0, 0)$ respectively, which we abbreviate throughout this dissertation as $f_q, F_{qr}$. All of the above terms involving $\mathbf{b}$ vanish and the simplified expression

is

$$E[(z_i - \nu_i)(z_j - \nu_j)] = \omega_{ij} + \sum_{q=1}^{d} \omega_{iq} \Big\{ \frac{-\nu_q \omega_{jq}}{\omega_{qq}} f_q + \sum_{r \neq q} \Big( \omega_{jr} - \frac{\omega_{qr}\omega_{jq}}{\omega_{qq}} \Big) f_{qr} \Big\}.$$

In matrix notation we can write:

$$E\{(\mathbf{z} - \boldsymbol{\nu})(\mathbf{z} - \boldsymbol{\nu})'\} = \boldsymbol{\Omega} - \boldsymbol{\Omega}\mathbf{W}\text{diag}(\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega} + \boldsymbol{\Omega}\mathbf{F}\boldsymbol{\Omega} - \boldsymbol{\Omega}\text{diag}(\mathbf{F}\boldsymbol{\Omega})\text{diag}(\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}, \quad \text{(A.6)}$$

where $\mathbf{W}$ is diagonal with $\mathbf{W}_{ii} = \nu_i f_i$ and

$$\mathbf{F}_{ij} = \begin{cases} F_{ij}, & i \neq j, \\ 0, & i = j. \end{cases}$$

### *A.2.1 Single event process*

Our expectation-maximization algorithm and score function for the single event case do not require $E\{(\mathbf{z} - \boldsymbol{\nu})(\mathbf{z} - \boldsymbol{\nu})'\}]$ as such, only the trace of its product with $\boldsymbol{\Omega}^{-1}$, i.e.

$$\mathbf{tr}[\boldsymbol{\Omega}^{-1}E\{(\mathbf{z} - \boldsymbol{\nu})(\mathbf{z} - \boldsymbol{\nu})'\}] = \mathbf{tr}[\mathbf{I}_Y - \mathbf{W}\text{diag}(\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega} + \mathbf{F}\boldsymbol{\Omega} - \text{diag}(\mathbf{F}\boldsymbol{\Omega})\text{diag}(\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}],$$

the $i$th diagonal element of which is

$$1 - \mathbf{W}_{ii}\frac{1}{\boldsymbol{\Omega}_{ii}}\boldsymbol{\Omega}_{ii} + (\mathbf{F}\boldsymbol{\Omega})_{ii} - (\mathbf{F}\boldsymbol{\Omega})_{ii}\frac{1}{\boldsymbol{\Omega}_{ii}}\boldsymbol{\Omega}_{ii} = 1 - \mathbf{W}_{ii} = 1 - \nu_i f_i.$$

Thus

$$\mathbf{tr}[\boldsymbol{\Omega}^{-1}E\{(\mathbf{z} - \boldsymbol{\nu})(\mathbf{z} - \boldsymbol{\nu})'\}] = Y - \boldsymbol{\nu} \cdot \mathbf{f}. \quad \text{(A.7)}$$

Since all of the terms involving $F_{ij}$ vanish, we do not actually compute any bivariate marginal densities for the single event case.

### *A.2.2 Dependent competing processes*

The matrix of second moments, $E\{(\mathbf{z} - \boldsymbol{\nu})(\mathbf{z} - \boldsymbol{\nu})'\}]$, also appears in the derivations of the score function and expectation-maximization algorithm for the dual-event case. In those

settings we do need to compute elements of $\mathbf{F}$, but only for the upper right (equivalently, the lower left) quadrant. Specifically, we compute bivariate marginal densities only of the form

$$F_{jk} = \mathrm{P}(z_{a;j} = 0, z_{b;k} = 0 | Y, \Delta, \mathbf{X}, \boldsymbol{\theta}, \mathbf{W}_0), \tag{A.8}$$

that is to say the density of the event that the random walk touches, but does not cross, the $a$ threshold at time $j$ and the $b$ threshold at time $k$ on the way to outcome $\Delta$ at time $Y$. All terms from the upper left and lower right quadrants of $\mathbf{F}$, corresponding to events where a walk touches the same threshold at two different times, cancel from the expressions of interest and we do not calculate such terms.

Considering (A.8), where without loss of generality, $j < k$, i.e. the walk touches the $a$ threshold strictly before it touches the $b$ theshold, and letting $\mathcal{C}$ represent all of the conditioning information $\mathbf{X}, \mathbf{W}_0, \boldsymbol{\theta}$

$$\begin{aligned} F_{jk} = \mathrm{P}(z_{a;j} = 0, z_{b;k} = 0 \mid Y, \Delta, \mathcal{C}) &= \frac{\mathrm{P}(Y, \Delta, z_{a;j} = 0, z_{b;k} = 0 \mid \mathcal{C})}{\mathrm{P}(Y, \Delta \mid \mathcal{C})} \\ &= \frac{\int_{z_{b;j}} \int_{z_{a;k}} \mathrm{P}(Y, \Delta, z_{a;j} = 0, z_{b;k} = 0, z_{b;j}, z_{a;k} \mid \mathcal{C}) dz_{b;j} dz_{a;k}}{\mathrm{P}(Y, \Delta \mid \mathcal{C})}. \end{aligned}$$

The numerator expands to

$$\begin{aligned} &\int_{z_{b;j}} \int_{z_{a;k}} \mathrm{P}(Y, \Delta, z_{a;j} = 0, z_{b;k} = 0, z_{b;j}, z_{a;k}) dz_{b;j} dz_{a;k} \\ = &\int_{z_{b;j}} \int_{z_{a;k}} \mathrm{P}(z_{a;j} = 0, z_{b;j} \mid \mathcal{C}) \cdot \mathrm{P}(z_{b;k} = 0, z_{a;k} \mid z_{a;j} = 0, z_{b;j}, \mathcal{C}) \\ &\qquad \cdot \mathrm{P}(Y - k, \Delta \mid z_{b;k} = 0, z_{a;k}, (\nu_{k+1}, \ldots, \nu_Y)', \mathcal{C}) dz_{b;j} dz_{a;k}. \end{aligned}$$

Computing the double integral involves the nested "density stripes" as described in Section 4.3.1. Computation is considerably simpler for the special case

$$F_{jj} = \mathrm{P}(z_{a;j} = z_{b;j} = 0 \mid Y, \Delta, \mathcal{C}) = \frac{\mathrm{P}(z_{a;j} = z_{b;j} = 0, Y, \Delta \mid \mathcal{C})}{P(Y, \Delta \mid \mathcal{C})}.$$

Here the numerator is simplified and is similar to the univariate marginal for the single event case, (A.3):

$$F_{jj} = \frac{\mathrm{P}(z_{a;j} = z_{b;j} = 0 \mid \mathcal{C}) \cdot \mathrm{P}(Y - j, \Delta \mid \mathbf{W}_0 = \mathbf{0}, (\nu_{j+1}, \ldots, \nu_Y)', \boldsymbol{\theta})}{P(Y, \Delta \mid \mathcal{C})}.$$

# Appendix B

# SHORTCUT THEOREM

With this theorem we present conditions under which first hitting time probabilities for a Gaussian random walk with time varying drift are bounded by easily expressed probabilities of a univariate Gaussian.

Let $W$ be a Gaussian random walk with time-varying drift and, without loss of generality, unit variance. i.e; for some $W_0 > 0$ and sequence of means $\boldsymbol{\mu} = \mu_1, \mu_2, \ldots$:

$$W_t = W_{t-1} + w_t \qquad \forall t > 0,$$
$$w_t \sim N(\mu_t, 1).$$

$W$ stops upon falling below the threshold of 0. Let $T$ be the stopping time, $T = \min\{t \mid W_t < 0\}$.

Let $X$ be the non-stopping dual of $W$, i.e. it is governed by the same $W_0, \boldsymbol{\mu}$ and generating mechanism as $W$, but an instance of $X$ continues indefinitely without stopping. Then for all $t = 1, 2, \ldots,$

$$X_t \sim N(M_t, t) \qquad \text{where } M_t = W_0 + \sum_{j=1}^{t} \mu_j$$

We use *almost never* to mean with probability $< \epsilon$ for some small $\epsilon$, while *nearly always* means with probability $> 1 - \epsilon$.

An informal statement of the Shortcut Theorem is the following claims:

1. If $X$ is almost never negative at $t$, then $W$ almost never first crosses at $t$.

2. If $X$ is nearly always negative at some time strictly before $t$, then $W$ almost never either (a) first crosses at $t$, or (b) survives at $t$.

3. If $X$ is almost never negative at or before $t$ then $W$ nearly always survives at $t$.

Now to state the Shortcut Theorem more formally.

**Definitions 1.**

- A *path* is a realization of the multivariate random variables $W$ or $X$.
- $\mathcal{W}, \mathcal{X}$ are the sets of all paths of the respective random variables, $W, X$.
- $\mathcal{W}_t^-$ is the set of all paths in $\mathcal{W}$ which stop (first hit 0) at $t$. $\mathcal{W}_t^+$ is the set of all paths in $\mathcal{W}$ which are as yet unstopped at $t$, i.e. $W_t > 0$ for all $i = 1, \ldots, t$.
- $\mathcal{X}_t^-$ is the set of all paths in $\mathcal{X}$ such that $X_t < 0$. $\mathcal{X}_t^+$ is the set of all paths in $\mathcal{X}$ such that $X_t \geq 0$.
- $\psi^-(t) = P(W_t < 0)$, $\psi^+(t) = P(W_t \geq 0)$, i.e. respectively failure and survival probabilities at $t$, for $t = 0, 1, \ldots$
- $\zeta^-(t) = P(X_t < 0) = \mathbf{\Phi}(-M_t/\sqrt{t})$, $\zeta^+(t) = P(X_t \geq 0) = 1 - \mathbf{\Phi}(-M_t/\sqrt{t})$, for $t = 1, 2, \ldots$, where $\mathbf{\Phi}$ is the standard normal cumulative distribution. Define $\zeta^+(0) = 1$, $\zeta^-(0) = 0$.

**Theorem 1** ("Shortcut")**.** For $\epsilon > 0$

Claim 1: If $\zeta^-(t) < \epsilon$ then $\psi^-(t) < \epsilon$.

Claim 2: If $\zeta^+(s) < \epsilon$ for some $s < t$, then $\psi^+(t) < \epsilon$ and $\psi^-(t) < \epsilon$.

Claim 3: If $\zeta^-(s) < \epsilon/t$ for every $s \leq t$ then $\psi^+(t) > 1 - \epsilon$.

*Proof.* Plainly, $\mathcal{W}_s^- \subset \mathcal{X}_s^-$ for all $s = 0, 1, \ldots$ (For $s > 1$ it is a proper subset, since $\mathcal{X}_s^-$ can contain paths which hit 0 before $s$ and therefore are not in $\mathcal{W}_s$), therefore $\psi^-(s) \leq \zeta^-(s)$, establishing Claim 1.

Similarly, $\mathcal{W}_s^+ \subset \mathcal{X}_s^+$ so $\psi^+(s) \leq \zeta^+(s)$.

Any path of $\mathcal{W}$ that either first hits or survives at time $t$ must have survived at time $t-1$, therefore

$$\mathcal{W}_t^- \subset \mathcal{W}_{t-1}^+ \subset \mathcal{W}_{t-2}^+ \subset \cdots \subset \mathcal{W}_1^+ \subset \mathcal{W},$$
$$\mathcal{W}_t^+ \subset \mathcal{W}_{t-1}^+ \subset \mathcal{W}_{t-2}^+ \subset \cdots \subset \mathcal{W}_1^+ \subset \mathcal{W}.$$

It follows that for all $s < t$:

$$\mathcal{W}_t^- \subset \mathcal{W}_s^+ \subset \mathcal{X}_s^+,$$
$$\implies \psi^-(t) \leq \min_{s<t} \zeta^+(s).$$

Similarly,

$$\mathcal{W}_t^+ \subset \mathcal{W}_s^+ \subset \mathcal{X}_s^+,$$
$$\implies \psi^+(t) \leq \min_{s<t} \zeta^+(s),$$

establishing both parts of Claim 2.

Since every path in $\mathcal{W}$ either survives at $t$ or has crossed 0 at some time $s \in (1, 2, \ldots, t)$

$$\mathcal{W} = \mathcal{W}_t^+ + \mathcal{W}_1^- + \cdots + \mathcal{W}_t^-,$$
$$1 = \psi^+(t) + \sum_{s=1}^{t} \psi^-(t),$$
$$1 \leq \psi^+(t) + \sum_{s=1}^{t} \zeta^-(t) < \psi^+(t) + \epsilon,$$
$$\psi^+(t) \geq 1 - \epsilon,$$

establishing Claim 3. □

*Remark.* Let $s_0$ be the maximum time $0 \leq s_0 < t$ where at all times $s$ up through $s_0$, $X_s$ has negligible probability of being negative. Then we can avoid all convolutions before step $s_0 + 2$.

*Proof.* Let

$$s_0 = \max\{s < t \ \mid \ \forall\ 0 \leq s' \leq s, \quad \zeta^-(s') \leq \epsilon/t\}.$$

By Claim 3, $\psi^+(s_0) > 1 - \epsilon$, therefore $W_{s_0+1} \approx X_{s_0+1} \sim N(M_{s_0+1}, s_0 + 1)$, so we can skip the first $s_0 + 1$ convolution steps and approximate $g_{s_0+1}$ with the Gaussian density of $X_{s_0+1}$, truncated to the region above 0. The first convolution would be for $g_{s_0+2} = \phi * g_{s_0+1}$. In fact, if $s_0 = t - 1$ then $X_t \approx N(M_t, t)$, so all convolutions may be avoided and we approximate the outcome probabilities:

$$\psi(1, t) = \Phi(-M_t/\sqrt{t}),$$
$$\psi(0, t) = 1 - \Phi(-M_t/\sqrt{t}).$$

□

# VITA

Stefan Sharkansky was born in 1963 to Ira and Ina Sharkansky in Madison, Wisconsin. He spent most of his school years in Madison, and also attended schools during shorter sojourns in Athens, Georgia, Nairobi, Kenya, Melbourne, Australia and Jerusalem, Israel. He graduated from James Madison Memorial High School in Madison in 1980, and from the University of Wisconsin-Madison in 1984 with a BA in Mathematics. He earned an MS in Computer Science from Stanford University in 1987. There he was privileged to serve as a teaching assistant to pioneering computer scientists Donald E. Knuth and Robert W Floyd. After Stanford, he spent the next several years in the San Francisco Bay Area working as a software developer. He also taught the course UNIX Network Programming for the Extensions of the University of California Berkeley and Santa Cruz. In 1995 he sold the Internet chat company he started to Quarterdeck Corporation. In April 1996 *Websight* magazine named him one of the 100 most influential people on the World Wide Web. Since 1998 he has operated PersonalFund.com, which provides web-based software and services to financial advisers. In 1999 he married Irene Song. Their son David was born in 2001. They have lived in Seattle since 2003.

Stefan commenced part-time graduate studies in Statistics at the University of Washington in 2008 and completed an MS degree in 2011.

He speaks fluent but rusty Hebrew and has working knowledge of Russian and German. Outside of his academic and professional careers he enjoys hiking, skiing, cooking, reading history and classic literature and writing non-fiction essays and humorous travelogues, some of which have been published. One summer vacation while at Stanford he rode his bicycle from California to Nova Scotia. He has been a contestant on *Jeopardy!*, taking second place and winning a week in St. Thomas. Menahem Begin's car once ran over his foot.