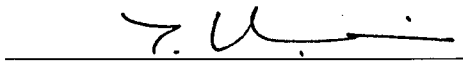
REGRESSION MODELING OF COMPETING RISKS WITH APPLICATIONS TO BONE MARROW TRANSPLANTATION STUDIES AND MORTGAGE PREPAYMENT AND DEFAULT ANALYSIS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Yuxue Jin

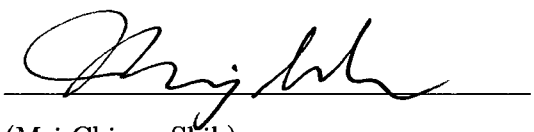
June 2009

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.



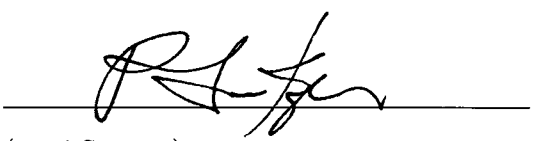
(Tze Leung Lai) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.



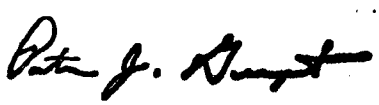
(Mei-Chiung Shih)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.



(Paul Switzer)

Approved for the University Committee on Graduate Studies.



Abstract

Competing risks frequently arise in medical applications when the subject under study may fail from more than one cause. Typically, regression models for competing risks are based on cause-specific hazards. However, the cause-specific hazard does not have a direct interpretation in terms of survival probability of a particular failure type. In recent years, many researchers have begun using the cumulative incidence function, i.e., the marginal failure probability of a particular cause, to model competing risks.

In the literature several methods have been suggested for direct regression modeling of the cumulative incidence function, such as Fine and Gray (1999), Klein and Andersen (2005) and Scheike and Zhang (2008). Fine and Gray's and Scheike and Zhang's methods require estimating the censoring distribution; hence, their performance is highly sensitive to how well the censoring distribution can be estimated. Although Klein and Andersen (2005) avoid estimating the censoring distribution by using pseudo-values, their method suffers from a loss in efficiency.

We propose an iterative maximum likelihood method to directly model the cumulative incidence function in the first part of this thesis. It involves iterating between two steps: the first step estimates the baseline subdistribution hazards using the current estimate of the regression coefficients; the second step updates the estimate of the coefficients by maximizing the log-likelihood with the baseline hazards fixed. We derive its asymptotic normality and illustrate the method on a real dataset to compare the risks of relapse and death in remission after bone marrow transplants

from different types of donors. Simulation studies show that our method outperforms the methods of Fine and Gray (1999) and Klein and Andersen (2005).

Competing risks also arise in mortgage data, which involves two mutually exclusive endpoints, prepayment and default. Many U.S. mortgages issued in recent years were made to subprime borrowers. As the house prices began to decline in mid 2006, subprime mortgage delinquencies soared, which made subprime-mortgage-backed securities almost worthless and led to a global credit crunch. A quantitative model to accurately predict the mortgage prepayment and default rates based on the loan level information and the state of the economy is therefore very important for both risk management and pricing mortgage-backed securities. In the second part of this thesis, we propose a neural network model to model the prepayment and default probabilities. We apply the model to a large dataset consisting of subprime loans originated from 2004 to 2006. Our analysis shows that the neural network model has better performance than the multilogit regression model for predicting mortgage prepayment and default rates.

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible.

My deepest gratitude goes to my advisor Professor Tze Leung Lai. Throughout my doctorate study, he provided encouragement, sound advice and good ideas. I feel very honored for having had the opportunity to work with him. I appreciate his gracious guidance in my research and choice of career path.

I would like to thank Professor Philip Lavori, Professor Paul Switzer, Professor Guenther Walther and Professor Mei-Chiung Shih for serving on my oral defense committee and giving me good suggestions for future research. Professor Philip Lavori gave me especially priceless advice which helped to improve the analysis of the bone marrow transplantation data. Special thanks go to Professor Paul Switzer and Professor Mei-Chiung Shih for carefully reading my thesis, raising stimulating questions and pointing out the typos.

I am also grateful to Ruby Wong and Jerry Halpern for providing me with the bone marrow transplantation data and sharing with me their knowledge about modeling competing risk data. I learned a lot from interacting with the BMT group.

I am indebted to David Li, Jason Jiang and Sean Chen for mentoring me during my internship at Barclays Capital and kindly letting me use the subprime mortgage data for my research. The enlightening discussions with them provided the basis of the research in the second part of my thesis.

I wish to thank my friends in the Department of Statistics for providing a stimulating and fun environment in which I learn and grow. Their friendship is very important to me.

Lastly, but most important, I want to thank my husband and my parents for their continued support through all these years. Their enduring love helped me through the difficult times. Without them I would never have been able to finish my doctoral study. To my husband Chao Li and my parents Xiuping and Guangyao Jin, I dedicate this thesis.

Contents

List of Tables

List of Figures

Part I

Direct Modeling of the Cumulative Incidence Function

Chapter 1

Introduction

1.1 Competing Risks: Cause-specific Hazard vs Cumulative Incidence Function

Problems involving competing risks are common in medical applications. In such problems there are K competing causes of failure that may occur. The occurrence of failure from one cause precludes that due to other causes. One observes for each subject a failure time and a cause of failure. An example of competing risk data is the study of different causes of death. Another example, which is quite common in cancer studies, involves relapse of the cancer as one competing risk and death in remission as another competing risk. For example, after bone marrow transplantations, patients may either die from treatment-related toxicity or relapse into leukemia. Interest is often in estimating the probability of occurrence of the competing risks, comparing these probabilities between treatment groups and modeling the effects of covariates on the probability of occurrence of the competing risks.

Competing risks are typically represented by a set of positive random variables $X_1, \ldots, X_K$, where X_j is the potential (unobservable) time to failure from the jth

cause. We observe $T^0 = \min(X_1, \ldots, X_K)$ and an indicator J which tells the cause of failure, $J = j$ if $T^0 = X_j$.

A basic quantity of interest in competing risk data is the cause-specific hazard rate $\lambda_j(t)$, also known as the crude hazard rate. It is the instantaneous probability of occurrence of failure from the jth cause in the presence of all causes of failure:

$$\lambda_j(t) = \lim_{\Delta \to 0} \frac{P(t \leq X_j \leq t + \Delta | T^0 \geq t)}{\Delta}. \tag{1.1}$$

The cause-specific hazard rate can be computed from the joint survival function of the X's, $G(x_1, \ldots, x_K) = P(X_1 \leq x_1, \ldots, X_K \leq x_K)$, as

$$\lambda_j(t) = -\frac{\partial \log G(x_1, \ldots, x_K)}{\partial x_j}, \quad \text{at } x_1 = \ldots = x_K = t. \tag{1.2}$$

When the potential failure times are independent, the cause-specific hazard rates are the same as the marginal hazard rates of the X_j's, which are defined as

$$\lim_{\Delta \to 0} \frac{1}{\Delta} P(t \leq X_j \leq t + \Delta | X_j \geq t).$$

However, this is in general not true if the potential failure times are dependent. As Cox (1959) and Tsiatis (1975) pointed out, it is also not possible to identify from competing risk data whether the X's are independent because for every dependent system of X's there is an independent set of random variables which have the same cause-specific hazard rates. This independent system of risks, however, has different marginal distributions from the original dependent set of variables.

Another quantity to summarize competing risk data is the cumulative incidence function (CIF), i.e., $F_j(t) = P(T^0 \leq t, J = j)$. It can be computed as a function of

all K cause-specific hazards by

$$F_j(t) = \int_0^t \lambda_j(u) \exp\left(-\int_0^u \sum_{h=1}^{K} \lambda_h(v)dv\right) du. \tag{1.3}$$

The cumulative incidence function represents the chance of failure from the jth cause in a world where individuals can fail from any of the causes. It is a subdistribution function with $F_j(\infty) = P(J = j)$, and depends on the cause-specific hazards of all risks.

An alternative formulation of competing risks uses a multistate model, as discussed in Andersen et al. (1992). It does not require the construction of potential failure times for each cause of failure. There are $K+1$ states a subject may be in at any point in time, denoted as state 0 to state K. State 0 is the transient state that the subject is alive. The other K states are absorbing states, each representing death from a given cause. The cause-specific hazard $\lambda_j(t)$ is the transition intensity from state 0 to state j, while the cumulative incidence function $F_j(t)$ is the transition probability from state 0 to state j, i.e., $P_{0j}(0,t) = P($ in state j at time $t|$ in state 0 at time 0).

If interest is in estimating and modeling the cause-specific hazard rates, failure from any cause other than the cause of interest can be treated as a censored observation. The usual Nelson-Aalen estimator can be applied to estimate the cumulative cause-specific hazard. Weighted log rank tests can be used to compare two or more groups. Regression analysis can also be performed under a proportional hazards assumption (Larson 1984; Prentice et al. 1978) or under Aalen's (1989) additive hazards model. Standard techniques suffice to fit the models, as the competing risks formulation does not add any additional complexity to the analysis.

While these methods are easy to apply using standard software, they are not modeling quantities of direct interest. Unlike survival data without competing risks, the exponential of the negative of the cumulative cause-specific hazard is a meaningless

quantity. Differences in the cause-specific hazards for particular risks do not translate directly into differences between the cumulative incidence functions. Although it is possible to obtain an estimate of the cumulative incidence function by combining estimates from a regression model of all the cause-specific hazards and plugging the hazards into (1.3), the resulting regression model for the CIF is a complicated nonlinear function of the covariates and is difficult to interpret.

Many authors have noted that the effects of covariates on the cumulative incidence function may be very different from those on the corresponding cause-specific hazard function. Furthermore, the cumulative incidence function is intuitively more appealing and well suited to graphical display. Therefore, it is of interest to develop regression techniques which allow for direct inference about the effects of covariates on the cumulative incidence function.

1.2 Direct Modeling of the Cumulative Incidence Function

In what follows, we use C_i to denote the censoring time, $T_i = \min(T_i^0, C_i)$ to denote the observed event or censoring time. Let Δ_i be the indicator of censoring, $\Delta_i = I_{(T_i^0 \leq C_i)}$. $N_{ij}(t)$ is the counting process of the ith subject associated with failure from the jth cause, $N_{ij} = I(T_i \leq t, \Delta_i = 1, J_i = j)$. $N_{\cdot 1}(t) = \sum_i N_{i1}(t)$ is the number of failures from the jth cause before time t. Z_i is a $p \times 1$ vector of covariates. Without loss of generality, we assume interest is in modeling the cumulative incidence function of the first cause of failure $F_1(t)$.

Previous work on the cumulative incidence function includes Aalen and Johansen (1978) and Aalen (1978), who proposed an efficient nonparametric estimate of the cumulative incidence function based on the transition intensities from a multistate

model (also see Andersen et al. 2003, pages 296-297):

$$\widehat{F}_j(t) = \int_0^t \prod_{T_i<u} \left(1 - \frac{\sum_{h=1}^{K} dN_{\cdot h}(T_i)}{Y(T_i)}\right) \frac{dN_{\cdot 1}(u)}{Y(u)}, \tag{1.4}$$

where $Y(t)$ is the number at risk at time t. Gray (1988) and Pepe (1991) developed tests for equality of the CIFs across treatment groups. Their methods are useful but restricted to data with discrete covariates.

Recently some work has been done on direct regression modeling of the cumulative incidence function. Most of this work assumes semi-parametric models, one baseline for each risk. The difficulty of estimating the semi-parametric models when the data are censored is that the parametric and nonparametric parameters have to be estimated simultaneously, since the partial likelihood principle no longer works in this case. The existing methods basically fall into two categories depending on how censored data is handled. Methods in the first category, such as Fine and Gray (1999), Fine (2001) and Scheike and Zhang (2008), use the inverse probability of censoring weighting (IPCW) technique (Robins and Rotnitzky 1992) to get around censored data. An alternative way to handle censored data, proposed by Klein and Andersen (2005), is based on jack-knife pseudo-values of the nonparametric estimate of the cumulative incidence function. In what follows we summarize these methods which model the CIF directly.

1.2.1 Proportional Subdistribution Hazard Model

Fine and Gray (1999) proposed a proportional subdistribution hazard model and extended the partial likelihood principle to competing risk data by modifying the risk

sets. The subdistribution hazard is defined as

$$a_1(t) = \frac{dF_1(t)/dt}{1 - F_1(t)}. \tag{1.5}$$

Under the proportional hazards assumption,

$$a_1(t; Z) = a_{10}(t) \exp(Z^T(t)\beta). \tag{1.6}$$

If the censoring times C_i's are known, in the absence of covariates it can be estimated by

$$\widehat{a}_1(t) = \frac{dN_{\cdot 1}(t)}{\sum_i I(C_i \geq t, N_{i1}(t) = 0)}. \tag{1.7}$$

It is easy to show that $1 - \exp(-\int_0^t \widehat{a}_1(s)ds)$ is asymptotically equivalent to the nonparametric maximum likelihood estimate of $F_1(t)$ given in (1.4).

The risk set at time t

$$\sum_i I(C_i \geq t, N_{i1}(t) = 0) \tag{1.8}$$

includes two distinct groups, the subjects who have not failed from any cause prior to time t and the subjects who have failed previously from other competing causes and are not censored at time t. If the censoring times are unknown, a subject who failed prior to time t due to other competing risks is weighted by $w_i(t)$, the probability that it is not censored at time t given that it was not censored at the failure time T_i:

$$w_i(t) = \widehat{G}(t)/\widehat{G}(T_i \wedge t), \tag{1.9}$$

where $\widehat{G}(t)$ is the Kaplan-Meier estimate of $P(C > t)$.

The unconventional risk set (1.8) leads to a partial likelihood for the subdistribution function $F_1(t; Z)$. Estimates of the regression coefficients can be obtained by

solving the estimating equation:

$$U_1(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \left(Z_i(s) - \frac{\sum_j w_j(s) Y_{j1}(s) Z_j(s) \exp(Z_j^T(s)\beta)}{\sum_j w_j(s) Y_{j1}(s) \exp(Z_j^T(s)\beta)} \right) \times w_i(s) dN_{i1}(s), \tag{1.10}$$

where $Y_{i1}(t) = (1 - N_{i1}(t))I(C_i \geq T_i^0 \wedge t)$.

1.2.2 Rank-Based Least Squares Regression

Fine (2001) proposed a semi-parametric transformation model of the CIF for time-independent covariates. Estimation of the regression coefficients is achieved with a rank-based least square criterion based on the correlated Bernoulli variables $u_{ij} = I\{\min(T_i^*, t_0) \geq T_j^*\}, 1 \leq i \neq j \leq n$, where $T_i^* = T_i^0 \times I(J_i = 1) + \infty \times I(J_i \neq 1)$. $t_0 < \infty$ is chosen so that $P(\min(T^0, C) > t_0) > 0$.

Since T_i^* is unknown for censored data, the IPCW technique is again adapted to adjust for missing data. Define

$$s_{1,ij} = I(\min(X_i, t_0) \geq X_j, J_j = 1)\Delta_j(\widehat{G}(X_j, Z_i)\widehat{G}(X_j, Z_j))^{-1},$$

$$s_{2,ij} = I(t_0 \geq X_j \geq X_i, J_j = 1, J_i \neq 1)\Delta_i\Delta_j(\widehat{G}(X_i, Z_i)\widehat{G}(X_j, Z_j))^{-1},$$

where $\widehat{G}(t, Z)$ is an estimate of the censoring distribution conditioned on the covariates. Let $\widehat{u}_{ij} = s_{1,ij} + s_{2,ij}$, it is easy to see that $\widehat{u}_{ij}$ is unbiased for u_{ij} conditional on the covairates.

Assume the existence of a known, differentiable function $g(\cdot)$, such that

$$g(F_1(t; Z)) = h(t) - Z^T\beta.$$

$h(t)$ determines the baseline failure probability when $Z = 0$; it is unspecified, invertible and strictly increasing in t. Possible choices of $g(\cdot)$ are $g(x) = \log(x/(1-x))$

and $g(x) = \log(-\log(1-x))$. The conditional expectation of u_{ij} is $\eta_{ij}(\alpha,\beta) = \int_{-\infty}^{\alpha}(1-g^{-1}(t-Z_i^T\beta))d\,g^{-1}(t-Z_j^T\beta)$ evaluated at the true values $\alpha = h(t_0)$ and β. A reasonable estimate of $\theta = (\alpha,\beta)$ is the minimizer of the squared error criterion $Q(\theta) = \sum_{i\neq j}(\hat{u}_{ij} - \eta_{ij}(\theta))^2$. A weighted version of the squared error criterion is also considered.

1.2.3 Direct Binomial Regression

A direct binomial regression method to model the CIF has been proposed in Scheike and Zhang (2008). $\frac{N_{i1}(t)}{G(T_i)}$ is unbiased for $F_1(t,Z_i)$ conditional on the covariate Z_i, since

$$E\left(\frac{N_{i1}(t)}{G(T_i)}|Z_i\right) = \int_0^\infty \frac{I(s\leq t)}{G(s)}G(s)dF_1(s,Z_i) = F_1(t,Z_i).$$

$\frac{N_{i1}(t)}{\widehat{G}(T_i)}$, therefore, can be used as the response in a generalized estimating equation. Suppose $F_1(t,Z_i)$ is modeled with a link function and a parameter vector η, η can be estimated by finding the zero of the score function

$$U(t,\eta(t)) = \sum_i D_i(t,\eta(t))w_i(t)\left(\frac{N_{i1}(t)}{\widehat{G}(T_i)} - F_1(t,Z_i)\right),$$

where $w_i(t)$ is the weight and $D_i(t,\eta(t))$ is the derivative of $F_1(t,Z_i)$ with respect to η at time t.

1.2.4 Pseudo-value Approach

Another recent method for direct regression of the CIF is the pseudo-value approach proposed by Klein and Andersen (2005). It is based on jackknife pseudo-values constructed from the Aalen-Johansen (AJ) nonparametric estimate (1.4). For a fixed grid of time points $\tau_1,\ldots,\tau_M$, the jackknife pseudo-value for the ith subject at time

τ_h is

$$\widehat{\theta}_{ih} = n\widehat{F}_1(\tau_h) - (n-1)\widehat{F}_1^{(i)}(\tau_h), \qquad 1 \leq h \leq M, \tag{1.11}$$

where $\widehat{F}_1(\tau_h)$ is the AJ estimator based on the complete data set and $\widehat{F}_1^{(i)}(\tau_h)$ is the estimator based on the sample of size $n-1$ leaving out the ith subject.

When there is no censoring,

$$\widehat{\theta}_i = (\widehat{\theta}_{i1}, \ldots, \widehat{\theta}_{iM})^T = (I(T_i \leq \tau_1, J_i = 1), \ldots, I(T_i \leq \tau_M, J_i = 1))^T.$$

Under independent right censoring, the AJ estimator can be expressed in terms of influence curves as in van der Laan and Robins (2003):

$$\widehat{F}_1(t) - F_1(t) = \frac{1}{n}\sum_{i=1}^{n} IC(T_i, \Delta_i, J_i; t) + o_p(n^{-1/2}).$$

Graw et al. (2008) have shown that

$$\widehat{\theta}_{ih} = IC(T_i, \Delta_i, J_i; \tau_h) + F_1(\tau_h) + o_p(n^{-1/2}).$$

Therefore the pseudo-values $\widehat{\theta}_i$'s are approximately independent and asymptotically unbiased for the cumulative incidence function conditioned on the covariates.

These pseudo-values are used in a generalized linear model to model the effects of covariates. Let $g(\cdot)$ be a link function; possible choices are the logit link $g(x) = \log(x/(1-x))$, or the complementary log-log link $g(x) = \log(-\log(1-x))$, which gives the proportional hazards representation as in Fine and Gray (1999). We assume

$$g(\theta_{ih}) = \alpha_h + Z_i^T\gamma = Z_{ih}^T\beta, \quad i = 1, \ldots, n, \quad h = 1, \ldots, M.$$

Define the inverse link by

$$\theta_{ih} = g^{-1}(Z_{ih}^T\beta) = \mu(Z_{ih}^T\beta).$$

Since pseudo-values for the same subject at different time points are correlated, the generalized estimating equation approach of Liang and Zeger (1986) is applied to estimate β. Let $\theta_i = (\theta_{i1}, \ldots, \theta_{iM})^T$, $d\mu_i(\beta)$ be the $(M + p) \times M$ matrix of partial derivatives of θ_i with respect to the parameters, $V_i(\beta)$ be a working covariance matrix. The estimating equations to be solved are

$$U(\beta) = \sum_i d\mu_i(\beta)V_i^{-1}(\beta)(\hat{\theta}_i - \theta_i). \tag{1.12}$$

Klein and Andersen (2005) suggested three possible choices of the working covariance matrix: the identity matrix, an "exact" working covariance matrix computed by the covariance between Bernoulli variables, and the usual sample covariance matrix of the pseudo-values. They "found no systematic differences between different choices of working GEE covariance" and "recommend that the simple independence working covariance matrix be used".

1.3 Outline

The methods that use the inverse probability of censoring weighting technique all require estimating the censoring distribution. Their performance is very sensitive to how well the censoring distribution can be estimated. Although the pseudo-value approach does not require estimating the censoring distribution, the pseudo-values for the same subject at different time points are highly correlated. Hence the generalized estimating equation approach is inefficient in this case; see Lai and Small (2007).

To get an efficient estimate and avoid estimating the censoring distribution, we

propose to maximize the likelihood iteratively. We assume a proportional subdistribution hazard model as in Fine and Gray (1999). The algorithm consists of iterating between two steps. The first step is to estimate the baseline subdistribution hazards by the Nelson-Aalen type estimates assuming the regression parameters are known; the second step is to update the regression parameters by maximizing the log-likelihood with the baseline hazards fixed.

In Chapter 2, we describe the algorithm and its relationship to the Cox proportional hazards model for survival data without competing risks. In Chapter 3, using similar techniques as in Lai and Ying (1988, 1991), we express the estimate as the solution to a system of estimating equations and derive its asymptotic normality. Simulation studies comparing the suggested iterative maximum likelihood method with the methods of Fine and Gray (1999) and Klein and Andersen (2005) are presented in Chapter 4. In Chapter 5, the iterative maximum likelihood method is applied to analyze data from a bone marrow transplantation study using different types of donors. In Chapter 6, we summarize the merits of the suggested iterative maximum likelihood method.

Chapter 2

Iterative Maximum Likelihood Method

2.1 Model and Algorithm

Let T^0 and C be the failure time and the censoring time, $J \in (1, \ldots, K)$ be the cause of failure. Without loss of generality, we assume there are two causes of failure, i.e. $K = 2$. Z is a $p \times 1$ vector of covariates. The algorithm and proof in this chapter and Chapter 3 are not restricted to time-independent covariates, but for simplicity of notation, we assume Z to be independent on t. In section 3.5, we briefly discuss how to extend the method to situations with time-varying covariates. Under independent right censoring, $T = \min(T^0, C)$, $\Delta = I_{(T^0 \leq C)}$, observe $\{T_i, \Delta_i, \Delta_i J_i, Z_i\}$ for $i = 1, \ldots, n$. Our interest is in modeling the cumulative incidence functions for the two causes of failure conditioned on the covariates, i.e., $F_j(t, Z) = P(T^0 \leq t, J = j|Z), j = 1, 2$.

We assume a proportional subdistribution hazard model with different regression parameters and baselines for different causes of failure. The cumulative incidence

functions are modelled as

$$F_j(t, Z) = 1 - \exp(-\exp(Z^T\beta_j)A_j(t)), j = 1, 2, \tag{2.1}$$

where $A_j(t), j = 1, 2$ are nonparametric baselines, $\beta_j, j = 1, 2$ are $p \times 1$ regression coefficients. For convenience, we use bold letters to denote vectors which concatenate the corresponding quantities of the two risks; for example, $\boldsymbol{\beta} = (\beta_1^T, \beta_2^T)^T$ is a $2p \times 1$ vector representing the regression parameters in the two competing risks. Assuming $A_i(t), j = 1, 2$ are absolutely continuous with $dA_j(t) = a_j(t)dt$, the likelihood of the data on the interval $[0, \tau]$ under independent right censoring is

$$\prod_i \prod_t d\Lambda_{i1}(t)^{dN_{i1}(t)} d\Lambda_{i2}(t)^{dN_{i2}(t)} \exp(-\Lambda_{i1}(\tau) - \Lambda_{i2}(\tau)), \tag{2.2}$$

where $N_{ij}(t) = I_{(T_i \le t, \Delta_i = 1, J_i = j)}$ is the counting process of the ith subject associated with failures from cause j, $\Lambda_{ij}(t)$ is its compensator, $Y_i(s) = I_{(T_i \ge s)}$,

$$d\Lambda_{ij}(t) = \frac{Y_i(t)(1 - F_j(t, Z_i)) \exp(Z_i^T\beta_j)a_j(t)dt}{1 - F_1(t, Z_i) - F_2(t, Z_i)}.$$

When there is censored data, the partial likelihood principle is no longer applicable. We suggest to maximize the likelihood in an iterative way which consists of two steps.

The first step is to estimate the baselines using Nelson-Aalen type estimates, assuming the regression coefficients are known. If the regression coefficients are $\mathbf{b} = (b_1^T, b_2^T)^T$, $A_j(t)$ is estimated by $\widehat{A}_{nj}(t, \mathbf{b}), j = 1, 2$.

$$\widehat{A}_{nj}(t, \mathbf{b}) = \int_0^t \frac{\sum_i dN_{ij}(s)}{\widehat{S}_{nj}(s-, \mathbf{b})}, \tag{2.3}$$

where

$$\widehat{S}_{nj}(s-, \mathbf{b}) = \sum_{i=1}^{n} \frac{Y_i(s)(1 - \widehat{F}_{nj}(s-, \mathbf{b}, Z_i)) \exp(Z_i^T b_j)}{1 - \widehat{F}_{n1}(s-, \mathbf{b}, Z_i) - \widehat{F}_{n2}(s-, \mathbf{b}, Z_i)}, \tag{2.4}$$

$$\widehat{F}_{nj}(s, \mathbf{b}, Z_i) = 1 - \exp\big(-\exp(Z_i^T b_j)\widehat{A}_{nj}(s, \mathbf{b})\big), \quad j = 1, 2. \tag{2.5}$$

The second step is to update estimates of the regression coefficients by maximizing the log-likelihood with the baselines fixed. The new estimates of the regression coefficients $\mathbf{b} = (b_1^T, b_2^T)^T$ are

$$\begin{aligned}(b_1, b_2) = \arg\max_{b_1, b_2} & \sum_i \sum_{j=1}^{2} I_{(\Delta_i=1, J_i=j)}(Z_i^T b_j - \exp(Z_i^T b_j)\widehat{A}_{nj}(T_i-)) \\ & - \sum_i I_{(\Delta_i=1)} \log(1 - \widehat{F}_{n1}(T_i, \mathbf{b}, Z_i) - \widehat{F}_{n2}(T_i, \mathbf{b}, Z_i)) \\ & - \sum_i \sum_{t \le T_i} \frac{\sum_j (1 - \widehat{F}_{nj}(t, \mathbf{b}, Z_i)) \exp(Z_i^T b_j) d\widehat{A}_{nj}(t, \mathbf{b})}{1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i)}.\end{aligned} \tag{2.6}$$

$\widehat{A}_{nj}(t, \mathbf{b})$ are held to be constant in the above equation. Up to first order,

$$-\frac{\sum_j (1 - \widehat{F}_{nj}(t, \mathbf{b}, Z_i)) \exp(Z_i^T b_j) d\widehat{A}_{nj}(t, \mathbf{b})}{1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i)} = d\log(1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i)). \tag{2.7}$$

Hence the last term in (2.6) is up to first order equal to

$$\sum_i \log(1 - \widehat{F}_{n1}(T_i, \mathbf{b}, Z_i) - \widehat{F}_{n2}(T_i, \mathbf{b}, Z_i)). \tag{2.8}$$

Thus (2.6) is up to first order equal to

$$\begin{aligned}(b_1, b_2) = \arg\max_{b_1, b_2} & \sum_i \sum_{j=1}^{2} I_{(\Delta_i=1, J_i=j)}(Z_i^T b_j - \exp(Z_i^T b_j)\widehat{A}_{nj}(T_i-)) \\ & + \sum_i I_{(\Delta_i=0)} \log(1 - \widehat{F}_{n1}(T_i, \mathbf{b}, Z_i) - \widehat{F}_{n2}(T_i, \mathbf{b}, Z_i)).\end{aligned} \tag{2.9}$$

However, the first-order approximation in (2.9) leads to a much poorer performance in our simulation studies than (2.6); hence, the method is implemented with the

likelihood represented in (2.6).

(2.6) can also be written as a stochastic integral:

$$(b_1, b_2) = \arg\max_{b_1,b_2} \sum_i \sum_{j=1}^{2} \int_0^\tau \log(d\widehat{\Lambda}_{ij}(t)) dN_{ij}(t) - \int_0^\tau d\widehat{\Lambda}_{ij}(t), \tag{2.10}$$

where

$$\begin{aligned} d\widehat{\Lambda}_{ij,n}(t, \mathbf{b}) &= \frac{Y_i(t)(1 - \widehat{F}_{nj}(t, \mathbf{b}, Z_i)) \exp(Z_i^T b_j) d\widehat{A}_{nj}(t, \mathbf{b})}{1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i)} \\ &= \frac{Y_i(t) d\widehat{F}_{nj}(t, \mathbf{b}, Z_i)}{1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i)}. \end{aligned} \tag{2.11}$$

The final estimates of the regression coefficients and baselines are obtained after iterating between these two steps until convergence. In the absence of covariates, it is easy to show that the Nelson-Aalen type estimate is asymptotically equivalent to the nonparametric maximum likelihood estimate of the baseline. Therefore, the final estimates of the regression coefficients and baselines fully maximize the log-likelihood.

Next we will express the final estimate of the coefficients as a solution to a system of estimating equations, which is the basis of the argument in Chapter 3. If $\widehat{A}_{nj}(t, \mathbf{b}), j = 1, 2$ are taken to be constant when taking derivatives with regard to $\mathbf{b}$, define

$$\begin{aligned} \widehat{\eta}_{ij,n}(t, \mathbf{b}) &= \frac{\partial}{\partial b_j} \log d\widehat{\Lambda}_{ij,n}(t, \mathbf{b}) \\ &= Z_i - Z_i \exp(Z_i^T b_j) \widehat{A}_{nj}(t, \mathbf{b}) + \frac{(1 - \widehat{F}_{nj}(t, \mathbf{b}, Z_i)) \exp(Z_i^T b_j) Z_i \widehat{A}_{nj}(t, \mathbf{b})}{1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i)}, \end{aligned} \tag{2.12}$$

$$\begin{aligned}\widehat{\phi}_{ij,n}(t,\mathbf{b}) &= \frac{\partial}{\partial b_j}\log d\widehat{\Lambda}_{ij',n}(t,\mathbf{b}) \\ &= \frac{(1-\widehat{F}_{nj}(t,\mathbf{b},Z_i))\exp(Z_i^T b_j)Z_i\widehat{A}_{nj}(t,\mathbf{b})}{1-\widehat{F}_{n1}(t,\mathbf{b},Z_i)-\widehat{F}_{n2}(t,\mathbf{b},Z_i)},\end{aligned} \tag{2.13}$$

where

$$j' = \begin{cases} 1, & j=2, \\ 2, & j=1. \end{cases}$$

The estimate of the coefficients $\widehat{\boldsymbol{\beta}}_n = (\widehat{\beta}_{n1}^T, \widehat{\beta}_{n2}^T)^T$ solves the following estimating equations, $j = 1, 2$

$$\begin{aligned}U_{nj}(\mathbf{b}) &= \sum_i \int_0^\tau \widehat{\eta}_{ij,n}(t,\mathbf{b})(dN_{ij}(t) - d\widehat{\Lambda}_{ij}(t,\mathbf{b})) \\ &\quad + \sum_i \int_0^\tau \widehat{\phi}_{ij,n}(t,\mathbf{b})(dN_{ij'}(t) - d\widehat{\Lambda}_{ij'}(t,\mathbf{b})),\end{aligned} \tag{2.14}$$

2.2 Relationship to the Cox Regression in Survival Data Without Competing Risks

If there is only one cause of failure, the iterative maximum likelihood method solves the same estimating equation as the Cox regression. The Cox regression solves the following estimating equation (see Andersen et al. 1992, page 486):

$$U(b) = \sum_i \int_0^\tau (Z_i(t) - E(t,b))dN_i(t), \tag{2.15}$$

where

$$E(t,b) = \frac{\sum_i Z_i(t)\exp(Z_i^T(t)b)Y_i(t)}{\sum_i \exp(Z_i^T(t)b)Y_i(t)}.$$

Since $\sum_i Z_i(t)\exp(Z_i^T(t)b)Y_i(t) = E(t,b)\sum_i \exp(Z_i^T(t)b)Y_i(t)$, the score function (2.15) can be reexpressed as

$$U(b) = \sum_i \int_0^\tau (Z_i(t) - E(t,b))(dN_i(t) - \exp(Z_i^T(t)b)Y_i(t)d\widehat{A}(t,b)), \qquad (2.16)$$

where $\widehat{A}(t,b)$ is the Nelson-Aalen estimate of the baseline defined by

$$\widehat{A}(t,b) = \int_0^t \frac{\sum_i dN_i(s)}{\sum_i \exp(Z_i^T(t)b)Y_i(s)}.$$

From the fact that $\sum_i dN_i(s) = \sum_i \exp(Z_i^T(t)b)Y_i(s)d\widehat{A}(s,b)$, the score function (2.16) can be simplified as

$$U(b) = \sum_i \int_0^\tau Z_i(t)(dN_i(t) - \exp(Z_i^T(t)b)Y_i(t)d\widehat{A}(t,b)). \qquad (2.17)$$

This is the same as the score function (2.14) of the iterative maximum likelihood method if there is no competing risk. Therefore the suggested iterative maximum likelihood method can be considered as an extension of the Cox regression to situations where the estimation of the regression parameters and nonparametric baselines cannot be separated.

Chapter 3

Asymptotic Normality

In this chapter we derive the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$. For simplicity, we only consider data observed in a finite time interval $[0, \tau]$ and estimates in a small neighborhood of the true parameters $\{\mathbf{b} : |\mathbf{b} - \boldsymbol{\beta}| < \rho\}$. The technical assumptions made in this chapter are collected and given below for easy reference:

A 1. $|Z_i| \leq B$ *for all* i *and some nonrandom constant* B,

A 2. *The limits in* (3.11) *and* (3.12) *exist in probability, hence* $S_j^0(t)$ *and* $h_j(t, \mathbf{b}, \widetilde{A}_1, \widetilde{A}_2)$, $j = 1, 2$, $0 \leq t \leq \tau$ *are well defined.*

A 3. $h_j(t, \mathbf{b}, \widetilde{A}_1, \widetilde{A}_2)$, $j = 1, 2$, $0 \leq t \leq \tau$ *are bounded away from zero.*

A 4. $1/h_j(t, \mathbf{b}, \widetilde{A}_1, \widetilde{A}_2)$, $j = 1, 2$, $0 \leq t \leq \tau$, *are Lipschitz continuous in* A_1 *and* A_2.

A 5. $1/h_j(t, \mathbf{b}, \widetilde{A}_1, \widetilde{A}_2)$, $j = 1, 2$, $0 \leq t \leq \tau$, *are continuously differentiable in* $\mathbf{b}$, $\widetilde{A}_1$ *and* $\widetilde{A}_2$.

A 6. *The limits defined in* (3.36) *exist in probability.*

A 7. *The limits defined in* (3.37)-(3.45) *exist in probability.*

A 8. Σ *defined in* (3.36) *is nonsingular.*

$1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i)$ frequently appears in the denominators in (2.4), (2.12) and (2.13). To bound it away from zero without losing efficiency, we apply similar techniques in Lai and Ying (1988, 1991) and introduce a smooth weight function of the form

$$p_n(x) = p(n^{\alpha}(x - cn^{-\alpha})), \tag{3.1}$$

with $c > 0$, $0 < \alpha < 1$, and p being a twice-continuously differentiable and nondecreasing function on the real line such that

$$p(u) = 0 \quad \text{for } u \leq 0, \qquad p(u) = 1 \quad \text{for } u \geq 1. \tag{3.2}$$

The idea is to multiply $1/(1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i))$ by a smooth weight function

$$\begin{aligned} \widehat{p}_n(t, \mathbf{b}, Z_i) &= p_n\left(1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i)\right) \\ &= \begin{cases} 1 & \text{if } 1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i) \geq (c+1)n^{-\alpha}, \\ 0 & \text{if } 1 - \widehat{F}_{n1}(t, \mathbf{b}, Z_i) - \widehat{F}_{n2}(t, \mathbf{b}, Z_i) \leq cn^{-\alpha}. \end{cases} \end{aligned} \tag{3.3}$$

For simplicity of notation, also define

$$\widetilde{p}_n(t, \mathbf{b}, Z_i) = p_n(1 - \widetilde{F}_1(t, \mathbf{b}, Z_i) - \widetilde{F}_2(t, \mathbf{b}, Z_i)), \tag{3.4}$$

$$p_n(t, Z_i) = p_n(1 - F_1(t, Z_i) - F_2(t, Z_i)), \tag{3.5}$$

where $\widetilde{F}$ will be defined in (3.13). The following modifications are made to (2.3), (2.4), (2.12) and (2.13).

$$\widehat{A}_{nj}(t, \mathbf{b}) = \int_0^t \frac{\sum_i \widehat{p}_n(s-, \mathbf{b}, Z_i) dN_{ij}(s)}{\widehat{S}_{nj}(s-, \mathbf{b})}, \tag{3.6}$$

$$\widehat{S}_{nj}(s-,\mathbf{b}) = \sum_{i=1}^{n} \widehat{p}_n(s-,\mathbf{b},Z_i)\frac{Y_i(s)(1-\widehat{F}_{nj}(s-,\mathbf{b},Z_i))\exp(Z_i^T b_j)}{1-\widehat{F}_{n1}(s-,\mathbf{b},Z_i)-\widehat{F}_{n2}(s-,\mathbf{b},Z_i)}, \tag{3.7}$$

$$\widehat{\eta}_{ij,n}(t,\mathbf{b}) = \widehat{p}_n(t,\mathbf{b},Z_i)\left(Z_i - Z_i\exp(Z_i^T b_j)\widehat{A}_{nj}(t,\mathbf{b}) + \frac{(1-\widehat{F}_{nj}(t,\mathbf{b},Z_i))\exp(Z_i^T b_j)Z_i\widehat{A}_{nj}(t,\mathbf{b})}{1-\widehat{F}_{n1}(t,\mathbf{b},Z_i)-\widehat{F}_{n2}(t,\mathbf{b},Z_i)}\right), \tag{3.8}$$

$$\widehat{\phi}_{ij,n}(t,\mathbf{b}) = \frac{\widehat{p}_n(t,\mathbf{b},Z_i)(1-\widehat{F}_{nj}(t,\mathbf{b},Z_i))\exp(Z_i^T b_j)Z_i\widehat{A}_{nj}(t,\mathbf{b})}{1-\widehat{F}_{n1}(t,\mathbf{b},Z_i)-\widehat{F}_{n2}(t,\mathbf{b},Z_i)}. \tag{3.9}$$

Let $U_n(\mathbf{b}) = (U_{n1}^T(\mathbf{b}), U_{n2}^T(\mathbf{b}))^T$, we establish in section 3.1 the asymptotic equivalent of $U_n(\mathbf{b})$ and consistency of $\widehat{\boldsymbol{\beta}}_n$. In section 3.2, we establish the asymptotic linearity of $U_n(\mathbf{b})$ in some neighborhood of $\boldsymbol{\beta}$. In section 3.3, the asymptotic normality of $U_n(\boldsymbol{\beta})$ is derived. Finally in section 3.4, we derive the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$.

3.1 Consistency of $\widehat{\boldsymbol{\beta}}_n$

Approximate $\widehat{A}_1(t,\mathbf{b})$ with a nonrandom function $\widetilde{A}_1(t,\mathbf{b})$, which is defined as

$$\widetilde{A}_1(t,\mathbf{b}) = \int_0^t \frac{S_1^0(s)a_1(s)ds}{h_1(s,\mathbf{b},\widetilde{A}_1(s,\mathbf{b}),\widetilde{A}_2(s,\mathbf{b}))}, \tag{3.10}$$

where

$$S_1^0(t) = \lim_{n\to\infty}\frac{1}{n}\sum_i \frac{p_n(t,Z_i)Y_i(t)(1-F_1(t,Z_i))\exp(Z_i^T\beta_1)}{1-F_1(t,Z_i)-F_2(t,Z_i)}, \tag{3.11}$$

$$\begin{aligned} & h_1(t,\mathbf{b},\widetilde{A}_1(t,\mathbf{b}),\widetilde{A}_2(t,\mathbf{b})) \\ = & \lim_{n\to\infty}\frac{1}{n}\sum_i \frac{\widetilde{p}_n(t,\mathbf{b},Z_i)Y_i(t)(1-\widetilde{F}_1(t,\mathbf{b},z))\exp(Z_i^T b_1)}{1-\widetilde{F}_1(t,\mathbf{b},z)-\widetilde{F}_2(t,\mathbf{b},z)}, \end{aligned} \tag{3.12}$$

$$\widetilde{F}_j(s,\mathbf{b},z) = 1 - \exp\big(-\exp(z^T b_j)\widetilde{A}_j(s,\mathbf{b})\big), \quad j = 1,2. \tag{3.13}$$

In similar fashion we define $\widetilde{A}_2$:

$$\widetilde{A}_2(t,\mathbf{b}) = \int_0^t \frac{S_2^0(s)a_2(s)ds}{h_2(s,\mathbf{b},\widetilde{A}_1(s,\mathbf{b}),\widetilde{A}_2(s,\mathbf{b}))}. \tag{3.14}$$

Under assumption A4, $1/h_j(s,\mathbf{b},\widetilde{A}_1,\widetilde{A}_2)$, $j = 1,2$ are Lipschitz continuous in $\widetilde{A}_1$ and $\widetilde{A}_2$. By Picard's Existence Theorem (Coddington and Levinson 1955), the solution to the integral equations (3.10) and (3.14) exists and is unique. Thus $\widetilde{A}_j(t,\mathbf{b}), j = 1,2$ are well defined. In this subsection, for simplicity of notation, we denote $h_j(t,\mathbf{b},\widetilde{A}_1(t,\mathbf{b}),\widetilde{A}_2(t,\mathbf{b}))$ by $h_j(t,\mathbf{b})$.

Lemma 3.1. *Under assumptions A1-A4, for $\alpha < 1/6$ and ρ_n such that*

$$\sup_{0<t<\tau,|\mathbf{b}-\beta|\le\rho_n,z} |\widetilde{p}_n(t,\mathbf{b},z) - \widetilde{p}_n(t,\boldsymbol{\beta},z)| = O(n^{-1/2}), \tag{3.15}$$

the following holds

$$\sup_{0<t<\tau,|\mathbf{b}-\beta|\le\rho_n} |\widehat{A}_{nj}(t,\mathbf{b}) - \widetilde{A}_j(t,\mathbf{b})| = O_p(n^{-1/2}), \quad j = 1,2. \tag{3.16}$$

Proof. The proof is based on mathematical induction. Chop the interval $[0,\tau]$ into $0 = t_0 \le t_1 \ldots \le t_m = \tau$ which is fine enough so that for $1 \le k \le m$,

$$\sup_{s,s'\in(t_{k-1},t_k],|\mathbf{b}-\beta|\le\rho_n} |h_1(s,\mathbf{b}) - h_1(s',\mathbf{b})| = O(n^{-1/2}), \tag{3.17}$$

$$\sup_{s,s'\in(t_{k-1},t_k],z} |p_n(s,z) - p_n(s',z)| = O(n^{-1/2}), \tag{3.18}$$

and the counting processes $N_{ij}(t)$, $1 \leq i \leq n, j = 1,2$ have no jump within the subintervals (t_{k-1}, t_k), $k = 1, \ldots, m$. Jumps can only occur on the endpoints of the subintervals.

Suppose

$$\sup_{0 \leq t \leq t_{k-1}, |\mathbf{b}-\beta| \leq \rho_n} |\widehat{A}_{nj}(t, \mathbf{b}) - \widetilde{A}_j(t, \mathbf{b})| = O_p(n^{-1/2}), \quad j = 1,2. \tag{3.19}$$

For $t_{k-1} < s < t_k$, $\widehat{S}_{nj}(s-, \mathbf{b}) = \widehat{S}_{nj}(t_{k-1}, \mathbf{b})$, $\widehat{p}_n(s-, \mathbf{b}, Z_i) = \widehat{p}_n(t_{k-1}, \mathbf{b}, Z_i)$. The following decomposition plays a major role in the proof to extend the result in (3.19) onto the interval $(t_{k-1}, t_k]$:

$$\begin{aligned}
& \sqrt{n}(d\widehat{A}_{nj}(s, \mathbf{b}) - d\widetilde{A}_j(s, \mathbf{b})) \\
= \; & \frac{1}{\frac{1}{n}\widehat{S}_{nj}(t_{k-1}, \mathbf{b})} \frac{1}{\sqrt{n}} \sum_i \widehat{p}_n(t_{k-1}, \mathbf{b}, Z_i)(dN_{ij}(s) - \lambda_{ij}(s)ds) \\
& + \frac{1}{\frac{1}{n}\widehat{S}_{nj}(t_{k-1}, \mathbf{b})} \frac{1}{\sqrt{n}} \sum_i (\widehat{p}_n(t_{k-1}, \mathbf{b}, Z_i) - p_n(t_{k-1}, Z_i))\lambda_{ij}(s)ds \\
& + \frac{1}{\frac{1}{n}\widehat{S}_{nj}(t_{k-1}, \mathbf{b})} \frac{1}{\sqrt{n}} \sum_i \left(p_n(t_{k-1}, Z_i)\lambda_{ij}(s)ds - S_j^0(s)a_j(s)ds\right) \\
& + \frac{\sqrt{n}S_j^0(s)a_j(s)ds}{\frac{1}{n}\widehat{S}_{nj}(t_{k-1}, \mathbf{b})} - \frac{\sqrt{n}S_j^0(s)a_j(s)ds}{h_j(s, \mathbf{b})}.
\end{aligned} \tag{3.20}$$

Combining the fact that the derivative of $\frac{1}{n}\widehat{S}_{nj}(t_{k-1}, \mathbf{b})$ with regard to $\widehat{A}_{nj}(t_{k-1}, \mathbf{b})$ is $O_p(n^{2\alpha})$ with the induction assumption (3.19), it follows that

$$\frac{1}{n}\widehat{S}_{nj}(t_{k-1}, \mathbf{b}) = h_j(t_{k-1}, \mathbf{b}) + o_p(1). \tag{3.21}$$

The first term in (3.20) is obviously $O_p(1)$. The second term can be further

decomposed into

$$\frac{1}{\frac{1}{n}\widehat{S}_{nj}(t_{k-1},\mathbf{b})}\frac{1}{\sqrt{n}}\sum_i(\widehat{p}_n(t_{k-1},\mathbf{b},Z_i)-\widetilde{p}_n(t_{k-1},\mathbf{b},Z_i))\lambda_{ij}(s)ds$$
$$+(\widetilde{p}_n(t_{k-1},\mathbf{b},Z_i)-\widetilde{p}_n(t_{k-1},\beta,Z_i))\lambda_{ij}(s)ds. \quad (3.22)$$

The derivative of $\widetilde{p}_n$ is n^α if and only if $1-\widetilde{F}_1(t,\mathbf{b},Z_i)-\widetilde{F}_2(t,\mathbf{b},Z_i)$ is between $(c+1)n^{-\alpha}$ and $cn^{-\alpha}$, hence $\sum_i \frac{\partial\widetilde{p}_n}{\partial\widetilde{A}_j}(t_{k-1},\mathbf{b},Z_i)/\widehat{S}_{nj}(t_{k-1},\mathbf{b})=O_p(1)$. By first order Taylor expansion, together with the induction assumption (3.19) and condition (3.15), the second term is $O_p(1)$. By condition (3.18), the third term in (3.20) is also $O_p(1)$.

By first order Taylor expansion, the fourth term in (3.20) is equivalent to

$$-\frac{\sqrt{n}q_j(s)a_j(s)ds}{\frac{1}{n^2}\widehat{S}^2_{nj}(t_{k-1},\mathbf{b})}\left(\frac{1}{n}\widehat{S}_{nj}(t_{k-1},\mathbf{b})-h_j(s,\mathbf{b})\right) \quad (3.23)$$
$$=-\frac{\sqrt{n}q_j(s)a_j(s)ds}{\frac{1}{n^2}\widehat{S}^2_{nj}(t_{k-1},\mathbf{b})}\left(\frac{1}{n}\widehat{S}_{nj}(t_{k-1},\mathbf{b})-h_j(t_{k-1},\mathbf{b})+h_j(t_{k-1},\mathbf{b})-h_j(s,\mathbf{b})\right).$$

It is easy to show that the derivatives of $\frac{1}{n}\widehat{S}_{nj}(t_{k-1},\mathbf{b})$ with regard to $\widehat{A}_l$, $l=1,2$ devided by $\frac{1}{n^2}\widehat{S}^2_{nj}(t_{k-1},\mathbf{b})$ are $O_p(1)$. Together with the induction assumption (3.19) and condition (3.17), it follows that the above term is $O_p(1)$. Thus we have shown that all the terms in the decomposition (3.20) are $O_p(1)$, which completes the proof. □

Replacing $\widehat{A}_{nj}$ in $U_{nj}(\mathbf{b})$ with $\widetilde{A}_j$, we get the asymptotic equivalent of the score functions:

$$\begin{aligned}\widetilde{U}_{nj}(\mathbf{b}) &= \sum_i\int_0^\tau \widetilde{\eta}_{ij,n}(t,\mathbf{b})(dN_{ij}(t)-d\widetilde{\Lambda}_{ij}(t,\mathbf{b}))\\ &\quad+\sum_i\int_0^\tau \widetilde{\phi}_{ij,n}(t,\mathbf{b})(dN_{ij'}(t)-d\widetilde{\Lambda}_{ij'}(t,\mathbf{b})),\end{aligned} \quad (3.24)$$

where

$$d\widetilde{\Lambda}_{ij,n}(t,\mathbf{b}) = \frac{Y_i(t)(1-\widetilde{F}_{nj}(t,\mathbf{b},Z_i))\exp(Z_i^T b_j)d\widetilde{A}_j(t,\mathbf{b})}{1-\widetilde{F}_{n1}(t,\mathbf{b},Z_i)-\widetilde{F}_{n2}(t,\mathbf{b},Z_i)}, \tag{3.25}$$

$$\widetilde{\eta}_{ij,n}(t,\mathbf{b}) = \widetilde{p}_n(t,\mathbf{b},Z_i)\Big(Z_i - Z_i\exp(Z_i^T b_j)\widetilde{A}_j(t,\mathbf{b}) + \frac{(1-\widetilde{F}_{nj}(t,\mathbf{b},Z_i))\exp(Z_i^T b_j)Z_i\widetilde{A}_j(t,\mathbf{b})}{1-\widetilde{F}_{n1}(t,\mathbf{b},Z_i)-\widetilde{F}_{n2}(t,\mathbf{b},Z_i)}\Big), \tag{3.26}$$

$$\widetilde{\phi}_{ij,n}(t,\mathbf{b}) = \frac{\widetilde{p}_n(t,\mathbf{b},Z_i)(1-\widetilde{F}_{nj}(t,\mathbf{b},Z_i))\exp(Z_i^T b_j)Z_i\widetilde{A}_j(t,\mathbf{b})}{1-\widetilde{F}_{n1}(t,\mathbf{b},Z_i)-\widetilde{F}_{n2}(t,\mathbf{b},Z_i)}. \tag{3.27}$$

Theorem 3.1. *Under assumptions A1-A4, for $\alpha < 1/6$ and ρ_n defined in* (3.15),

$$\sup_{|\mathbf{b}-\boldsymbol{\beta}|\le\rho_n} |U_{nj}(\mathbf{b}) - \widetilde{U}_{nj}(\mathbf{b})| = o_p(n), \quad j=1,2 \tag{3.28}$$

Proof. It follows from Lemma 3.1 and the fact that the derivatives of $U_{nj}(\mathbf{b})$ with regard to $\widehat{A}_j(t,\mathbf{b}), j=1,2, 0\le t\le\tau$ are $O(n^{3\alpha+1})$.

□

It is obvious that $A_1(t)$ and $A_2(t)$ satisfy the equations (3.10) and (3.14) when $\mathbf{b}=\boldsymbol{\beta}$; therefore, $\widetilde{A}_j(t,\boldsymbol{\beta}) = A_j(t)$, $\widetilde{F}_j(t,\boldsymbol{\beta},z) = F_j(t,z)$, $j=1,2$. In what follows, when $\mathbf{b}=\boldsymbol{\beta}$, the tilde in the notations can be removed and they are equal to the corresponding quantities evaluated at the true parameters.

$\widetilde{U}_{nj}(\boldsymbol{\beta})$ can be written as a sum of local square integrable martingales:

$$\widetilde{U}_{nj}(\boldsymbol{\beta}) = \sum_i \int_0^\tau \eta_{ij}(t)dM_{ij}(t) + \sum_i \int_0^\tau \phi_{ij}(t)dM_{ij'}(t), \tag{3.29}$$

where

$$\eta_{ij}(t) = p_n(t, Z_i)\left(Z_i - Z_i \exp(Z_i^T \beta_j) A_j(t) + \frac{Z_i(1 - F_j(t, Z_i)) \exp(Z_i^T \beta_j) A_j(t)}{1 - F_1(t, Z_i) - F_2(t, Z_i)}\right), \tag{3.30}$$

$$\phi_{ij}(t) = \frac{p_n(t, Z_i) Z_i (1 - F_j(t, Z_i)) \exp(Z_i^T \beta_j) A_j(t)}{1 - F_1(t, Z_i) - F_2(t, Z_i)}, \tag{3.31}$$

$$dM_{ij}(t) = dN_{ij}(t) - \lambda_{ij}(t)dt$$

$$\lambda_{ij}(t) = \frac{Y_i(t)(1 - F_1(t, Z_i)) \exp(Z_i^T \beta_j) a_j(t)}{1 - F_1(t, Z_i) - F_2(t, Z_i)}.$$

Theorem 3.2. *Under assumptions A1-A4, for $\alpha < 1/6$, $\widehat{\beta}_{nj}, j = 1, 2$ are consistent.*

Proof. The consistency of $\widehat{\beta}_{nj}$ follows from Theorem 3.1 and the fact that $E(\widetilde{U}_{nj}(\boldsymbol{\beta})) = 0$. □

3.2 Asymptotic Linearity of $U_n(\mathbf{b})$

Under assumption A5, $\widetilde{A}_j, j = 1, 2$ are also differentiable in $\mathbf{b}$. The derivatives are defined as the solution to the following ordinary differential equations:

$$\frac{\partial \widetilde{A}_j(t, \mathbf{b})}{\partial \mathbf{b}} = \int_0^t S_j^0(s) a_j(s) ds \frac{\partial}{\partial \mathbf{b}} \left(\frac{1}{h_j(s, \mathbf{b}, \widetilde{A}_1(s, \mathbf{b}), \widetilde{A}_2(s, \mathbf{b}))}\right), \quad j = 1, 2. \tag{3.32}$$

We are interested in the derivatives evaluated at the true parameters $\boldsymbol{\beta}$.

$$\frac{\partial h_1(t, \mathbf{b})}{\partial b_1}\Big|_{\mathbf{b}=\boldsymbol{\beta}} = D_1(t) + R_1(t)\frac{\partial \widetilde{A}_1(t, \boldsymbol{\beta})}{\partial b_1} + Q_1(t)\frac{\partial \widetilde{A}_2(t, \boldsymbol{\beta})}{\partial b_1}, \tag{3.33}$$

$$\frac{\partial h_2(t,\mathbf{b})}{\partial b_1}\Big|_{\mathbf{b}=\boldsymbol{\beta}} = E_2(t) + Q_2(t)\frac{\partial \widetilde{A}_1(t,\boldsymbol{\beta})}{\partial b_1} + R_2(t)\frac{\partial \widetilde{A}_2(t,\boldsymbol{\beta})}{\partial b_1}, \tag{3.34}$$

where $D_j, E_j, j = 1,2$ are $p \times 1$ vectors, $R_j, Q_j, j = 1,2$ are real valued functions. They are defined as

$$D_j(t) = \frac{\partial h_j(t,\mathbf{b},\widetilde{A}_1(t,\boldsymbol{\beta}),\widetilde{A}_2(t,\boldsymbol{\beta}))}{\partial b_j}\Big|_{\mathbf{b}=\boldsymbol{\beta}},$$

$$E_j(t) = \frac{\partial h_j(t,\mathbf{b},\widetilde{A}_1(t,\boldsymbol{\beta}),\widetilde{A}_2(t,\boldsymbol{\beta}))}{\partial b_{j'}}\Big|_{\mathbf{b}=\boldsymbol{\beta}},$$

$$R_j(t) = \frac{\partial h_j(t,\mathbf{b},\widetilde{A}_1(t,\boldsymbol{\beta}),\widetilde{A}_2(t,\boldsymbol{\beta}))}{\partial \widetilde{A}_j}\Big|_{\mathbf{b}=\boldsymbol{\beta}},$$

$$Q_j(t) = \frac{\partial h_j(t,\mathbf{b},\widetilde{A}_1(t,\boldsymbol{\beta}),\widetilde{A}_2(t,\boldsymbol{\beta}))}{\partial \widetilde{A}_{j'}}\Big|_{\mathbf{b}=\boldsymbol{\beta}}.$$

Plugging (3.33) and (3.34) into (3.32), we have

$$\frac{d}{dt}\frac{\partial \widetilde{A}_1(t,\boldsymbol{\beta})}{\partial b_1} = -\frac{a_1(t)}{S_1^0(t)}\left(D_1(t) + R_1(t)\frac{\partial \widetilde{A}_1(t,\boldsymbol{\beta})}{\partial b_1} + Q_1(t)\frac{\partial \widetilde{A}_2(t,\boldsymbol{\beta})}{\partial b_1}\right),$$

$$\frac{d}{dt}\frac{\partial \widetilde{A}_2(t,\boldsymbol{\beta})}{\partial b_1} = -\frac{a_2(t)}{S_2^0(t)}\left(E_2(t) + Q_2(t)\frac{\partial \widetilde{A}_1(t,\boldsymbol{\beta})}{\partial b_1} + R_2(t)\frac{\partial \widetilde{A}_2(t,\boldsymbol{\beta})}{\partial b_1}\right).$$

A similar system of equations exist for the derivatives of $\widetilde{A}$ with regard to b_2.

$$\frac{d}{dt}\frac{\partial \widetilde{A}_1(t,\boldsymbol{\beta})}{\partial b_2} = -\frac{a_1(t)}{S_1^0(t)}\left(E_1(t) + R_1(t)\frac{\partial \widetilde{A}_1(t,\boldsymbol{\beta})}{\partial b_2} + Q_1(t)\frac{\partial \widetilde{A}_2(t,\boldsymbol{\beta})}{\partial b_2}\right),$$

$$\frac{d}{dt}\frac{\partial \widetilde{A}_2(t,\boldsymbol{\beta})}{\partial b_2} = -\frac{a_2(t)}{S_2^0(t)}\left(D_2(t) + Q_2(t)\frac{\partial \widetilde{A}_1(t,\boldsymbol{\beta})}{\partial b_2} + R_2(t)\frac{\partial \widetilde{A}_2(t,\boldsymbol{\beta})}{\partial b_2}\right).$$

These ordinary differential equations can be solved numerically with initial values

$$\frac{\partial \widetilde{A}_j(0,\boldsymbol{\beta})}{\partial b_l} = 0, \quad j,l = 1,2.$$

$\widetilde{U}_n(\mathbf{b})$, which is a smooth function of $\mathbf{b}$ and $\widetilde{A}_j$, $j = 1,2$, is also differentiable in $\mathbf{b}$. Its derivatives can be readily obtained once the derivatives of $\widetilde{A}$ with regard to $\mathbf{b}$ are known.

Theorem 3.3. *Under assumptions A1-A6, for $\alpha < 1/6$,*

$$\lim_{n\to\infty}\frac{1}{n}\frac{\partial U_n(\boldsymbol{\beta})}{\partial \mathbf{b}} = \lim_{n\to\infty}\frac{1}{n}\frac{\partial \widetilde{U}_n(\boldsymbol{\beta})}{\partial \mathbf{b}}, \qquad \textit{in probability.} \tag{3.35}$$

Denote the matrix on the right hand side by Σ,

$$\Sigma = \lim_{n\to\infty}\begin{pmatrix} \frac{1}{n}\frac{\partial \widetilde{U}_{n1}(\boldsymbol{\beta})}{\partial b_1} & \frac{1}{n}\frac{\partial \widetilde{U}_{n1}(\boldsymbol{\beta})}{\partial b_2} \\ \frac{1}{n}\frac{\partial \widetilde{U}_{n2}(\boldsymbol{\beta})}{\partial b_1} & \frac{1}{n}\frac{\partial \widetilde{U}_{n2}(\boldsymbol{\beta})}{\partial b_2} \end{pmatrix}. \tag{3.36}$$

Proof. It follows directly from Theorem 3.1. □

3.3 Asymptotic Normality of $U_n(\boldsymbol{\beta})$

Assume the following limits exist in probability, $j,k = 1,2$,

$$S_j^{\eta_j}(t) = \lim_{n\to\infty 0}\frac{1}{n}\sum_i \eta_{ij}(t)\frac{Y_i(t)(1-F_j(t,Z_i))\exp(Z_i^T\beta_j)}{1-F_1(t,Z_i)-F_2(t,Z_i)}, \tag{3.37}$$

$$S_{j'}^{\phi_j}(t) = \lim_{n\to\infty} \frac{1}{n} \sum_i \phi_{ij}(t) \frac{Y_i(t)(1 - F_{j'}(t, Z_i)) \exp(Z_i^T \beta_{j'})}{1 - F_1(t, Z_i) - F_2(t, Z_i)}, \tag{3.38}$$

$$S_j^{2\eta_j}(t) = \lim_{n\to\infty} \frac{1}{n} \sum_i \eta_{ij}(t) \eta_{ij}(t)^T \frac{Y_i(t)(1 - F_j(t, Z_i)) \exp(Z_i^T \beta_j)}{1 - F_1(t, Z_i) - F_2(t, Z_i)}, \tag{3.39}$$

$$S_{j'}^{2\phi_j}(t) = \lim_{n\to\infty} \frac{1}{n} \sum_i \phi_{ij}(t) \phi_{ij}(t)^T \frac{Y_i(t)(1 - F_{j'}(t, Z_i)) \exp(Z_i^T \beta_{j'})}{1 - F_1(t, Z_i) - F_2(t, Z_i)}, \tag{3.40}$$

$$S_j^{\eta_1 \phi_2}(t) = \lim_{n\to\infty} \frac{1}{n} \sum_i \eta_{i1}(t) \phi_{i2}(t)^T \frac{Y_i(t)(1 - F_j(t, Z_i)) \exp(Z_i^T \beta_j)}{1 - F_1(t, Z_i) - F_2(t, Z_i)}, \tag{3.41}$$

$$S_j^{\phi_1 \eta_2}(t) = \lim_{n\to\infty} \frac{1}{n} \sum_i \phi_{i1}(t) \eta_{i2}(t)^T \frac{Y_i(t)(1 - F_j(t, Z_i)) \exp(Z_i^T \beta_j)}{1 - F_1(t, Z_i) - F_2(t, Z_i)}, \tag{3.42}$$

$$\begin{aligned} K_{ijj}(t) &= \frac{\partial}{\partial A_j} \left(\frac{1 - F_j(t, Z_i)}{1 - F_1(t, Z_i) - F_2(t, Z_i)} \right) \\ &= -\frac{(1 - F_j(t, Z_i)) \exp(Z_i^T \beta_j)}{1 - F_1(t, Z_i) - F_2(t, Z_i)} + \frac{(1 - F_j(t, Z_i))^2 \exp(Z_i^T \beta_j)}{(1 - F_1(t, Z_i) - F_2(t, Z_i))^2}, \end{aligned}$$

$$\begin{aligned} K_{ijj'}(t) &= \frac{\partial}{\partial A_{j'}} \left(\frac{1 - F_j(t, Z_i)}{1 - F_1(t, Z_i) - F_2(t, Z_i)} \right) \\ &= \frac{(1 - F_1(t, Z_i))(1 - F_2(t, Z_i)) \exp(Z_i^T \beta_{j'})}{(1 - F_1(t, Z_i) - F_2(t, Z_i))^2}, \end{aligned}$$

$$\nu_{kj}(t) = \lim_{n\to\infty} \frac{1}{\widehat{S}_{nj}(t, \boldsymbol{\beta})} \sum_i p_n(t, Z_i) Y_i(t) \exp(Z_i^T \beta_j) K_{ijk}(t), \tag{3.43}$$

$$\xi_{jk}^{\eta_j}(t) = \lim_{n\to\infty} \frac{1}{n} \sum_i \eta_{ij}(t) Y_i(t) \exp(Z_i^T \beta_j) K_{ijk}(t) - S_j^{\eta_j}(t) \nu_{kj}(t), \tag{3.44}$$

$$\xi_{j'k}^{\phi_j}(t) = \lim_{n\to\infty} \frac{1}{n} \sum_i \phi_{ij}(t) Y_i(t) \exp(Z_i^T \beta_{j'}) K_{ij'k}(t) - S_{j'}^{\phi_j}(t) \nu_{kj'}(t). \tag{3.45}$$

First we derive the asymptotic equivalent of $\widehat{A}_{nj}(t, \boldsymbol{\beta})$.

Lemma 3.2. *Under assumptions A1-A4 and A7, for $\alpha < 1/6$, Define*

$$\mathbf{A}_n(t) = \left(\sqrt{n}(\widehat{A}_{n1}(t,\boldsymbol{\beta}) - A_1(t)), \quad \sqrt{n}(\widehat{A}_{n2}(t,\boldsymbol{\beta}) - A_2(t))\right), \tag{3.46}$$

$$\mathbf{W}_n(t) = \left(\frac{1}{\sqrt{n}}\int_0^t \sum_i \frac{p_n(s,Z_i)}{S_1^0(s)} dM_{i1}(s), \quad \frac{1}{\sqrt{n}}\int_0^t \sum_i \frac{p_n(s,Z_i)}{S_2^0(s)} dM_{i2}(s)\right), \tag{3.47}$$

$$d\boldsymbol{\nu}(t) = \begin{pmatrix} \nu_{11}(t)dA_1(t) & \nu_{12}(t)dA_2(t) \\ \nu_{21}(t)dA_1(t) & \nu_{22}(t)dA_2(t) \end{pmatrix}. \tag{3.48}$$

$\mathbf{A}_n(t)$ *follows the following Volterra equation asymptotically:*

$$\mathbf{A}_n(t) = \mathbf{W}_n(t) - \int_0^t \mathbf{A}_n(s-)d\boldsymbol{\nu}(s) + o_p(1). \tag{3.49}$$

Its unique solution is

$$\mathbf{A}_n(t) = \int_0^t \mathbf{W}_n(ds)\Phi(s,t) + o_p(1), \tag{3.50}$$

where

$$\Phi(s,t) = \boldsymbol{\pi}_{(s,t]}(\mathrm{I} - d\boldsymbol{\nu}). \tag{3.51}$$

Proof. The proof is based on the following decomposition:

$$\begin{aligned}
&\sqrt{n}(d\widehat{A}_{nj}(t,\boldsymbol{\beta}) - dA_j(t)) \\
&= \frac{1}{\frac{1}{n}\widehat{S}_{nj}(t-,\boldsymbol{\beta})}\frac{1}{\sqrt{n}}\sum_i \widehat{p}_n(t-,\boldsymbol{\beta},Z_i)(dN_{ij}(t) - \lambda_{ij}(t)dt) \\
&\quad - \widehat{p}_n(t-,\boldsymbol{\beta},Z_i)Y_i(t)\exp(Z_i^T b_j)\left(\frac{1-\widehat{F}_{nj}(t-,\boldsymbol{\beta},Z_i)}{1-\widehat{F}_{n1}(t-,\boldsymbol{\beta},Z_i)-\widehat{F}_{n2}(t-,\boldsymbol{\beta},Z_i)}\right. \\
&\qquad \left. - \frac{1-F_j(t,Z_i)}{1-F_1(t,Z_i)-F_2(t,Z_i)}\right) dA_j(t).
\end{aligned} \tag{3.52}$$

Since the derivatives of $\widehat{p}_n(t,\boldsymbol{\beta},Z_i)$ with regard to $\widehat{A}_{nk}(t,\boldsymbol{\beta}), k = 1,2$ are $O(1)$, by

the first-order Taylor expansion and Lemma 3.1, we can replace $\widehat{p}_n(t-,\boldsymbol{\beta},Z_i)$ with $p_n(t,Z_i)$. Similarly, $\frac{1}{n}\widehat{S}_{nj}(t-,\boldsymbol{\beta})$ can be replaced with $S_j^0(t)$, since the derivatives of $\frac{1}{n}\widehat{S}_{nj}(t,\boldsymbol{\beta})$ with regard to $\widehat{A}_{nk}(t,\boldsymbol{\beta}), k=1,2$ are $O_p(n^{2\alpha})$.

Hence, replacing $\frac{1}{n}\widehat{S}_{nj}(t,\boldsymbol{\beta})$ with $S_j^0(t)$ and $\widehat{p}_n(t,\boldsymbol{\beta},Z_i)$ with $p_n(t,Z_i)$, it follows immediately that the first term in (3.52) is equivalent to

$$\frac{1}{\sqrt{n}}\sum_i \frac{p_n(t,Z_i)}{S_j^0(t)} dM_{ij}(t). \tag{3.53}$$

By Taylor expansion we have

$$\begin{aligned}
&\sqrt{n}\left(\frac{1-\widehat{F}_{nj}(t,\boldsymbol{\beta},Z_i)}{1-\widehat{F}_{n1}(t,\boldsymbol{\beta},Z_i)-\widehat{F}_{n2}(t,\boldsymbol{\beta},Z_i)} - \frac{1-F_j(t,Z_i)}{1-F_1(t,Z_i)-F_2(t,Z_i)}\right) \\
= \;& \sqrt{n}K_{ij1}(t)(\widehat{A}_{n1}(t)-A_1(t)) + \sqrt{n}K_{ij2}(t)(\widehat{A}_{n2}(t)-A_2(t)) + o_p(1).
\end{aligned} \tag{3.54}$$

The second-order term in the Taylor expansion (3.54) is $o_p(1)$, since the second-order derivatives are $O(n^{3\alpha})$ and $(\widehat{A}_{nj}(t,\boldsymbol{\beta})-A_j(t))^2 = O_p(n^{-1})$. By (3.54), the second term in (3.52) is asymptotically equivalent to

$$\begin{aligned}
&\frac{1}{\widehat{S}_{nj}(t,\boldsymbol{\beta})}\sum_{k=1}^{2}\sum_i p_n(t,Z_i)\mathrm{Y}_i(t)\exp(Z_i^T\beta_j)K_{ijk}(t)\sqrt{n}(\widehat{A}_{nk}(t-,\boldsymbol{\beta})-A_k(t))dA_j(t) \\
=&\sum_{k=1}^{2}\nu_{kj}(t)\sqrt{n}(\widehat{A}_{nk}(t-,\boldsymbol{\beta})-A_k(t))dA_j(t)+o_p(1),
\end{aligned} \tag{3.55}$$

where $\nu_{kj}(t)$ is defined in (3.43).

Hence we have shown that $\mathbf{A}_n(t)$ follows the Volterra equation (3.49). By Theorem II.6.3 in Andersen et al. (1992), its unique solution is given by (3.50).

□

Next we derive the asymptotic equivalent of $U_{nj}(\boldsymbol{\beta})/\sqrt{n}, j=1,2$.

Theorem 3.4. *Under assumptions A1-A4 and A7, for $\alpha < 1/6$,*

$$\begin{aligned} U_{nj}(\boldsymbol{\beta})/\sqrt{n} &= \frac{1}{\sqrt{n}}\sum_i \int_0^\tau \left(\eta_{ij}(t) - \frac{p_n(t,Z_i)S_j^{\eta_j}(t)}{S_j^0(t)}\right) dM_{ij}(t) \\ &\quad + \frac{1}{\sqrt{n}}\sum_i \int_0^\tau \left(\phi_{ij}(t) - \frac{p_n(t,Z_i)S_{j'}^{\phi_j}(t)}{S_{j'}^0(t)}\right) dM_{ij'}(t) \\ &\quad - \sum_{k=1}^2 \sqrt{n}\int_0^\tau \xi_{jk}^{\eta_j}(t)(\widehat{A}_{nk}(t-,\boldsymbol{\beta}) - A_k(t))dA_j(t) \\ &\quad - \sum_{k=1}^2 \sqrt{n}\int_0^\tau \xi_{j'k}^{\phi_j}(t)(\widehat{A}_{nk}(t-,\boldsymbol{\beta}) - A_k(t))dA_{j'}(t) + o_p(1). \end{aligned} \tag{3.56}$$

Proof. Recall that the score function in (2.14) has two parts. First we derive the asymptotic equivalent of the first part $\frac{1}{\sqrt{n}}\sum_i \int_0^\tau \widehat{\eta}_{ij,n}(t,\boldsymbol{\beta})(dN_{ij}(t) - d\widehat{\Lambda}_{ij,n}(t,\boldsymbol{\beta}))$ based on the following decomposition:

$$\begin{aligned} &\frac{1}{\sqrt{n}}\sum_i \int_0^\tau \widehat{\eta}_{ij,n}(t,\boldsymbol{\beta})(dN_{ij}(t) - d\widehat{\Lambda}_{ij,n}(t,\boldsymbol{\beta})) - \eta_{ij}(t)dM_{ij}(t) \\ =&\frac{1}{\sqrt{n}}\sum_i \int_0^\tau (\widehat{\eta}_{ij,n}(t,\boldsymbol{\beta}) - \eta_{ij}(t))dM_{ij}(t) \\ &- \sqrt{n}\int_0^\tau \frac{1}{n}\sum_i \widehat{\eta}_{ij,n}(t,\boldsymbol{\beta})\frac{Y_i(t)(1-\widehat{F}_{nj}(t,\boldsymbol{\beta},Z_i))\exp(Z_i^T\beta_j)}{1-\widehat{F}_{n1}(t,\boldsymbol{\beta},Z_i)-\widehat{F}_{n2}(t,\boldsymbol{\beta},Z_i)}(d\widehat{A}_{nj}(t) - dA_j(t)) \\ &+ \frac{1}{\sqrt{n}}\sum_i \int_0^\tau \widehat{\eta}_{ij,n}(t,\boldsymbol{\beta})Y_i(t)\exp(Z_i^T\beta_j)\left(\frac{1-F_j(t,Z_i)}{1-F_1(t,Z_i)-F_2(t,Z_i)}\right. \\ &\qquad \left. - \frac{1-\widehat{F}_{nj}(t,\boldsymbol{\beta},Z_i)}{1-\widehat{F}_{n1}(t,\boldsymbol{\beta},Z_i)-\widehat{F}_{n2}(t,\boldsymbol{\beta},Z_i)}\right) dA_j(t). \end{aligned} \tag{3.57}$$

By first-order Taylor expansion and Lemma 3.1, it can be shown that $\widehat{\eta}_{ij,n}(t,\boldsymbol{\beta}) - \eta_{ij}(t) = O_p(n^{2\alpha-1/2})$, hence the first term in (3.57) is $o_p(1)$.

Furthermore,

$$\begin{aligned}&\widehat{\eta}_{ij,n}(t,\boldsymbol{\beta})\frac{Y_i(t)(1-\widehat{F}_{nj}(t,\boldsymbol{\beta},Z_i))\exp(Z_i^T\beta_j)}{1-\widehat{F}_{n1}(t,\boldsymbol{\beta},Z_i)-\widehat{F}_{n2}(t,\boldsymbol{\beta},Z_i)}\\ =\;&\eta_{ij}(t)\frac{Y_i(t)(1-F_j(t,Z_i))\exp(Z_i^T\beta_j)}{1-F_1(t,Z_i)-F_2(t,Z_i)}+O_p(n^{3\alpha-1/2}),\end{aligned}\tag{3.58}$$

thus we can replace the integrand in the second term with $S_j^{\eta_j}(t)$. Combining with the Volterra equation in (3.49), the second term is asymptotically equivalent to

$$-\frac{1}{\sqrt{n}}\sum_i\int_0^\tau\frac{p_n(t,Z_i)S_j^{\eta_j}(t)}{S_j^0(t)}dM_{ij}(t)+\sum_{k=1}^2\sqrt{n}\int_0^\tau S_j^{\eta_j}(t)\nu_{kj}(t)(\widehat{A}_{nk}(t-,\boldsymbol{\beta})-A_k(t))dA_j(t).\tag{3.59}$$

Using the first-order approximation in (3.54) for the third term and by (3.59), the second and the third terms together are asymptotic equivalent to

$$-\frac{1}{\sqrt{n}}\sum_i\int_0^\tau\frac{p_n(t,Z_i)S_j^{\eta_j}(t)}{S_j^0(t)}dM_{ij}(t)-\sum_{k=1}^2\sqrt{n}\int_0^\tau\xi_{jk}^{\eta_j}(t)(\widehat{A}_{nk}(t-,\boldsymbol{\beta})-A_k(t))dA_j(t).\tag{3.60}$$

Similar asymptotic equivalence can be derived for the second part in (2.14). □

To derive the asymptotic variance of $\frac{1}{\sqrt{n}}U_{nj}(\boldsymbol{\beta})$, first we should note that M_{i1} and M_{i2} are orthogonal for every i. Next, we will show that the first term in (3.56) is asymptotically uncorrelated with the third and the fourth terms. For $l,k=1,2$, by (3.50),

$$\begin{aligned}&\mathrm{E}\left(\frac{1}{\sqrt{n}}\sum_i\int_0^\tau\left(\eta_{ij}(t)-\frac{p_n(t,Z_i)S_j^{\eta_j}(t)}{S_j^0(t)}\right)dM_{ij}(t)\right.\\ &\left.\cdot\sqrt{n}\int_0^\tau\xi_{lk}(t)(\widehat{A}_{nk}(t-,\boldsymbol{\beta})-A_k(t))dA_l(t)\right)\end{aligned}$$

is asymptotically equivalent to

$$\begin{aligned}&\frac{1}{n}\sum_i \int_0^\tau \mathrm{E}\left(\int_0^t \left(\eta_{ij}(u) - \frac{p_n(u,Z_i)S_j^{\eta_j}(u)}{S_j^0(u)}\right) dM_{ij}(u)\right.\\ &\qquad\left.\cdot \int_0^t \Phi_{jk}(v,t)\frac{p_n(v,Z_i)}{S_j^0(v)} dM_{ij}(v)\right)\xi_{lk}(t)dA_l(t)\\ =&\frac{1}{n}\sum_i \int_0^\tau \xi_{lk}(t)dA_l(t)\int_0^t \left(\eta_{ij}(u) - \frac{p_n(u,Z_i)S_j^{\eta_j}(u)}{S_j^0(u)}\right)\Phi_{jk}(u,t)\frac{p_n(u,Z_i)}{S_j^0(u)}\mathrm{E}(dM_{ij}^2(u)),\end{aligned} \tag{3.61}$$

where $\Phi_{jk}(u,t)$ is the element on the jth row and kth column in the 2×2 matrix $\Phi(u,t)$. Since $\mathrm{E}(dM_{ij}^2(u)) = \lambda_{ij}(u)$, (3.61) is equal to

$$\int_0^\tau \xi_{lk}(t)dA_l(t)\int_0^t \Phi_{jk}(u,t)\frac{1}{n}\sum_i \left(\eta_{ij}(u) - \frac{p_n(u,Z_i)S_j^{\eta_j}(u)}{S_j^0(u)}\right)\frac{p_n(u,Z_i)\lambda_{ij}(u)du}{S_j^0(u)}. \tag{3.62}$$

It is easy to show that

$$\lim_{n\to\infty}\frac{1}{n}\sum_i p_n(u,Z_i)\eta_{ij}(u)\lambda_{ij}(u) = S_j^{\eta_j}(u)a_j(u), \tag{3.63}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_i p_n^2(u,Z_i)\lambda_{ij}(u) = S_j^0(u)a_j(u). \tag{3.64}$$

Hence

$$\sum_i \left(\eta_{ij}(u) - \frac{p_n(u,Z_i)S_j^{\eta_j}(u)}{S_j^0(u)}\right)\frac{p_n(u,Z_i)\lambda_{ij}(u)du}{S_j^0(u)} = 0.$$

Therefore (3.62) is 0 and the first term is asymptotically uncorrelated with the other terms. Similarly, the second term is asymptotically uncorrelated with the other terms.

By the martingale central limit theorem, we have

$$\begin{aligned} V_j^{\eta_j} &= \lim_{n\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_i \int_0^\tau \left(\eta_{ij}(t) - \frac{p_n(t,Z_i)S_j^{\eta_j}(t)}{S_j^0(t)}\right) dM_{ij}(t)\right) \\ &= \int_0^\tau \left(\frac{S_j^{2\eta_j}(t)}{S_j^0(t)} - \frac{S_j^{\eta_j}(t)(S_j^{\eta_j}(t))^T}{(S_j^0(t))^2}\right) S_j^0(t)dA_j(t), \end{aligned} \tag{3.65}$$

$$\begin{aligned} V_{j'}^{\phi_j} &= \lim_{n\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_i \int_0^\tau \left(\phi_{ij}(t) - \frac{p_n(t,Z_i)S_{j'}^{\phi_j}(t)}{S_{j'}^0(t)}\right) dM_{ij'}(t)\right) \\ &= \int_0^\tau \left(\frac{S_{j'}^{2\phi_j}(t)}{S_{j'}^0(t)} - \frac{S_{j'}^{\phi_j}(t)(S_{j'}^{\phi_j}(t))^T}{(S_{j'}^0(t))^2}\right) S_{j'}^0(t)dA_{j'}(t). \end{aligned} \tag{3.66}$$

Define

$$d\boldsymbol{\xi}_j(t) = \left(\xi_{j1}^{\eta_j}(t)dA_j(t) + \xi_{j'1}^{\phi_j}(t)dA_{j'}(t) \quad \xi_{j2}^{\eta_j}(t)dA_j(t) + \xi_{j'2}^{\phi_j}(t)dA_{j'}(t) \right), \tag{3.67}$$

to simplify notation, write the last two terms in (3.56) in a matrix form as

$$-\int_0^\tau d\boldsymbol{\xi}_j(t)\mathbf{A}_n^T(t). \tag{3.68}$$

By (3.50),

$$\begin{aligned} & E(\mathbf{A}_n^T(t)\mathbf{A}_n(s)) \\ &= \int_0^t \int_0^s \Phi(u,t)E(\mathbf{W}_n^T(du)\mathbf{W}_n(dv))\Phi(v,s) + o_p(1) \\ &= \int_0^{s\wedge t} \Phi(u,t)E(\mathbf{W}_n^T(du)\mathbf{W}_n(du))\Phi(u,s) + o_p(1) \\ &= \int_0^{s\wedge t} \Phi(u,t) \begin{pmatrix} \frac{dA_1(u)}{S_1^0(u)} & 0 \\ 0 & \frac{dA_2(u)}{S_2^0(u)} \end{pmatrix} \Phi(u,s) + o_p(1). \end{aligned} \tag{3.69}$$

The last equality holds by (3.64) and the fact that $M_{i1}(t)$ and M_{i2} are orthogonal, with compensators $\lambda_{ij}(t), j = 1, 2$ respectively.

The asymptotic variances of the score functions are

$$V^{U_j} = \lim_{n\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{n}} U_{nj}(\boldsymbol{\beta})\right) \tag{3.70}$$
$$= V_j^{\eta_j} + V_{j'}^{\phi_j} + \int_0^\tau \int_0^\tau d\boldsymbol{\xi}_j(t) \int_0^{s\wedge t} \Phi(u,t) \begin{pmatrix} \frac{dA_1(u)}{S_1^0(u)} & 0 \\ 0 & \frac{dA_2(u)}{S_2^0(u)} \end{pmatrix} \Phi(u,s) d\boldsymbol{\xi}_j^T(s),$$

$$V^{U_1,U_2} = \lim_{n\to\infty} \mathrm{Cov}\left(\frac{1}{\sqrt{n}} U_{n1}(\boldsymbol{\beta}), \frac{1}{\sqrt{n}} U_{n2}(\beta)\right) \tag{3.71}$$
$$= \int_0^\tau \left(\frac{S_1^{\eta_1\phi_2}(t)}{S_1^0(t)} - \frac{S_1^{\eta_1}(t)(S_1^{\phi_2}(t))^T}{(S_1^0(t))^2}\right) S_1^0(t) dA_1(t) \tag{3.72}$$
$$+ \int_0^\tau \left(\frac{S_2^{\phi_1\eta_2}(t)}{S_2^0(t)} - \frac{S_2^{\phi_1}(t)(S_2^{\eta_2}(t))^T}{(S_2^0(t))^2}\right) S_2^0(t) dA_2(t)$$
$$+ \int_0^\tau \int_0^\tau d\boldsymbol{\xi}_1(t) \int_0^{s\wedge t} \Phi(u,t) \begin{pmatrix} \frac{dA_1(u)}{S_1^0(u)} & 0 \\ 0 & \frac{dA_2(u)}{S_2^0(u)} \end{pmatrix} \Phi(u,s) d\boldsymbol{\xi}_2^T(s). \tag{3.73}$$

Finally we arrive at the asymptotic variance of the score function:

Theorem 3.5. *Under assumptions A1-A4 and A7, for $\alpha < 1/6$,*

$$V = \lim_{n\to\infty} Var\left(\frac{1}{\sqrt{n}} U_n(\boldsymbol{\beta})\right) = \begin{pmatrix} V^{U_1} & V^{U_1,U_2} \\ (V^{U_1,U_2})^T & V^{U_2} \end{pmatrix}. \tag{3.74}$$

3.4 Asymptotic Normality of $\widehat{\boldsymbol{\beta}}_n$

Theorem 3.6. *Under assumptions A1-A8, for $\alpha < 1/6$,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = -\Sigma^{-1}\frac{1}{\sqrt{n}}U_n(\boldsymbol{\beta}) + o_p(1), \tag{3.75}$$

where Σ is defined in (3.36)*. Furthermore,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N(0, \Sigma^{-1}V(\Sigma^{-1})^T), \tag{3.76}$$

with V defined in (3.74)*.*

Proof. By the Taylor expansion and Theorem 3.2,

$$0 = \frac{1}{\sqrt{n}}U_n(\widehat{\boldsymbol{\beta}}_n) = \frac{1}{\sqrt{n}}U_n(\boldsymbol{\beta}) + \frac{1}{n}\frac{\partial U_n(\boldsymbol{\beta})}{\partial \mathbf{b}}\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) + o_p(1). \tag{3.77}$$

(3.75) follows immediately from Theorem 3.3. (3.75) and Theorem 3.5 lead to the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$ as given in (3.76). □

3.5 Discussion

In this subsection we revisit assumptions A2 and A7 and consider when these assumptions hold as well as providing a remedy to situations when they do not hold. If $Z_i, i = 1, \ldots, n$ are i.i.d. with distribution μ,

$$\begin{aligned} S_j^0(t) &= E\left(\frac{Y_i(t)(1 - F_j(t, Z_i))\exp(Z_i^T\beta_j)}{1 - F_1(t, Z_i) - F_2(t, Z_i)}\right) \\ &= \int (1 - F_j(t, z))\exp(z^T\beta_j)G(t, z)d\mu(z) \end{aligned} \tag{3.78}$$

is always finite. Similarly, $\nu_{kj}(t)$ in (3.43) is also well defined. However, if $1 - F_1(t, Z) - F_2(t, Z) = 0$ with nonzero probability, the limits defined in (3.36)-(3.42) and (3.44)-(3.45) do not exist, because for $m \geq 2$,

$$\begin{aligned} & E\left(\frac{Y_i(t)(1 - F_j(t, Z_i))\exp(Z_i^T\beta_j)}{(1 - F_1(t, Z_i) - F_2(t, Z_i))^m}\right) \\ = & \int \frac{(1 - F_j(t, z))\exp(z^T\beta_j)G(t, z)}{(1 - F_1(t, z) - F_2(t, z))^{m-1}} d\mu(z) \end{aligned} \tag{3.79}$$

is infinite.

To handle this problem, we redefine $\eta_{ij}(t)$ and $\phi_{ij}(t)$ by dividing the left-hand side of (3.30) and (3.31) by

$$\zeta_n(t) \stackrel{def}{=} 1 + \frac{1}{n}\sum_i \frac{p_n(t, Z_i)}{(1 - F_1(t, Z_i) - F_2(t, Z_i))^2}. \tag{3.80}$$

The limits in (3.36)-(3.42) and (3.44)-(3.45) are now well defined. In order for Theorem 3.4 to hold, we only need to divide $U_{nj}(\beta)$ by $\zeta_n(t)$. Similarly, redefine Σ and V as

$$\Sigma = \lim_{n\to\infty} \frac{1}{n\zeta_n(t)} \frac{\partial \widetilde{U}_n(\beta)}{\partial \mathbf{b}},$$

$$V = \lim_{n\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{n}\zeta_n(t)} U_n(\beta)\right).$$

Dividing both sides of (3.77) by $\zeta_n(t)$, we have

$$0 = \frac{1}{\sqrt{n}\zeta_n(t)} U_n(\beta) + \frac{1}{n\zeta_n(t)} \frac{\partial U_n(\beta)}{\partial \mathbf{b}} \sqrt{n}(\widehat{\beta}_n - \beta) + o_p(1). \tag{3.81}$$

Following simiar arguments in Theorem 3.6, the asymptotic normality of $\widehat{\beta}_n$ given in (3.76) still holds.

If $Z_i, i = 1, \ldots, n$ are treated as fixed instead of i.i.d., we would need more complicated conditions as those in Lai and Ying (1991, 1994) and apply the limit theorems

in the independent but non-identically distributed case.

Although we have assumed that the covariates are time independent, the suggested iterative maximum likelihood method can be applied to time-varying covariates. The cumulative incidence function in (2.1) would be expressed as

$$F_j(t, Z) = 1 - \exp\left(-\int_0^t \exp(Z^T(s)\beta_j) dA_j(s)\right), \quad j = 1, 2.$$

Its derivative with regard to t is

$$\frac{dF_j(t, Z)}{dt} = (1 - F_j(t, Z)) \exp(Z^T(t)\beta_j) dA_j(t), \quad j = 1, 2.$$

Without causing any confusion, the argument Z_i in $F_j(t, Z_i)$ and $\widehat{F}_{nj}(t, \mathbf{b}, Z_i)$ represent the history of Z_i up to time t, i.e. $\{Z_i(s), 0 \leq s \leq t\}$. The algorithm and proof in Chapter 2 and 3 still go through if Z_i is replaced with $Z_i(t)$.

Chapter 4

Simulation Studies

In this chapter, we present the results of simulation studies comparing the iterative maximum likelihood method with the methods of Fine and Gray (1999) and Klein and Andersen (2005). The iterative maximum likelihood method is implemented in Matlab with the optimization routine *fminunc*, which implements a subspace trust region method and is based on the interior-reflective Newton method described in Coleman and Li (1994). Fine and Gray's method is implemented with the R package *cmprsk*; Klein and Andersen's method is implemented based on the code in Klein et al. (2008).

We investigate the performance of the three methods in five scenarios. In all five scenarios, the cumulative incidence functions are given by

$$F_j(t;Z) = 1 - (1 - p_j(1 - e^{-t}))^{\exp(Z\beta_j)}, \qquad j = 1, 2. \tag{4.1}$$

To simulate a failure time and a failure type from these cumulative incidence functions, we first simulate the failure time T according to the total failure probability $F(t;Z) = F_1(t;Z) + F_2(t;Z)$. Then with probability $dF_j(T;Z)/dF(T;Z)$, the cause of failure is the jth cause.

Scenario	p_1	p_2	β_1	β_2	Z	LB	UB	Censor%
I: Moderate Censoring	0.25	0.5	1	-0.5	Bern(0.5)	0.045	4	0.371
II: Heavy Censoring	0.2	0.4	1	-0.5	Bern(0.5)	0.02	1	0.746
III: No censoring	0.6	0.6	1	-0.5	Unif(-1,1)	5	6	0
IV: Large Coef	0.4	0.3	-4	1	Unif(-1,1)	0.045	4	0.348
V:Censoring Dep. on Z	0.38	0.56	1	-0.5	Bern(0.5)		4	0.457

Table 4.1: Model parameters and the average censoring rates in the five scenarios.

In scenarios I, II and V, the covariate Z follows the Bernoulli distribution with a success probability of $1/2$; in scenarios III and IV, Z is uniformly distributed on the interval $(-1, 1)$. In scenarios I, II, III, IV, the censoring times are uniformly distributed on an interval (LB, UB); in scenario V, the censoring distribution depends on the covariate Z. The censoring time is $\min(C^0, UB)$, where C^0 follows the distribution $P(C^0 \leq t|Z) = 1 - \exp(-\frac{5t}{9}e^Z)$. The model parameters and the average censoring rates are listed in Table 4.1.

For each scenario and each sample size of 100, 200 and 500, 1000 parallel simulations are done to calculate the biases, variances, mean square errors of the estimates as well as the vector mean square errors $|\widehat{\beta} - \beta|^2$. The results are reported in Tables 4.2-4.4. In this section, ML, FG and PD denote the iterative maximum likelihood method, Fine and Gray's method and Klein and Andersen's pseudo-value approach respectively. For the iterative maximum likelihood method, $\alpha = 0.16$, $c = 0.1$; for Klein and Andersen's approach, the independent working covariance matrix is used.

In scenarios I and II, ML has smaller variances and mean square errors than FG and PD. The gain in efficiency of ML over FG and PD is especially obvious for sample sizes of 100 and 200. In scenario III, ML has similar variances as FG, both smaller than the variances of PD; however, FG has much larger biases than ML. The advantage of ML over FG and PD is even clearer in scenario IV, in which ML has smaller variances and biases than both FG and PD.

n	Bias			$Var(\widehat{\beta})$			MSE		
	ML	FG	PD	ML	FG	PD	ML	FG	PD
Scenario I: moderate censoring									
	-0.034	0.044	0.058	0.137(0.158)	0.172	0.197	0.138	0.174	0.200
100	0.002	-0.016	-0.020	0.124(0.125)	0.144	0.171	0.124	0.145	0.171
							0.262	0.319	0.371
	-0.037	0.003	0.010	0.073(0.074)	0.081	0.092	0.074	0.081	0.092
200	-0.001	-0.008	-0.014	0.062(0.064)	0.068	0.078	0.062	0.068	0.078
							0.136	0.149	0.170
	-0.015	-0.002	0.001	0.028(0.031)	0.030	0.034	0.028	0.030	0.034
500	0.004	0.002	-0.001	0.025(0.025)	0.026	0.030	0.025	0.026	0.030
							0.054	0.056	0.063
Scenario II: heavy censoring									
	0.043	0.056	0.068	0.243(0.218)	0.342	0.269	0.245	0.345	0.274
200	-0.055	-0.056	-0.072	0.205(0.192)	0.207	0.228	0.208	0.210	0.233
							0.410	0.498	0.435
	0.008	0.010	0.015	0.076(0.077)	0.076	0.081	0.076	0.077	0.081
500	-0.012	-0.013	-0.016	0.072(0.071)	0.072	0.079	0.072	0.072	0.080
							0.147	0.148	0.161

Table 4.2: Scenario I and II: bias, variance and mean square error of estimates using the iterative maximum likelihood method (ML), Fine and Gray's method (FG) and Klein and Andersen's method (PD). For each sample size, the first row is the results of estimates for β_1, the second row is for β_2, the third row is the vector mean square errors. The number in the bracket on each row is the mean of the estimated variances using ML.

n	Bias			$Var(\widehat{\beta})$			MSE		
	ML	FG	PD	ML	FG	PD	ML	FG	PD
Scenario III: no censoring									
	-0.003	-0.170	0.027	0.067 (0.132)	0.072	0.107	0.067	0.100	0.108
100	0.068	-0.173	-0.007	0.066 (0.079)	0.064	0.089	0.071	0.094	0.089
							0.137	0.194	0.197
	0.016	-0.170	0.022	0.035(0.035)	0.036	0.052	0.035	0.065	0.053
200	0.022	-0.177	-0.009	0.032(0.030)	0.033	0.043	0.033	0.064	0.043
							0.066	0.126	0.095
	0.017	-0.173	0.008	0.014(0.015)	0.013	0.019	0.014	0.043	0.019
500	-0.001	-0.179	-0.012	0.012(0.012)	0.012	0.017	0.012	0.045	0.017
							0.026	0.088	0.037
Scenario IV: large coefficient									
	0.064	0.352	-0.195	0.353(0.774)	0.451	0.869	0.357	0.575	0.907
100	-0.052	0.295	0.257	0.228(0.245)	0.171	0.838	0.231	0.258	0.904
							0.588	0.833	1.809
	-0.001	0.422	-0.006	0.173(0.287)	0.225	0.289	0.173	0.402	0.289
200	-0.004	0.310	0.187	0.106(0.106)	0.083	0.106	0.106	0.179	0.141
							0.279	0.581	0.429
	-0.017	0.482	0.058	0.059(0.084)	0.072	0.093	0.059	0.304	0.097
500	-0.010	0.292	0.160	0.037(0.038)	0.029	0.035	0.037	0.114	0.060
							0.097	0.418	0.157

Table 4.3: Scenario III and IV: bias, variance and mean square error of estimates using the iterative maximum likelihood method (ML), Fine and Gray's method (FG) and Klein and Andersen's method (PD). For each sample size, the first row is the results of estimates for β_1, the second row is for β_2, the third row is the vector mean square errors. The number in the bracket on each row is the mean of the estimated variances using ML.

n	Bias			$Var(\widehat{\beta})$			MSE		
	ML	FG	PD	ML	FG	PD	ML	FG	PD
Scenario V: censoring depending on Z									
	0.028	-0.098	0.082	0.159(0.166)	0.162	0.252	0.159	0.171	0.259
100	0.135	-0.232	-0.019	0.187(0.220)	0.227	0.235	0.205	0.281	0.235
							0.364	0.452	0.494
	0.014	-0.125	0.028	0.069(0.076)	0.071	0.104	0.069	0.086	0.105
200	0.094	-0.195	-0.002	0.088(0.098)	0.104	0.107	0.097	0.142	0.107
							0.165	0.229	0.211
	0.006	-0.127	0.015	0.028(0.033)	0.029	0.043	0.028	0.045	0.043
500	0.048	-0.179	0.020	0.033(0.038)	0.040	0.040	0.035	0.071	0.040
							0.063	0.117	0.083

Table 4.4: Scenario V: bias, variance and mean square error of estimates using the iterative maximum likelihood method (ML), Fine and Gray's method (FG) and Klein and Andersen's method (PD). For each sample size, the first row is the results of estimates for β_1, the second row is for β_2, the third row is the vector mean square errors. The number in the bracket on each row is the mean of the estimated variances using ML.

Scenarios III and IV are different from scenarios I and II in an important aspect. If we let $\tau_i = \inf\{t : F_1(t; Z_i) + F_2(t; Z_i) \geq 1\}$, in scenarios I and II, $\tau_i = \infty$ for all i; in scenarios III and IV, $\tau_i < \infty$ for some i and τ_i's may be different for different subjects. For the ith subject, the proportionality of the cumulative incidence functions (4.1) only holds on the interval $[0, \tau_i]$. For example, in scenario IV, for $Z = -0.5$, $\tau = 0.99$; for $Z = 1$, $\tau = \infty$. As shown in Figure 4.1, the CIF of the second risk for $Z = -0.5$ becomes flat for $t > 0.99$, while the CIF for $Z = 1$ is still increasing.

ML does not use any information beyond τ_i's and is still consistent in scenarios III and IV. On the contrary, FG assumes the proportionality to hold until the largest event or censoring time for all subjects; therefore, it has very large biases in scenarios III and IV. PD also suffers the same drawback, but its biases are smaller than FG, although larger than ML in scenario IV. FG appears to be more affected by this drawback than PD.

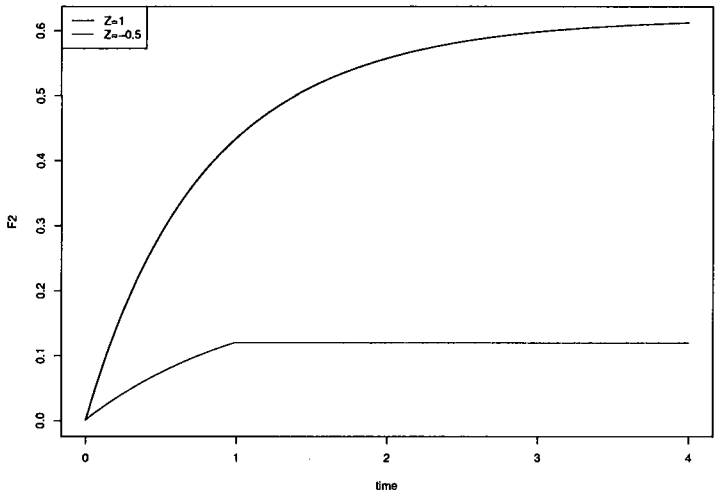


Figure 4.1: Cumulative incidence functions of the second risk for $Z = 1$ (black) and $Z = -0.5$ (red) in scenario IV .

In scenario V, we compare the performance of the three methods when the censoring distribution depends on the covariate. In the R package *cmprsk*, the censoring distribution is assumed to be identical for all subjects and is estimated by the Kaplan-Meier estimate. Hence, it is not surprising that FG has large biases in this scenario. Both ML and PD have small biases as neither of them requires estimating the censoring distribution, but ML has smaller variances than PD.

In summary, ML has better performance, i.e. smaller mean square errors, than FG and PD in all the five scenarios investigated. The estimated variances of the iterative maximum likelihood method are consistent with the simulated variances in all scenarios except scenario IV. In scenario IV, the estimated variance of $\widehat{\beta}_1$ tends to overestimate the true variance. This may be due to the very large absolute value of β_1 ($\beta_1 = -4$).

Chapter 5

Bone Marrow Transplantation Study

Most patients with leukemia who might benefit from a bone marrow transplant lack an HLA-identical sibling donor; hence, there is increasing interest in the use of alternative donors. It is important to know the relative efficacy of alternative donor transplants, especially when there is a choice between an HLA-mismatched related donor and a matched unrelated donor. In this chapter, we compare the outcomes of bone marrow transplants for leukemia using different types of donors. In section 1, we describe the data in this study and illustrate the difference between the Kaplan-Meier estimate and the cumulative incidence function. In section 2, we present the results of regression analysis using the iterative maximum likelihood method, Fine and Gray's method and the Cox proportional hazards model. Finally in section 3, we use the Nelson-Aalen plots to check if the model assumption is valid.

5.1 Data

The data set consists of 415 patients whose transplants were performed at Stanford Hospital and Clinics between 1986 and 1997. Among them, 303 transplants are from HLA identical sibling donors (SIBL), 87 from HLA-matched unrelated donors (URD), 25 from partially mismatched related donors (PMRD). 156 patients were treated for acute lymphoblastic leukemia (ALL), 130 for acute myelogenous leukemia (AML), 130 for chronic myelogenous leukemia (CML). Before transplants, 219 patients were at an early stage of disease (defined as in the first remission or chronic phase); 197 were at an advanced stage of disease (defined as in the second remission, accelerated phase, not in remission or in blast phase). The median age of patients at transplant is 27 years of age. Patients were followed up to five years after transplants; 93 patients relapsed and 139 died in remission.

Since relapse alters the probability of death in remission and death in remission precludes the occurrence of relapse, we consider both the occurrence of relapse and death in remission as absorbing states. In analysis of such multiple-endpoints data, the complement of a Kaplan-Meier estimate (1-KM) is often used to represent the probability of occurrence of a specified endpoint. As pointed out in Gooley et al. (1999), 1-KM is inappropriate and not interpretable when used in the presence of competing risks. 1-KM can only be interpreted as a hypothetical probability that assumes the probability of failure from the cause of interest would not change if competing risks were removed. In other words, 1-KM yields an unbiased estimate only if failure from the cause of interest is independent from failure from other competing risks. In other cases, 1-KM usually overestimates the true probability of failure of a specified type.

A more appropriate and interpretable estimate of the probability of failure of a specified type is the cumulative incidence function (CIF), which is usually obtained

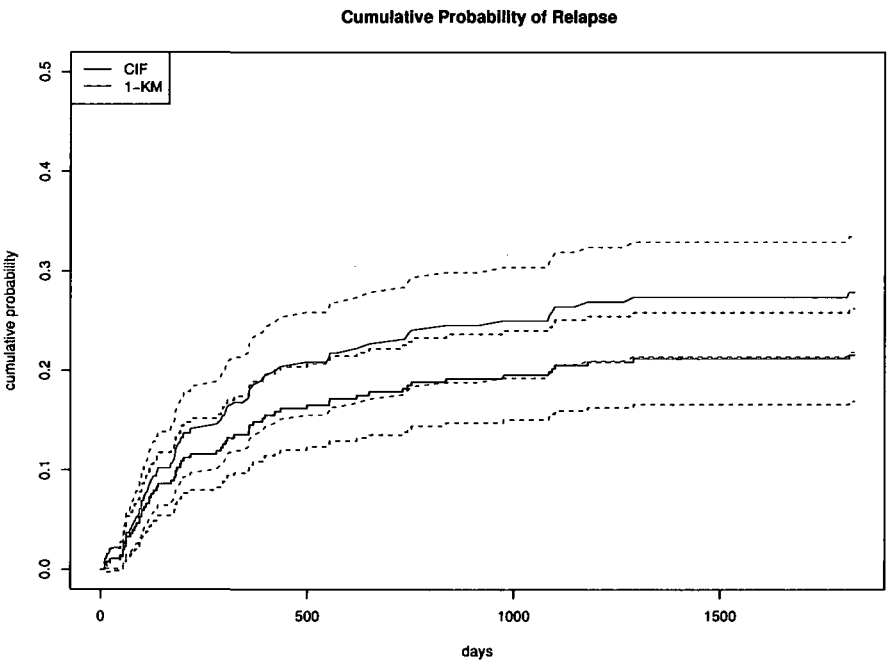


Figure 5.1: Cumulative probability of relapse after transplants from HLA-identical sibling donors. The red curves are 1-KM; the black curves are CIF. The solid lines are the estimates; the dashed lines are the 95% pointwise confidence bands.

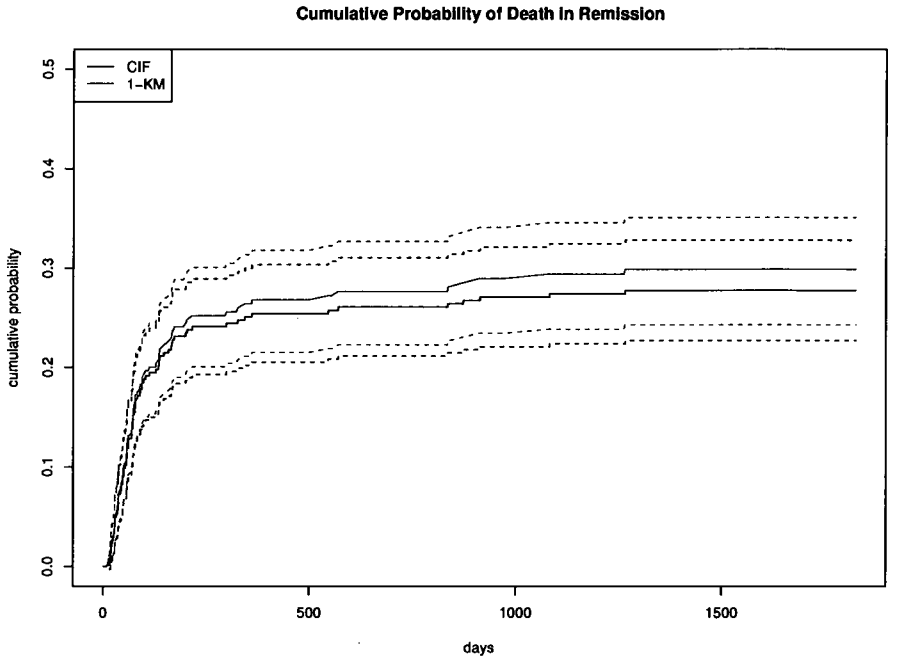


Figure 5.2: Cumulative probability of death in remission after transplants from HLA-identical sibling donors. The red curves are 1-KM; the black curves are CIF. The solid lines are the estimates; the dashed lines are the 95% pointwise confidence bands.

with the Aalen-Johansen estimate in (1.4). Figures 5.1 and 5.2 show the 1-KM and CIF estimates of the probabilities of relapse and death in remission after transplants from HLA-identical sibling donors, 1-KM clearly overestimating CIF.

5.2 Methods and Results

To assess the effects of donor type on the outcome of the transplants, we compare the CIFs of relapse and death in remission stratified by donor type, as shown in Figures 5.3 and 5.4. Donor type appears to have no significant effect on the risk of relapse, as the confidence bands of the CIFs of relapse using different types of donors overlap

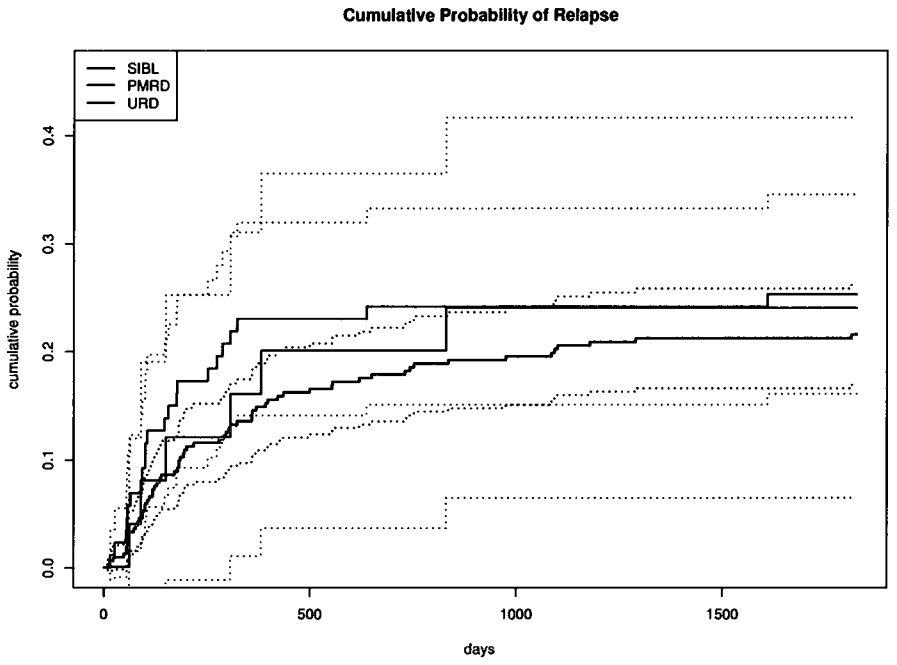


Figure 5.3: CIF of relapse using identical sibling donors (black), partially mismatched related donors (red) and matched unrelated donors (green). The solid lines are the estimates, the dashed lines are the 95% pointwise confidence bands.

with each other. However, the confidence band of the CIF of death in remission using identical sibling donors appears to be lower than those using partially mismatched related donors or matched unrelated donors, suggesting patients with identical sibling donors may have a significantly lower risk of death in remission.

To control for other covariates, we use regression methods to estimate the effects of donor type. Since the relative risks of alternative donors compared with identical sibling donors have been found to be different for chronic and acute diseases in Szydlo et al. (1997), we consider the interaction of the disease type and the donor type. The regression covariates are age at transplant (demeaned), the stage of disease (early or advanced) and a categorical variable representing the disease type and the donor type.

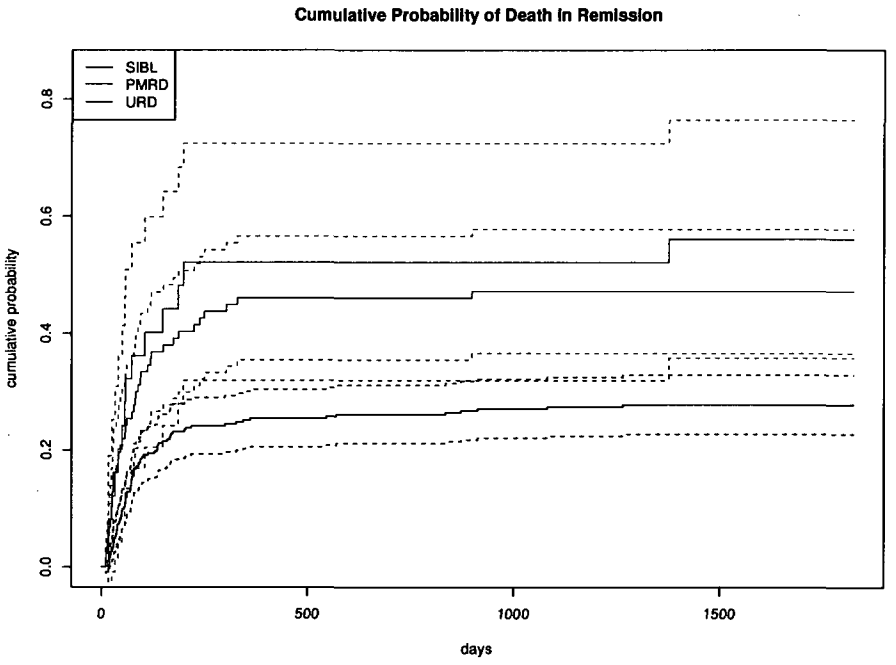


Figure 5.4: CIF of death in remission using identical sibling donors (black), partially mismatched related donors (red) and matched unrelated donors (green). The solid lines are the estimates, the dashed lines are the 95% pointwise confidence bands.

Donor type	Disease	Number of patients
SIBL	Acute	212
PMRD	Acute	17
URD	Acute	56
SIBL	CML	91
PMRD	CML	8
URD	CML	31

Table 5.1: The number of patients in each subgroup with a specified type of donor and disease. Acute stands for ALL or AML; SIBL stands for HLA-identical sibling donor; PMRD stands for partially mismatched related donor; URD stands for matched unrelated donor.

This categorical variable has six levels, a combination of two disease types (acute and chronic) and three donor types. The number of patients in each subgroup is shown in Table 5.1.

We estimate the CIFs of relapse and death in remission using the suggested iterative maximum likelihood method and Fine and Gray's method. As a comparison with the cause-specific hazards formulation, we also estimate the cause-specific hazards of relapse and death in remission using the Cox proportional hazards model. The estimated coefficients, their standard errors and p-values are shown in Table 5.2. The reference group is acute patients with HLA-identical sibling donors (SIBL-Acute), age 26, and at an early stage of disease.

The subgroup URD-CML has a much lower risk of relapse compared to the reference group SIBL-Acute. It has 31 patients: 22 died in remission; 9 survived beyond five years; none of them relapsed. All three methods give a very large negative regression coefficient; however, except Fine and Gray's method, the other two methods have very large estimated standard errors and their p-values of the Wald test are close to 1. As Professor Philip Lavori pointed out in a personal email communication, the Wald test is not robust for very large coefficients. Instead, the likelihood ratio test is more reliable in this case. The p-values for URD-CML by the likelihood ratio test,

		PMRD Acute	URD Acute	SIBL CML	PMRD CML	URD CML	Advanced Stage	Age
					Relapse			
$\hat{\beta}_R$	ML	-0.196	0.111	-0.457	-0.342	-33.961	0.914	-0.003
	FG	-0.120	0.387	-0.438	-0.357	-9.954	1.138	0.002
	Cox	0.326	0.628	-0.544	0.225	-17.074	1.295	0.012
SE	ML	0.458	0.246	0.394	1.098	8.484e+06	0.298	0.007
	FG	0.472	0.277	0.350	1.017	0.262	0.258	0.008
	Cox	0.475	0.265	0.364	1.030	2381.688	0.268	0.008
p-value	ML	0.669	0.651	0.246	0.755	0.000*	0.002	0.606
	FG	0.799	0.162	0.211	0.726	0.001*	0.000	0.823
	Cox	0.493	0.018	0.134	0.827	0.003*	0.000	0.119
					Death in Remission			
$\hat{\beta}_D$	ML	1.046	0.376	-0.252	0.870	1.005	0.071	0.027
	FG	1.063	0.719	-0.215	0.878	1.076	0.120	0.031
	Cox	1.110	0.857	-0.264	0.846	1.020	0.239	0.033
SE	ML	0.384	0.237	0.253	0.489	0.270	0.207	0.006
	FG	0.347	0.283	0.245	0.492	0.272	0.201	0.006
	Cox	0.367	0.272	0.250	0.478	0.269	0.195	0.007
p-value	ML	0.006	0.112	0.319	0.075	0.000	0.730	0.000
	FG	0.002	0.011	0.380	0.074	0.000	0.550	0.000
	Cox	0.003	0.002	0.290	0.077	0.000	0.219	0.000

Table 5.2: Estimated coefficients for relapse ($\hat{\beta}_R$) and death in remission ($\hat{\beta}_D$), their standard errors (SE) and p-values using the iterative maximum likelihood method (ML), Fine and Gray's method (FG) and the Cox proportional hazards model (Cox). The reference group is acute patients with identical sibling donors, age 26, at an early stage of disease. p-values with an asterisk are computed by the likelihood ratio test; the other p values are computed by the Wald test.

given in Table 5.2, are highly significant in all three methods.

The estimated coefficients and their p-values by the three methods are in general consistent except for the subgroup URD-Acute. For relapse, ML and FG give insignificant p-values, while Cox has a significant p-value. For death in remission, both FG and Cox have significant p-values, but ML has a slightly insignificant p-value. The reason for this discrepancy needs further investigation.

To summarize the findings in Table 5.2, patients at an advanced stage of disease have a significantly higher risk of relapse compared to those at an early stage of disease. Except the subgroup URD-CML, the other subgroups have no significant difference in the risk of relapse compared to the reference group SIBL-Acute. Age at transplant has no significant effect on the risk of relapse either.

Older patients have a slightly but significantly higher risk of death in remission compared to younger patients. The subgroups URD-CML and PMRD-Acute both have a significantly higher risk of death in remission compared to the reference group. The subgroups URD-Acute and PMRD-CML have marginally insignificant p-values, but this may change if the sample size increases. The stage of disease has no significant effect on the risk of death in remission.

The iterative maximum likelihood method converges after 331 iterations if the initial values for the coefficients of all covariates are zero in both models for relapse and death in remission. If the starting value is -1 for the subgroup URD-CML in the model for relapse, the algorithm will converge after 109 iterations. This suggests that a preliminary look at the data and a reasonable guess of the coefficients may help the iterative algorithm to converge much faster.

To see if the results will change for a different choice of the baseline reference group, we reanalyze the data using CML patients with identical sibling donors (SIBL-CML), aged 26 and at an early stage of disease as the reference group. The estimated coefficients, their standard errors and p-values are reported in Table 5.3. These results

		PMRD Acute	URD Acute	SIBL Acute	PMRD CML	URD CML	Advanced Stage	Age
Relapse								
$\hat{\beta}_R$	ML	-0.165	0.172	0.026	-0.122	-30.599	1.042	-0.006
	FG	0.318	0.825	0.438	0.081	-9.516	1.138	0.002
	Cox	0.870	1.172	0.544	0.770	-16.529	1.295	0.012
SE	ML	0.583	0.392	0.363	1.117	1.681e+06	0.267	0.007
	FG	0.572	0.430	0.350	1.054	0.370	0.258	0.008
	Cox	0.588	0.434	0.364	1.048	2381.688	0.268	0.008
p-value	ML	0.778	0.662	0.943	0.913	0.000*	0.000	0.418
	FG	0.578	0.055	0.211	0.939	0.009*	0.000	0.823
	Cox	0.139	0.007	0.134	0.463	0.030*	0.000	0.119
Death in Remission								
$\hat{\beta}_D$	ML	1.239	0.592	0.173	1.083	1.213	0.099	0.027
	FG	1.279	0.934	0.215	1.094	1.291	0.120	0.031
	Cox	1.374	1.122	0.264	1.110	1.284	0.239	0.033
SE	ML	0.443	0.289	0.257	0.502	0.298	0.199	0.006
	FG	0.390	0.352	0.245	0.491	0.283	0.201	0.006
	Cox	0.419	0.342	0.250	0.486	0.296	0.195	0.007
p-value	ML	0.005	0.041	0.501	0.031	0.000	0.619	0.000
	FG	0.001	0.008	0.380	0.026	0.000	0.550	0.000
	Cox	0.001	0.001	0.290	0.022	0.000	0.219	0.000

Table 5.3: Estimated coefficients, their standard errors (SE) and p-values. The reference group is CML patients with identical sibling donors, age 26, at an early stage of disease. p-values with an asterisk are computed by the likelihood ratio test; the other p values are computed by the Wald test.

in general agree with those in Table 5.2, except that PMRD-CML has a significantly higher risk of death in remission compared to SIBL-CML in all three methods, while its relative risk compared to SIBL-Acute is marginally insignificant with p-values around 0.07 in all three methods.

5.3 Model Diagnostic

The iterative maximum likelihood method and Fine and Gray's method assume a proportional hazards model for the cumulative incidence function, as given in (2.1). To check if the model assumption is valid, we use the Nelson-Aalen plot, suggested in Kay (1977) (also see Andersen et al. 1992, page 555).

The Nelson-Aalen plot is based on the residual $\widetilde{T}_{ij} = \Lambda_{ij}(T_i)$, where Λ_{ij} is the cumulative cause-specific hazard

$$\Lambda_{ij}(s) = \int_0^t \frac{1 - F_j(s; Z_i)}{1 - F_1(s; Z_i) - F_2(s; Z_i)} \exp(Z_i^T \beta_j) dA_j(s).$$

It is the residual in the sense of Cox and Snell (1968), further discussed by Kay (1977). Kay noted that in the uncensored case the true values $\widetilde{T}_{ij}, i = 1, \ldots, n$, for each j, constitute a sample of n independent unit exponential variables. If some of the observations are right censored, one may treat the corresponding residuals as right-censored. Kay (1977) suggested using the Nelson-Aalen plot, possibly within a number of strata, to check the approximate exponentiality of the residuals. If the residuals are truly unit exponentially distributed, the Nelson-Aalen plot should be very close to the unit slope.

The Nelson-Aalen plots of the residuals of relapse and death in remission stratified by the stage of disease are displayed in Figures 5.5 and 5.6. Although there is some deviation of the Nelson-Aalen plots from the unit slope, the unit slope lies within the

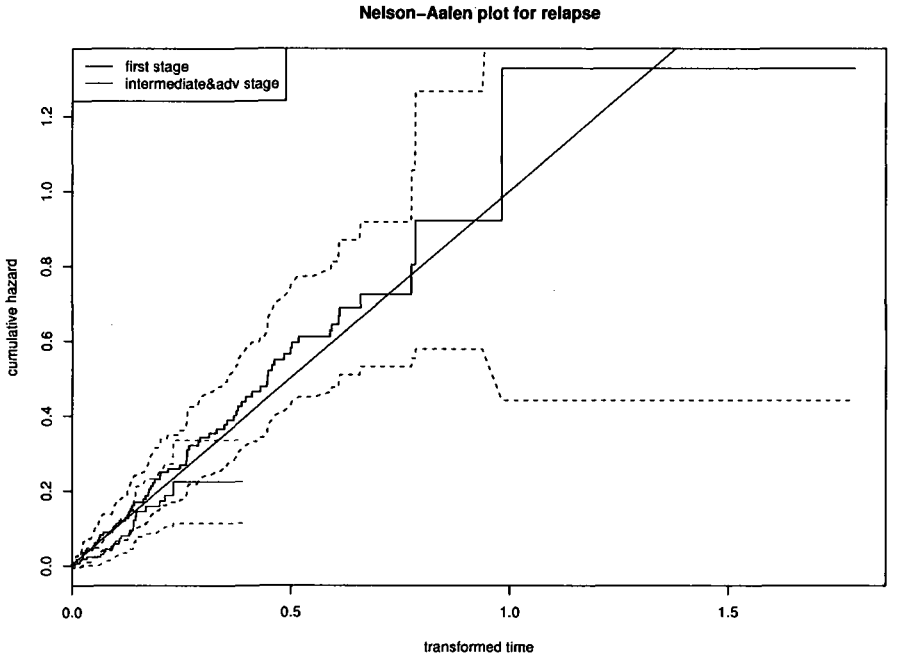


Figure 5.5: Nelson-Aalen plot of the residuals of relapse for patients at an advanced stage (green) vs those at an early stage (black). Red line is the unit slope. Dashed lines are the 95% pointwise confidence bands.

95% pointwise confidence bands. This is also true for the Nelson-Aalen plots stratified by the other covariates, which are not shown here. Hence, we conclude that there is no significant evidence against the proportional hazards model assumption.

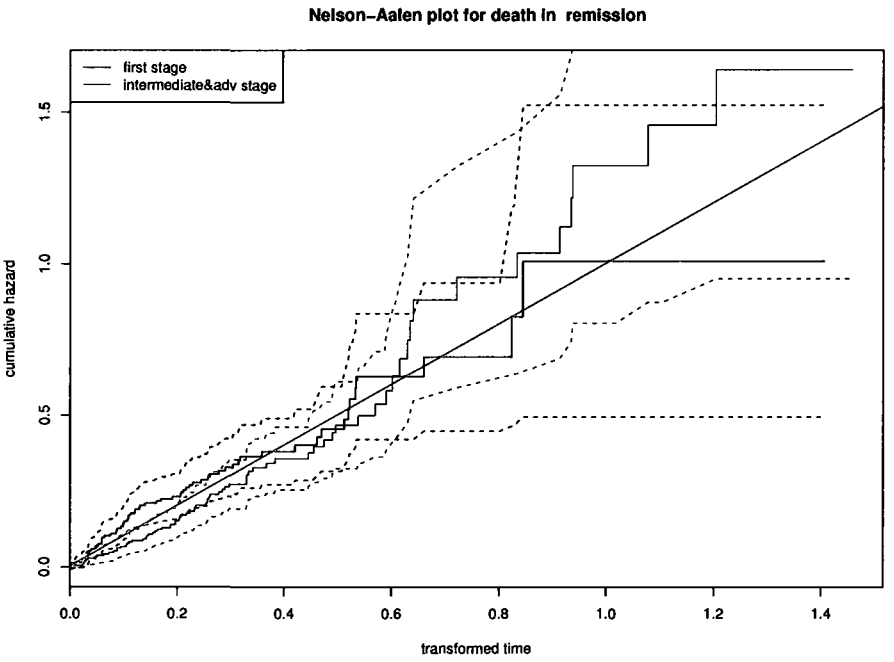


Figure 5.6: Nelson-Aalen plot of the residuals of death in remission for patients at an advanced stage (green) vs those at an early stage (black). Red line is the unit slope. Dashed lines are the 95% pointwise confidence bands.

Chapter 6

Conclusion

We propose an iterative maximum likelihood method to model the cumulative incidence function directly. It consists of iterating between two steps: the first step is to estimate the baseline subdistribution hazards with the Nelson-Aalen-type estimate using the current estimate of the regression coefficients; the second step is to update the estimate of the coefficients by maximizing the log-likelihood with the baseline hazards fixed. We derive the asymptotic normality of the maximum likelihood estimate by representing it as the solution to a system of estimating equations. The iterative maximum likelihood estimate has smaller mean square errors than the estimates of Fine and Gray (1999) and Klein and Andersen (2005) in the five scenarios investigated by simulation. The proposed method is further illustrated on a real dataset to compare the risks of relapse and death in remission after bone marrow transplants using different types of donors.

Simulation studies in the literature have shown that the methods of Scheike and Zhang (2008) and Fine (2001) have similar performance as the method of Fine and Gray (1999). Combined with our extensive simulation study, the proposed method is shown to outperform these other methods.

The proposed method has three major advantages. First, it can be shown to

be asymptotically efficient. Although Fine and Gray's method has similar variances on complete data, as shown in scenario III, it has larger variances than our method on censored data. Klein and Andersen's method also has larger variances than our method in all the scenarios investigated.

Second, it does not require estimating the censoring distribution. The IPCW-based methods would be biased if the censoring distribution is misspecified. Although Klein and Andersen (2005) avoid estimating the censoring distribution by resorting to pseudo-values, their method has to pay a price in efficiency loss, especially for moderately censored or complete data.

Furthermore, the proposed method has a more robust model assumption. Fine and Gray's and Klein and Andersen's methods require the model assumptions to hold until the largest event or censoring time for all subjects. They are biased in certain situations when some subjects fail from either cause with probability 1 before the largest event or censoring time; however, the iterative maximum likelihood method is still consistent in this case.

The iterative procedure may sometimes fail to converge for certain initial values, but convergence is always guaranteed if the initial values are within a small neighborhood of the true parameters. In case that convergence fails, we suggest trying other initial values or starting with a $\sqrt{n}$-consistent preliminary estimate.

Part II

Mortgage Prepayment and Default: Modeling and Applications

Chapter 7

Introduction

7.1 Subprime Crisis

Many U.S. mortgages issued in recent years were made to subprime borrowers. These borrowers usually have a weak credit history, limited income and FICO credit scores below 680 on a scale that ranges from 300 to 850. Subprime loans, which have a much higher rate of default than prime mortgage loans and carry higher mortgage rates, accounted for about 21% of all mortgage originations from 2004 to 2006, up from 9% from 1996 through 2004. As of March 2007, the value of U.S. subprime mortgages was estimated at $1.3 trillion.

When U.S. house prices began to decline in mid 2006, mortgage delinquencies soared. By October 2007, approximately 16% of subprime adjustable rate mortgages (ARM) were either 90-days delinquent or in foreclosure, roughly triple the rate of 2005. By January 2008, the delinquency rate had risen to 21%, and by May 2008 it was 25%. Securities backed by subprime mortgages, widely held by financial firms, lost most of their value. This resulted in a large decline in the capital of many banks and led to a global credit crunch.

The causes of this crisis are varied and complex. An important cause was the

boom in the housing market before 2006, when low interest rates and large inflows of foreign funds created easy credit conditions for a number of years prior to the crisis, thus fueling a housing market boom. For example, between 1997 and 2006, the price of a typical American house increased by 124%. A major contributor to the increase in the overall demand for housing was subprime mortgage lending. The house price and credit explosion led to a building boom and eventually to a surplus of unsold homes. As a result, U.S. housing prices peaked and began to decline in mid-2006. By September 2008, the average U.S. housing prices had declined by over 20% from their mid-2006 peak. As of March 2008, 10.8% of all homeowners had negative equity in the houses they owned. Consequently, many subprime borrowers of adjustable rate mortgages began to default after the rates were reset, because they could not use refinancing to escape higher monthly payments.

Another important cause of the crisis was in the securitization practices, which involve the pooling of financial assets, especially those for which there is no liquid secondary market, such as mortgages. The pooled assets are transferred to a structured investment vehicle and serve as collateral for new financial assets issued by the vehicle. While in the traditional mortgage model credit risk is retained by the bank originating the loan, with securitization, it is transferred to third-party investors. The total amount of mortgage-backed securities (MBS) almost tripled to $7.3 trillion between 1996 and 2007. On one hand, securitization allowed issuers to easily generate funding for new loans; on the other hand, it led to weak underwriting standards and irresponsible lending as each link in the mortgage chain made a profit while passing the associated risk to the next link in the chain.

A third important cause of the crisis is attributed to the increasingly complex and opaque financial products. These finanical products, instead of distributing the credit risk, hid the true risk from view. As the noted economist Arnold Kling said at congressional hearings on the collapse of Freddie Mac and Fannie Mae in December

2008, a highly-risky loan could be "laundered" by Wall Street and return to the banking system as a highly-rated security for sale to investors.

To summarize the causes of the subprime crisis, I quote the following paragraph in *the Declaration of the Summit on Financial Markets and the World Economy* by leaders of the G 20 countries dated November 15, 2008 :

"During a period of strong global growth, growing capital flows, and prolonged stability earlier this decade, market participants sought higher yields without an adequate appreciation of the risks and failed to exercise proper due diligence. At the same time, weak underwriting standards, unsound risk management practices, increasingly complex and opaque financial products, and consequent excessive leverage combined to create vulnerabilities in the system. "

7.2 Mortgage Valuation Models

Given the complexity of mortgage-related instruments, quantitative models to valuate MBS play a vital role in pricing MBS and assessing their risks. These mortgage valuation models have evolved along two related, but still separate, paths.

The first path views mortgages as long-term bonds issued by borrowers with embedded call (prepayment) and put (default) options. Mortgagors exercise the options optimally to minimize the mortgage value and maximize their own wealth. This line of research includes the models developed by Dunn and McConnell (1981a,b). Individuals can increase their wealth by defaulting on a mortgage when the market value of the mortgage equals or exceeds the value of the house and by prepaying the mortgage when the prevailing market rate is below the mortgage contract rate.

These value-minimizing models, although theoretically attractive, have several disadvantages. First, there is apparent suboptimal behavior of borrowers. They are not equally astute in exercising the options. Some borrowers do not default when the

mortgage value exceeds the house value; some do not prepay when the market rate is below the contract rate. Secondly, default and prepayment are not purely financial decisions. Borrowers face many complex constraints, such as residential immobility. Another technical drawback of these models is that they are not immediately applicable to the analysis of collateralized debt obligation (CDO) tranches. The hierarchical structure of CDO requires the knowledge of prior mortgage payments, while the value-minimizing models are solved by a backward finite difference procedure.

The second path, pioneered by Schwartz and Torous (1989, 1992), relies upon statistical methods to characterize the mortgagor's default and prepayment behavior. It is assumed that at each point in time there exist probabilities of prepayment and default depending upon the risk-free interest rate, the mortgaged house value and certain loan-specific variables. Under assumptions on the dynamics of the risk-free interest rate and the mortgaged house value, standard arbitrage arguments give the partial differential equation (PDE) satisfied by the value of any MBS. Monte Carlo methods can be applied to solve the PDE by simulating the risk-neutral paths of the interest rate and the mortgaged house value. The cash flows to the security holders are determined given the estimated prepayment and default probabilities, and the average of the discounted present value of future cash flows in many realizations gives the price of any MBS.

The price given by the model is often not the same as the observed market price. To equate the model price with the market price, a constant spread is added to all the risk-neutral spot interest rates in the Monte Carlo simulation. This spread is called the *option-adjusted spread* (OAS), which may be interpreted as the risk premium not explained by the risk-free interest rate. A detailed description about the origin and interpretation of OAS can be found in Kupiec and Kah (1999).

The basic ingredients of the second path are the estimated prepayment and default probabilities. The next section provides a brief review of methods in the literature to

estimate these probabilities.

7.3 Modeling Prepayment and Default Probabilities

Mortgage data involve competing risks. There are two end-points, prepayment and default, which are mutually exclusive. The occurrence of one end-point precludes that of the other. Since mortgage data are usually observed monthly or quarterly, for such discrete data with numerous ties, the popular Cox proportional hazards model is not immediately applicable. Instead, several parametric methods have been suggested to model the hazards.

Schwartz and Torous (1989) suggested to model the prepayment hazard under the proportional hazards assumption and model the baseline hazard with a log-logistic function:

$$\lambda_0(t) = \frac{\gamma p(\gamma t)^{p-1}}{1+(\gamma t)^p}.$$

It was implemented using pool-level covariates: the difference between the contract rate and long-term Treasury rate, the fraction of the pool outstanding and seasonality. Their model is parsimonious, but it imposes restrictions on the shape of the baseline hazard.

Deng et al. (2000) proposed to use a parameter γ_{jk} for each discrete time interval t_k and model the hazards with the Gumbel link:

$$\lambda_j(t_k, Z(t_k)) = 1 - \exp(-\exp(\gamma_{jk} + \beta_j Z(t_k) + \log(\eta_j))), \quad j = 1, 2.$$

η_1 and η_2 are unobserved heterogeneities associated with the hazard functions for default and prepayment respectively. They are distributed on two or three mass

points, representing two or three populations among borrowers. All the parameters, including γ's, β and the distribution of η_j, are estimated by maximizing the full likelihood. Although it allows more flexibility in modeling the baseline hazards, their method is computationally very expensive for long-duration data.

Given the inherent discreteness of mortgage data, a more appropriate method is the multilogit regression model. In each month the borrower makes a decision out of three choices, to continue payment, default or prepay. Let $Y(t)$ be 0, 1, 2 according to whether the borrower continues payment, defaults or prepays in month t, and let $Z(t)$ denote the vector of covariates in month t. The multilogit model estimates the conditional probability of each choice in any given month by

$$P(Y(t) = j | Y(t-1) = 0, Z(t)) = \frac{\exp(\beta_{0j} + \beta_j^T Z(t))}{1 + \sum_{l=1}^{2} \exp(\beta_{0l} + \beta_l^T Z(t))}, \quad j = 1, 2$$

Calhoun and Deng (2002) implemented the multilogit regression model on conforming fixed-rate and adjustable-rate residential mortgages between 1979 and 1993. Their covariates include loan age, dummy variabls for origination years, the loan-to-value ratio (LTV), seasonality, occupancy status of the property, the original loan size, the ratio of 10-year Constant Maturity Treasury rate to the 1-year Constant Maturity Treasury rate, mortgage premium value and the probability of negative equity based on estimated drift and volatility of house prices. Since prepayment and default probabilities depend on loan age in a clearly nonlinear way, they include a quadratic function of loan age.

7.4 Outline

To incorporate possible nonlinear dependency on the covariates and interactions among them, especially interactions between loan age and the other covarites, we

propose to extend the multilogit model to neural networks. In Chapter 8, we describe the neural network model to estimate the conditional prepayment and default probabilities. In Chapter 9, we apply both the multilogit regression model and the neural network model to a large subprime mortgage dataset. We predict the monthly prepayment and default rates on three test datasets of vintages 2004, 2005 and 2006 and compare the prediction accuracy of the neural network model with that of the multilogit model. In Chapter 10, we summarize the advantages of the neural network model over the multilogit model as well as findings from the analysis of the data.

Chapter 8

Neural Network Model

In this chapter, we provide a brief description of fitting the neural network model and some issues involved based on Hastie et al. (2001, Chapter 11). The most widely used "vanilla" neural network, sometimes called the single hidden layer network, is a regression or classification model, typically represented by a network diagram as in Figure 8.1. It applies both to regression or classification. For J-class classification, there are J output units, with the jth unit modeling the probability of class j. We will use a single hidden layer network to model the conditional probability of continuing payment, default or prepayment. In our case, $J = 3$.

The inputs to the network are the covariates $Z(t)$; the hidden neuron X_m is created as a function of linear combinations of the inputs. Recall that $Y_i(t)$ being $0, 1, 2$ are coded for the ith borrower's decision to continue payment, default or prepay in month t, the targets are modeled as a function of linear combinations of the X's.

$$X_m(t) = \sigma(\alpha_{0m} + \alpha_m^T Z(t)), \quad m = 1, \ldots, M,$$

$$R_j(t) = \beta_{0j} + \beta_j^T X(t), \quad j = 0, \ldots, J-1,$$

$$f_j(Z(t)) = P(Y(t) = j | Y(t-1) = 0, Z(t)) = g_j(R(t)), \quad j = 0, \ldots, J-1,$$

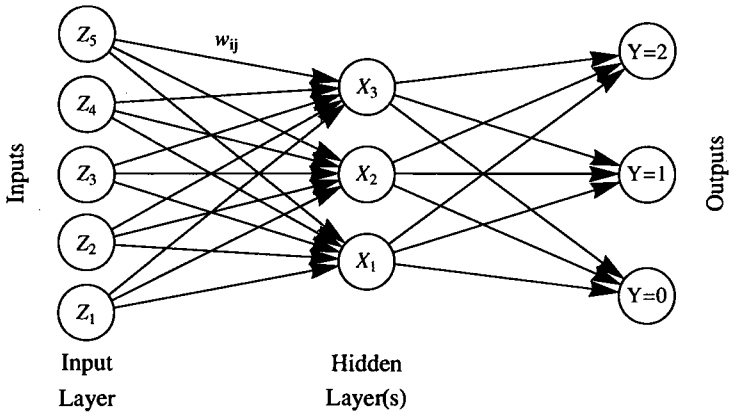


Figure 8.1: Schematic of a single hidden layer, feed-forward neural network.

where $X(t) = (X_1(t), \ldots, X_M(t))^T$, $R(t) = (R_0(t), \ldots, R_{J-1}(t))^T$. The activation function $\sigma(v)$ is usually chosen to be the sigmoid $\sigma(v) = 1/(1+e^{-v})$ or the Gaussian radial basis function. For classification, $g_j(\cdot)$ is usually chosen to be the softmax function

$$g_j(R) = \frac{e^{R_j}}{\sum_{l=0}^{J-1} e^{R_l}},$$

which is exactly the transformation used in the multilogit model. If σ is linear, the entire model collapses to a linear model in the inputs. By introducing the nonlinear transformation σ, a neural network can be thought of as a nonlinear extension of the linear model.

The complete set of unknown parameters, called *weights* and denoted by θ, consists of

$$\{\alpha_{0m}, \alpha_m; m = 1, 2, \ldots, M\},$$

$$\{\beta_{0j}, \beta_j; j = 0, 1, \ldots, J-1\},$$

with $M(p+1) + J(M+1)$ parameters in total.

We use cross-entropy as the error function:

$$L(\theta) = -\sum_{i=1}^{n}\sum_{t=1}^{T_i} \log f_{Y_i(t)}(Z(t)), \tag{8.1}$$

where T_i is the last month of observation for the ith loan. The generic approach to minimizing $L(\theta)$ is by gradient descent, called back-propagation in this setting. The gradient can be computed by a forward and backward sweep over the network, keeping track only of quantities local to each unit. A detailed description about back-propagation can be found in Hastie et al. (2001, page 354). Although back-propagation is simple and local, it can be very slow and hence is usually not the method of choice. Better approaches to fitting include conjugate gradients and variable metric methods, which avoid explicit computation of the second order derivative matrix while still providing faster convergence.

Neural networks are generally overparametrized and will overfit the data at the global minimum of $L(\theta)$. In early developments of neural networks, an early stopping rule was used to avoid overfitting. A more explicit method for regularization is *weight decay*, which is analogous to ridge regression for linear models. Hastie et al. (2001, page 356) suggested adding a penalty term to the error function

$$L(\theta) + \gamma\left(\sum_{jm}\beta_{jm}^2 + \sum_{ml}\alpha_{ml}^2\right),$$

where $\gamma \geq 0$ is a decay parameter. Larger values of γ will shrink the weights towards zero; typically cross-validation or a validation dataset is used to choose γ.

The scaling of the inputs determines the effective scaling of the weights in the input layer; therefore, it is best to standardize all inputs to have mean zero and standard deviation one. This ensures all inputs are treated equally in the regularization process.

Generally, it is better to have too many hidden units than too few. With too few

hidden units, the model might not have enough flexibility to capture the nonlinearities in the data; with too many hidden units, the weights can be shrunk toward zero with appropriate regularization. As stated in Hastie et al. (2001, page 358), it is most common to use a reasonably large number of units and train them with regularization. Typically the number of hidden units is in the range of 5 to 100, the number increasing with the number of inputs and the number of training cases.

Since the error function $L(\theta)$ is nonconvex with many local minima, the final solution is quite dependent on the choice of starting weights. One could try a number of randomly generated starting configurations, and choose the solution with the lowest error. A better approach is to use the average predictions over the collection of networks as the final prediction (Ripley 1996; Hastie et al. 2001). With standardized inputs, the starting weights are typically in the range $[-0.7, 0.7]$.

Chapter 9

Empirical Analysis

In this chapter, we apply both the neural network model and the multilogit model to a large subprime mortgage dataset. In section 1, we describe the data and the covariates used in the two regression models. In section 2, we interpret the effects of these covariates. In section 3, we use the two models to predict monthly prepayment and default rates on four test datasets and compare their prediction accuracy.

9.1 Data

The empirical analysis is based upon loan-level mortgage data consisting of 130,000 first-lien 2-28 adjustable rate subprime loans originated between 2004 and 2006, which were randomly chosen from the database maintained by *LoanPerformance* [1]. These loans have initial mortgage rates for the first two years, then the rates are reset according to some index, such as 1-year LIBOR, plus a margin. They usually have prepayment penalty for the first two years. For each loan, the available information includes the year and month of origination and termination (if it has been closed),

[1] *LoanPerformance* is a self-standing division of First American CoreLogic, Inc.; it creats and maintains the industry's largest mortgage securities and servicing databases.

Loan Purpose	Purchase	Refinance
	44.1%	55.9%
Income Documentation	Full	Limited or No
	66.5%	33.5%
Occupancy Status	Owner	Investor
	89.7%	10.3%

Table 9.1: The distributions of the categorical covariates in the data.

Covariate	5%	25%	50%	75%	95%
FICO	517	563	603	641	703
combined LTV (%)	62	80	90	99	100
loan size(thousand)	57	92	131.92	195.20	355.23
initial mortgage rate(%)	6.000	6.990	7.875	8.800	10.250

Table 9.2: The quantiles of FICO, LTV, loan size and initial mortgage rate in the data.

indicator of prepayment or default. The loan-specific covariates include the borrower's credit score FICO when the loan was originated, the combined loan-to-value ratio (LTV), the original loan amount (loan size) and the initial mortgage rate. Also available are three categorical covariates: loan purpose, representing whether the loan is for purchase or refinance; income documentation, representing whether the borrower has full or no documentation of income; occupancy status, representing whether the mortgaged home is owner occupied or for investment. Tables 9.2 and 9.1 show the distributions of the continuous and categorical covariates in the data.

Besides these loan-specific covariates, two macro-economic variables are also included, the Federal Housing Finance Agency (FHFA) MSA House Price Index (HPI)[2] and 1-year Constant Maturity Treasury rate (CMT). Figure 9.1 shows the average HPI in the data from 2004 to 2007. House prices increased dramatically from 2004 to 2006, peaked in mid-2006 and started to decline in 2007. Figure 9.2 shows the 1-year

[2]HPI is a weighted, repeat-sales index, measuring the average price changes in repeat sales or refinancings on the same properties.

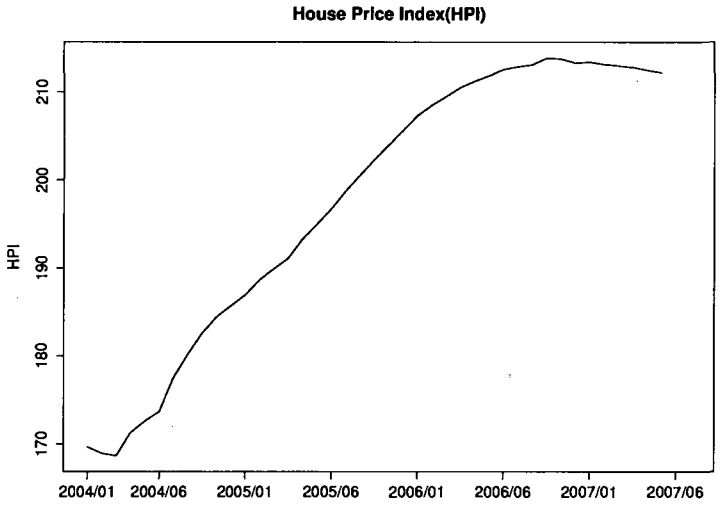


Figure 9.1: Average HPI 2004-2007.

CMT rate for the same period.

These two macro-economic variables do not directly enter into the regression model. Instead they are transformed into the mortgage premium value (MP) and the equity position(EQ), as suggested in Calhoun and Deng (2002). The call (pre-payment) option value of the mortgage is a function of the difference between the present value of the anticipated future stream of mortgage payments discounted at the current market rate, and the present value of the mortgage discounted at the current contract rate. Deng et al. (1996) suggested approximating the call option value using the relative spread between the current mortgage contract rate and the prevailing market rate. As the prevailing market rate is not available in the data and is highly correlated with the risk-free interest rate, we use 1-year CMT rate to proxy

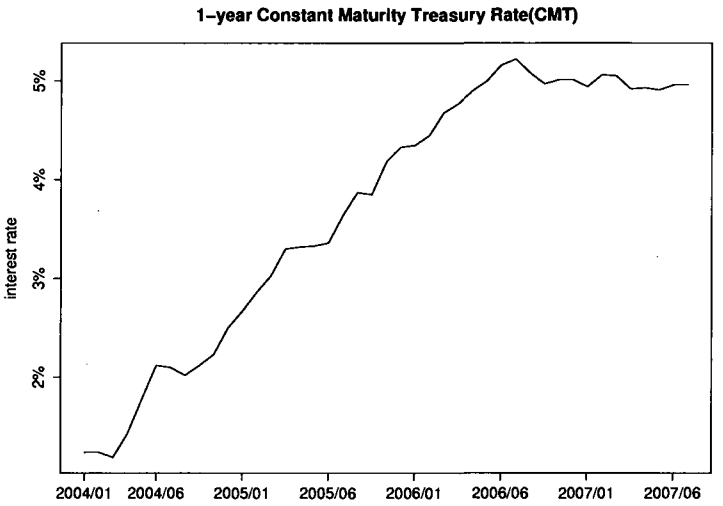


Figure 9.2: 1-year CMT rate 2004-2007.

the prevailing market rate and approximate the mortgage premium value by

$$\begin{aligned} \text{MP} &= 1 - \frac{\text{prevailing mortgage rate}}{\text{mortgage contract rate}} \\ &\approx 1 - \frac{\text{CMT}}{\text{mortgage contract rate}}. \end{aligned}$$

A similar approximation can be found in Schwartz and Torous (1989), where long-term Treasury rates are used to proxy refinancing rates.

The equity position of the borrower is determined by the difference between the market value of the property securing the loan and the unpaid mortgage balance. Since periodic observations on the values of individual properties are not available, we use the original loan size divided by the loan-to-value ratio and multiplied by the

appreciation in HPI to approximate a property's market value:

$$\begin{aligned} \text{EQ} &= \frac{\text{current market value of property}}{\text{unpaid balance}} - 1 \\ &\approx \frac{\text{HPI(t)/HPI(0)} \times \text{loan size/LTV}}{\text{unpaid balance}} - 1. \end{aligned}$$

We implement both the multilogit model and the neural network model, with covariates including loan age, loan size, FICO, mortgage premium value, equity position, loan purpose, income documentation and occupancy status. The training sample consists of 30,000 loans originated from the beginning of 2004 to the end of 2005 and observed until April 2006. For the multilogit model, we include a natural cubic spline in loan age with knots at 10 and 20 months to capture nonlinear dependency on loan age. Other placements of the knots do not affect the results significantly. Unlike fitting the neural network model, the likelihood in the multilogit model is not penalized, since the multilogit model has much fewer parameters than the neural network model and does not tend to overfit the data.

In the neural network model, all the continuous covariates, including loan age, are standardized before they are entered into the model. Two parameters need to be chosen, the number of hidden neurons and the decay parameter γ. Following the common practice described in Chapter 8, we put down 20 hidden neurons to allow for enough flexibility and use an independent validation dataset of 5889 loans to choose the optimal decay parameter. The neural network model is trained with six randomly generated starting configurations of the weights; the log-likelihoods of the six networks on the validatation dataset are computed and shown in Figure 9.3. $\gamma = 0.15$ gives consistently higher log-likelihoods than the other values of γ in the range from 0.03 to 0.18; hence it is chosen as the optimal decay parameter. The average of the predictions from the six networks is used as the final prediction. Both the multilogit model and the neural network model are implemented with the R

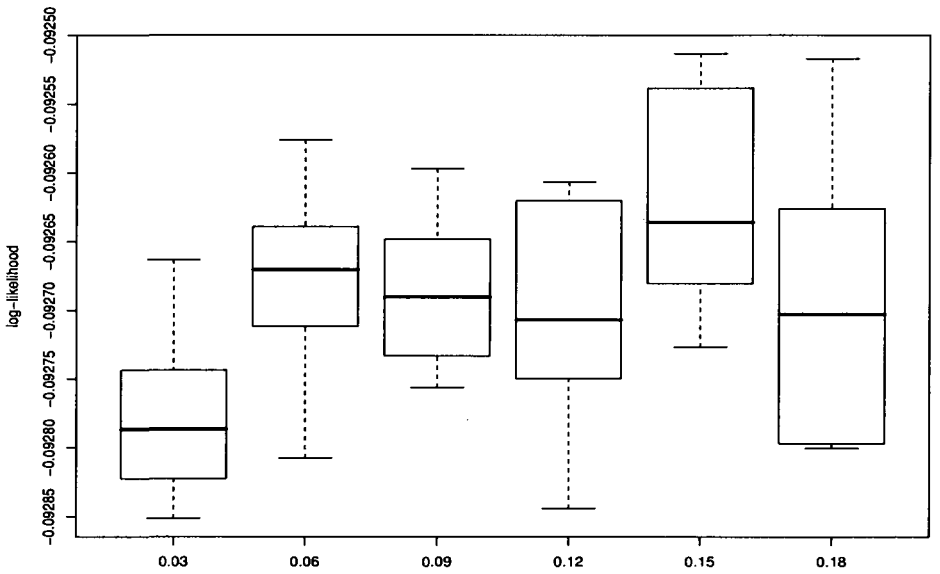


Figure 9.3: Boxplots of the log-likelihoods on the validation dataset for different values of the decay parameter γ from 0.03 to 0.18.

Covariate	Default			Prepayment		
	Coefficient	SE	p Value	Coefficient	SE	p Value
purpose:refinance	-0.303	0.060	0.000	0.199	0.029	0.000
document:no doc	0.432	0.060	0.000	0.105	0.028	0.000
occupancy:investor	0.778	0.084	0.000	0.024	0.046	0.606
loan size	0.066	0.031	0.031	0.271	0.010	0.000
FICO	-0.605	0.036	0.000	-0.009	0.015	0.521
mortgage premium	0.112	0.038	0.003	0.140	0.017	0.000
equity position	-0.610	0.051	0.000	0.148	0.009	0.000

Table 9.3: Estimated coefficients and their standard errors and p-values using the multilogit model. The references for the first three categorical covariates are purchase, full documentation of income and owner occupancy, respectively. The last four continuous covariates are standardized.

package *nnet*.

9.2 Covariate Effects

Table 9.3 reports the estimated coefficients, their standard errors and p-values in the multilogit regression model. Since loan age is included in the multilogit model in terms of a natural cubic spline, its effects on default and prepayment probabilities are displayed in Figure 9.4. As a comparison, the effects of the covariates in the neural network model are displayed in Figures 9.5–9.11.

In both models, the equity position has a positive effect on prepayment probability and a negative effect on default probability, which is consistent with the option theory of value-minimizing mortgage models. The mortgage premium value has positive effects on both prepayment and default probabilities. Its positive effect on default probability may be a little surprising, but it is consistent with the findings in Deng et al. (2000). They pointed out that a borrower may default as a means of prepayment when the interest rate is low enough. Furthermore, Figure 9.7 shows that the effects of the mortgage premium value on prepayment and default probabilities are nonlinear

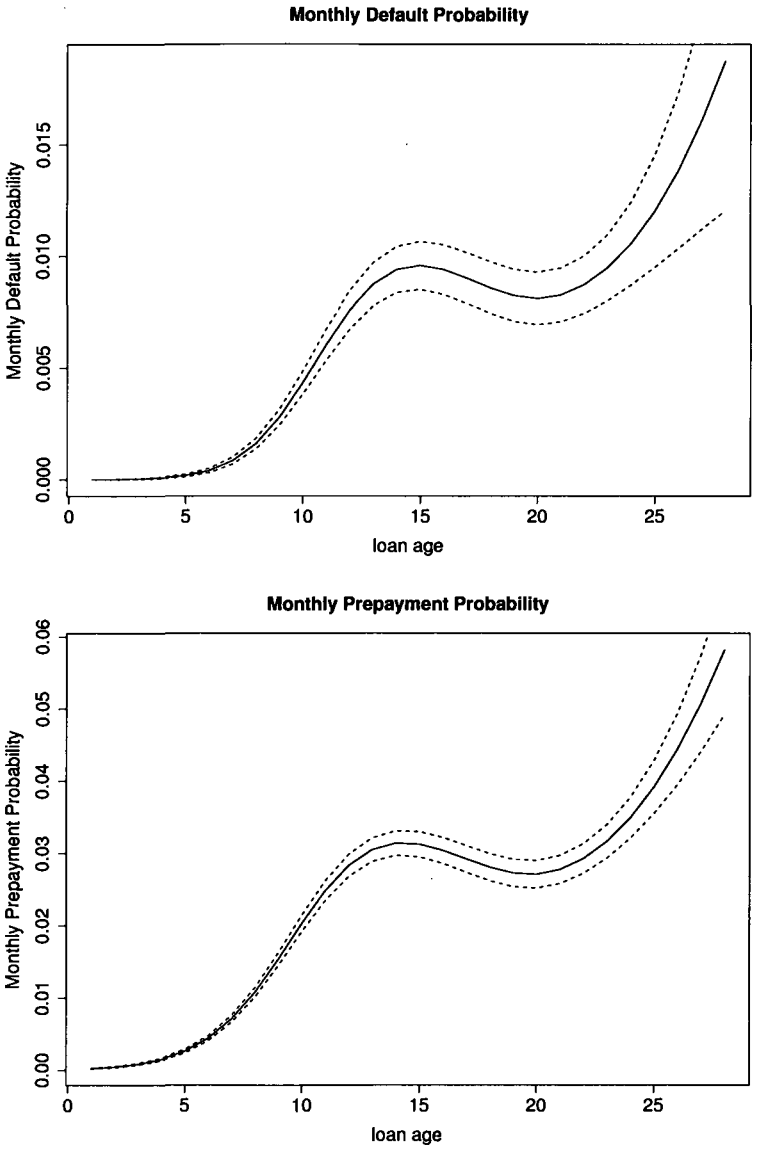


Figure 9.4: Effects of loan age on default (top) and prepayment (bottom) probabilities in the multilogit model. The other continuous covariates are at their sample means, the categorical covariates are purchase, full documentation of income and owner occupancy, respectively. The dashed lines are the 95% pointwise confidence bands.

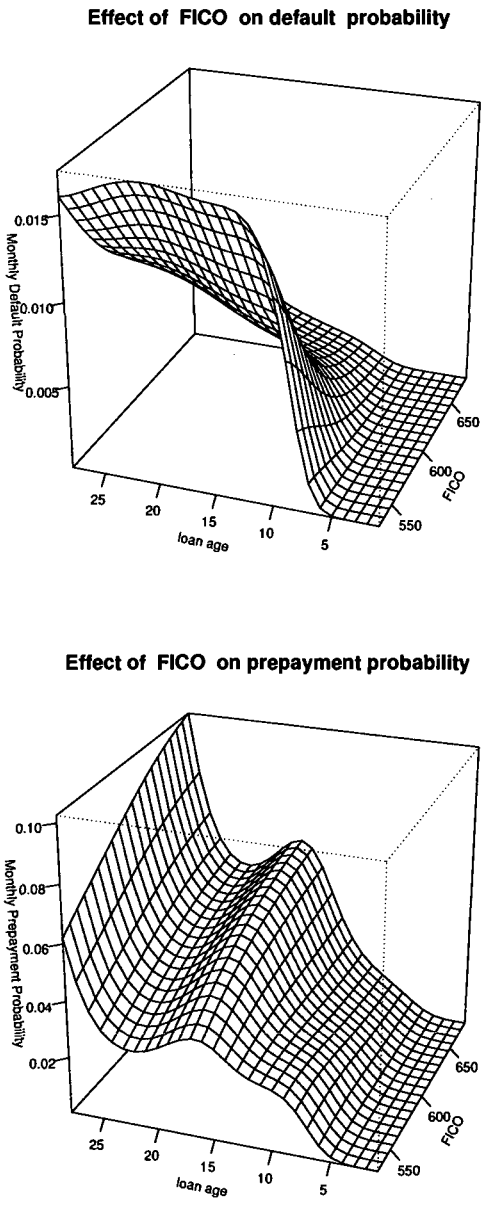


Figure 9.5: Effects of FICO on default (top) and prepayment (bottom) probabilities. The other continuous covariates are at their sample means, the categorical covariates are purchase, full documentation of income and owner occupancy, respectively.

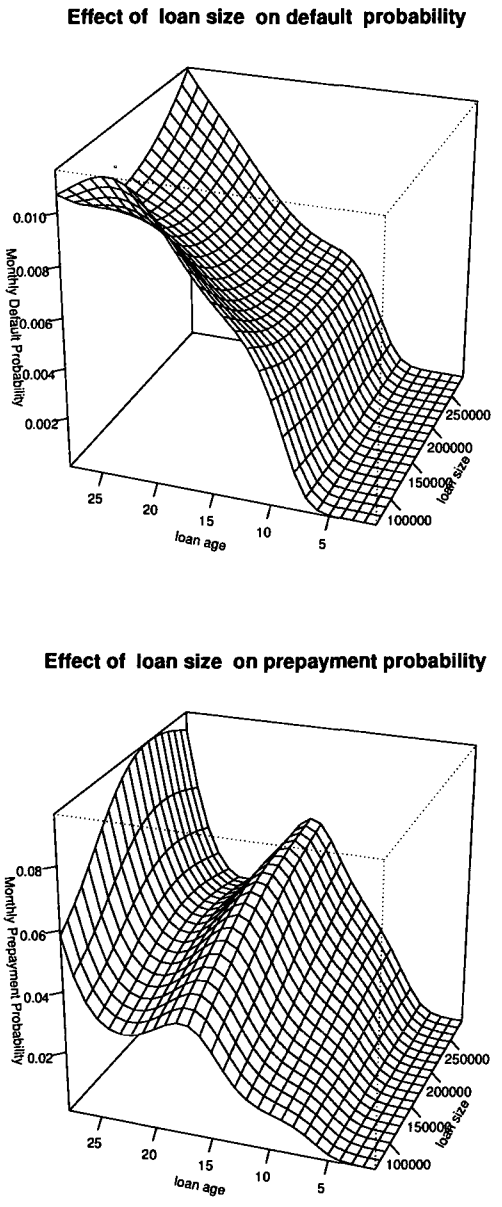


Figure 9.6: Effects of the loan size on default (top) and prepayment (bottom) probabilities. The other continuous covariates are at their sample means, the categorical covariates are purchase, full documentation of income and owner occupancy, respectively.

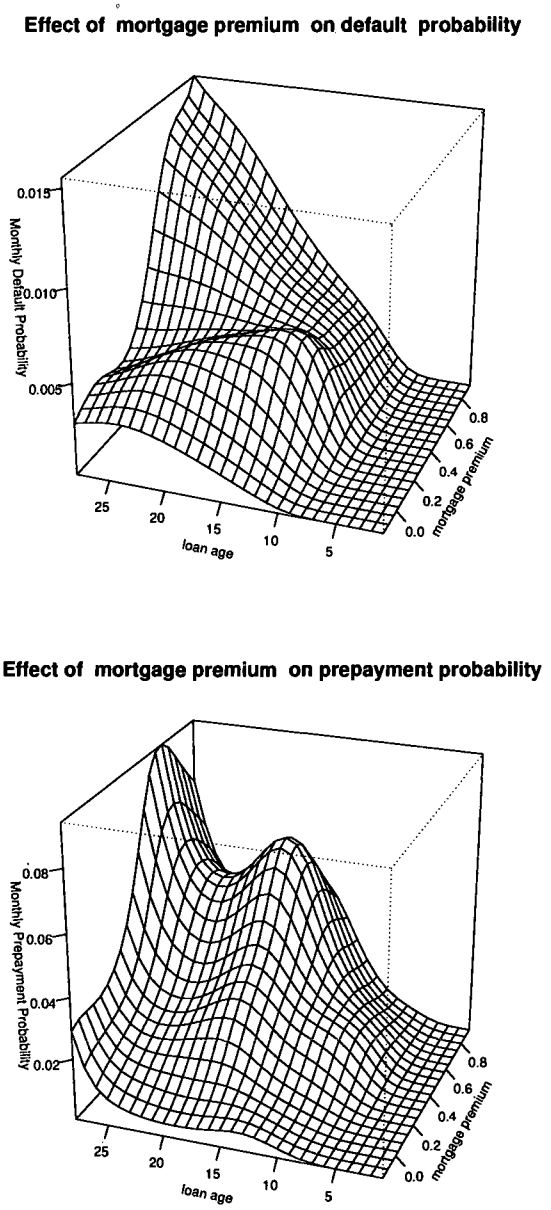


Figure 9.7: Effects of the mortgage premium value on default (top) and prepayment (bottom) probabilities. The other continuous covariates are at their sample means, the categorical covariates are purchase, full documentation of income and owner occupancy, respectively.

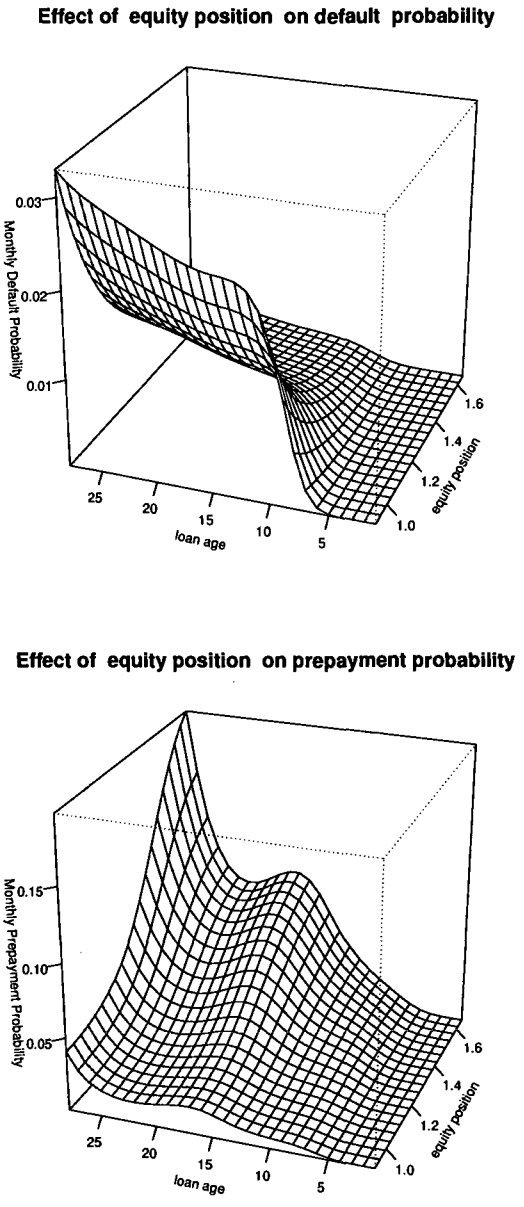


Figure 9.8: Effects of the equity position on default (top) and prepayment (bottom) probabilities. The other continuous covariates are at their sample means, the categorical covariates are purchase, full documentation of income and owner occupancy, respectively.

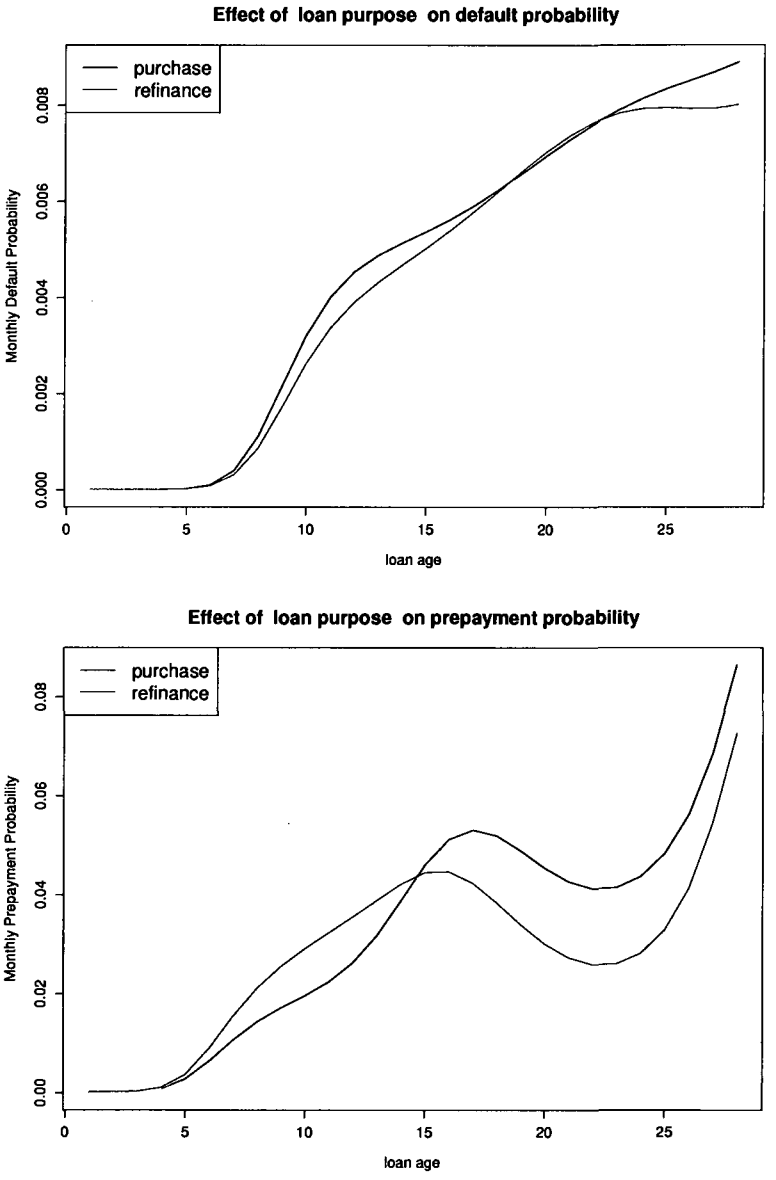


Figure 9.9: Effects of loan purpose on default (top) and prepayment (bottom) probabilities. The continuous covariates are at their sample means, the other categorical covariates are full documentation of income and owner occupancy.

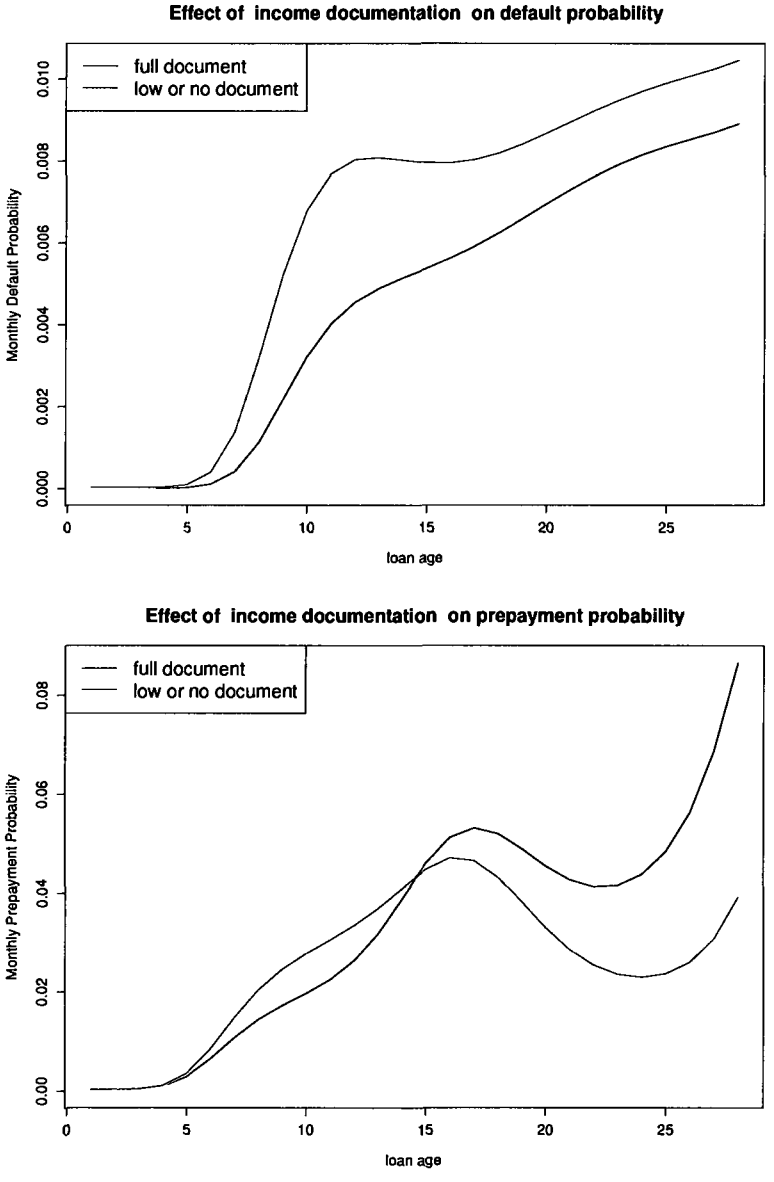


Figure 9.10: Effects of income documentation on default (top) and prepayment (bottom) probabilities. The continuous covariates are at their sample means, the other categorical covariates are purchase and owner occupancy.

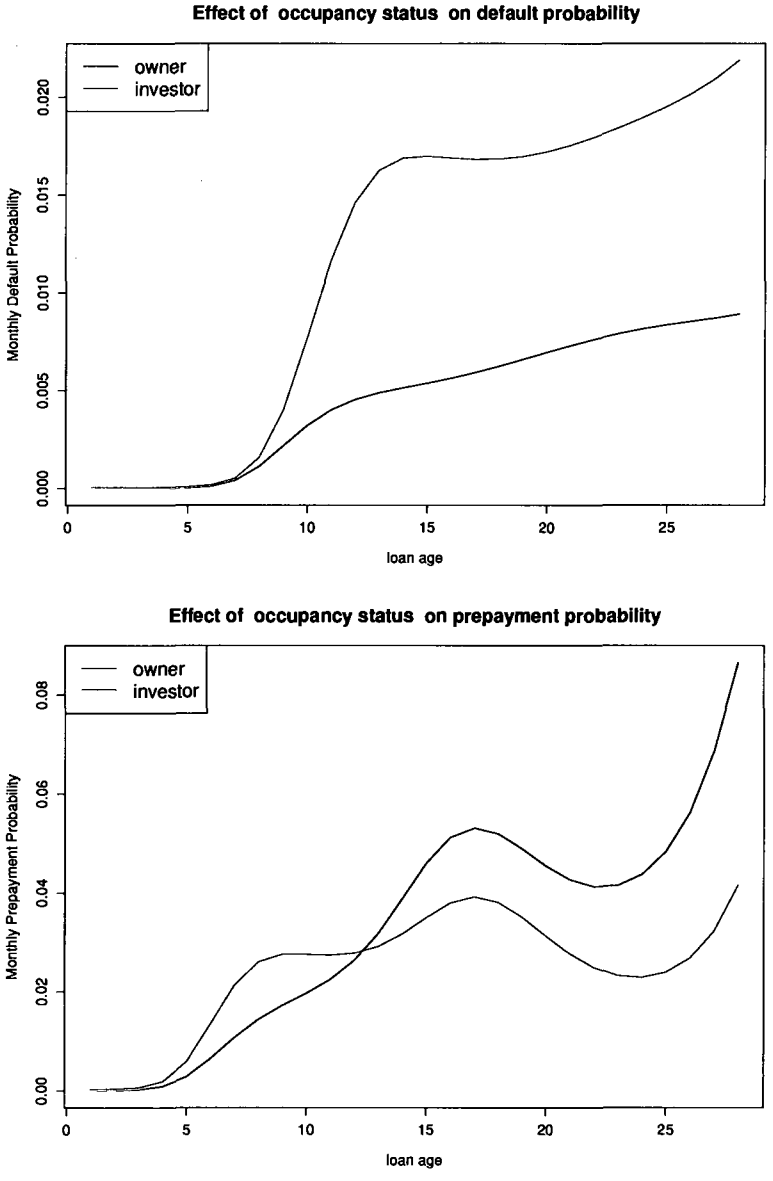


Figure 9.11: Effects of occupancy status on default (top) and prepayment (bottom) probabilities. The continuous covariates are at their sample means, the other categorical covariates are purchase and full documentation of income.

and interact with loan age.

Although the effects of the covariates in the multilogit model and the neural network model are in general consistent, there are a few differences. For example, FICO has a negative effect on default probability in both models; its effect on prepayment probability is insignificant in the multilogit model, but positive in the neural network model. In the multilogit model, loans for refinance have a lower probability of default and a higher probability of prepayment than loans for purchase. In the neural network model, loan purpose appears to have no significant effect on default probability; its effect on prepayment probability changes with loan age. Compared with purchase, refinance increases the conditional prepayment probability for the first 14 months after origination and then decreases the probability. The effects of income documentation and occupancy status on prepayment probability in the neural network model also change with loan age in a very similar way.

9.3 Predicting Prepayment and Default Rates

In this section we use the multilogit model and the neural network model to predict the fraction of the outstanding loans in a pool prepaid or defaulted on in a given month (hereafter called prepayment or default rate). Let $S_i(k)$ denote the probability that the ith loan is still current at the end of month k ; $D_i(k)$ denote the probability that the ith loan will default by the end of month k conditioned on that it is current at the end of month $k-1$; $P_i(k)$ denote the probability that the ith loan will prepay by the end of month k conditioned on that it is current at the end of month $k-1$.

$D_i(k)$ and $P_i(k)$ can be estimated directly from the models, given the loan-specific covariates, 1-year CMT rate and HPI at month k. $S_i(k)$ can be computed by induction

assuming that $S_i(1) = 1$ for all i:

$$S_i(k) = S_i(k-1)(1 - D_i(k) - P_i(k)). \tag{9.1}$$

For a pool of n loans, the estimated default rate at month k is

$$\widehat{D}(k) = \frac{\sum_{i=1}^{n} S_i(k-1)D_i(k)}{\sum_{i=1}^{n} S_i(k-1)}. \tag{9.2}$$

Similarly, the estimated prepayment rate at month k is

$$\widehat{P}(k) = \frac{\sum_{i=1}^{n} S_i(k-1)P_i(k)}{\sum_{i=1}^{n} S_i(k-1)}. \tag{9.3}$$

Note that there are two time scales involved; one is the calender time, the other is the loan age. These two time scales are related by the origination time of the loan, with the loan age being the difference between the calender time and the origination time. We first look at the monthly prepayment and default rates when all the loans are aligned by loan age. The test dataset for this purpose consists of all the loans originated in 2004, excluding those used in the training data. Figure 9.12 displays the actual and the predicted prepayment and default rates of this test dataset. After 24 months of age, the prepayment rates increase dramatically; the default rates also increase, but less dramatically. This is likely due to the reset of mortgage rates and release from prepayment penalty after two years. Table 9.4, which tabulates the L_1 errors of the two models, shows that the neural network model has a smaller L_1 error compared to the multilogit model.

To predict prepayment and default rates in a given calender month, we design three test datasets, consisting of loans originated in the first half of 2004, 2005, and 2006, respectively. Each pool has about $16,000$ loans, none of which are used in the training data. For each pool, we predict the monthly prepayment and default

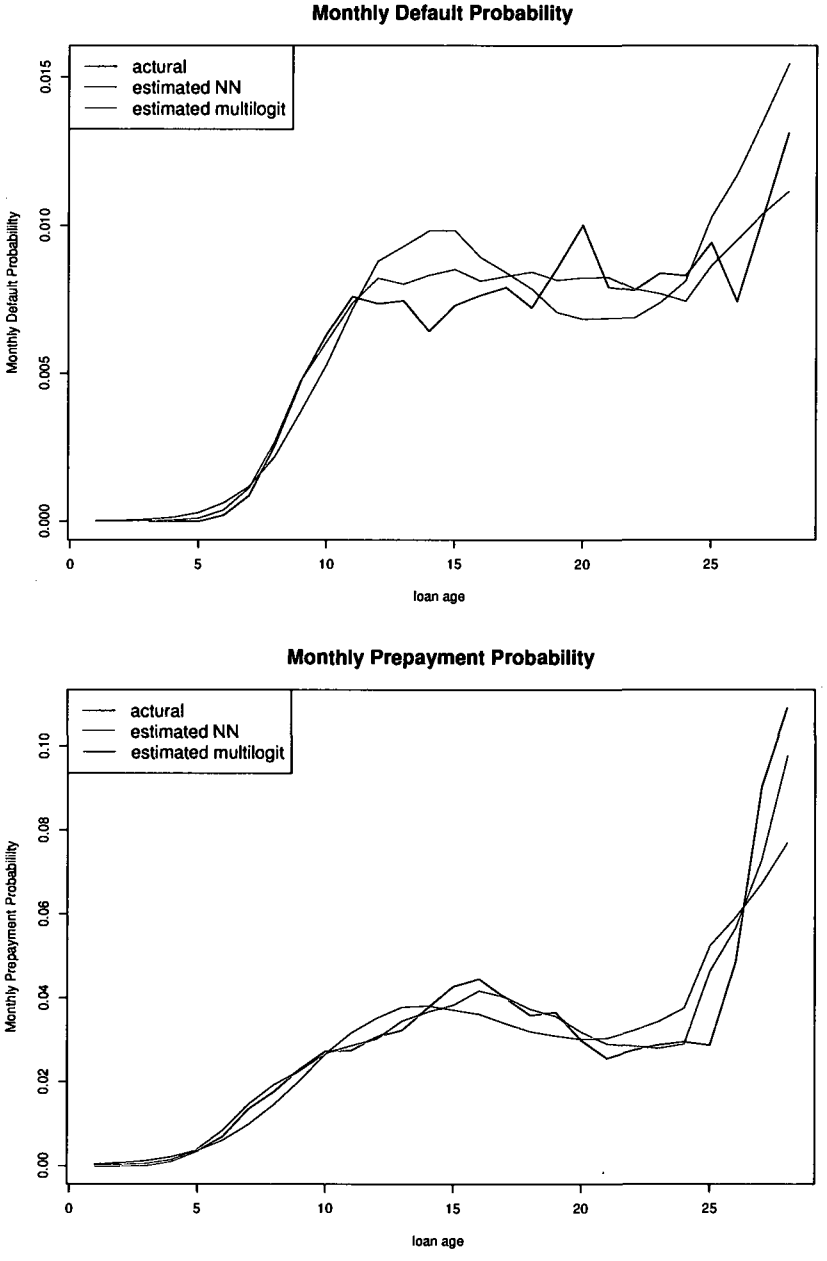


Figure 9.12: Actual vs predicted monthly default (top) and prepayment (bottom) rates of loans aligned by loan age: actual rates (black), predicted rates by the neural network model (red), predicted rates by the multilogit model (green).

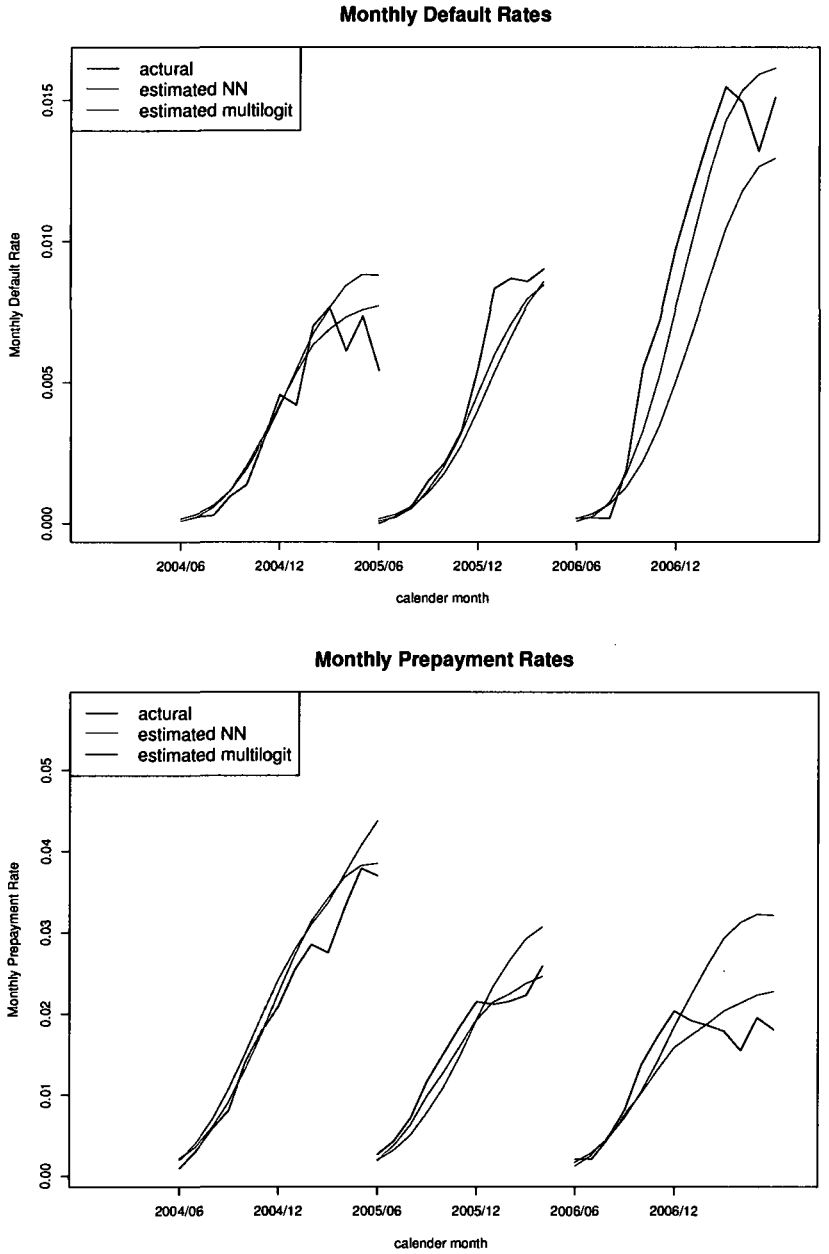


Figure 9.13: Actual vs predicted monthly default (top) and prepayment (bottom) rates on the three test datasets of vintages 2004, 2005 and 2006, respectively: the actual rates (black), the predicted rates by the neural network model (red) and the predicted rates by the multilogit model (green).

Model	Prepayment	Default
multilogit	0.00612	0.00122
neural network	0.00295	0.00060

Table 9.4: L_1 errors of the predicted monthly prepayment and default rates using the multilogit model and the neural network model on the same test dataset used in Figure 9.12.

Vintage	Prepayment		Default	
	neural network	multilogit	neural network	multilogit
2004	0.00271	0.00172	0.00061	0.00080
2005	0.00142	0.00336	0.00063	0.00086
2006	0.00246	0.00577	0.00119	0.00263

Table 9.5: L_1 errors of the predicted monthly prepayment and default rates on the three test datasets using the multilogit model and the neural network model.

rates for 12 months, the results of which are shown in Figure 9.13. The L_1 errors of the predictions are tabulated in Table 9.5. The neural network model has smaller L_1 errors than the multilogit model in all the comparisons except for predicting the prepayment rates of the pool originated in 2004.

It is worth noting that loans in the third pool, which originated in 2006, have different underwriting standards from those in the training data, which originated from 2004 to 2005 and were observed until April 2006. During the period of observation for the training data, the house prices were increasing; however, the prediction of the prepayment and default rates of the third pool is from June 2006 to May 2007, a period during which house prices began to decline. It is not surprising that the third pool has higher default rates than the pools originated in 2004 and 2005 and lower prepayment rates than the pool originated in 2004. Of special interest is that the neural network model can predict the prepayment and default rates of the third pool quite accurately under a very different macro-economic situation from that of the training data.

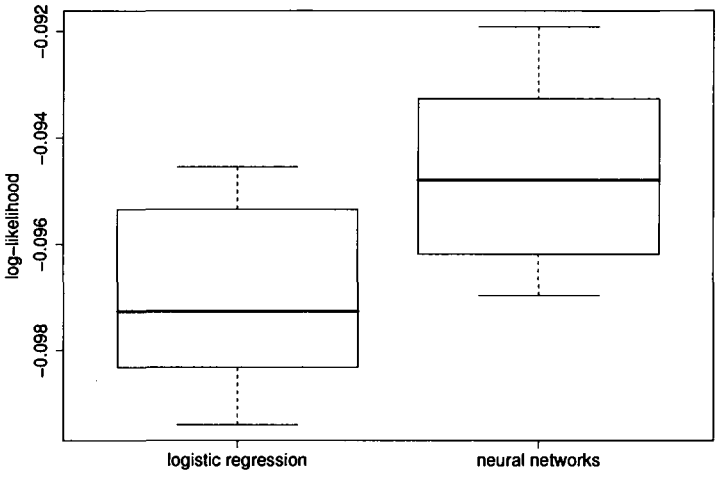


Figure 9.14: Boxplots of log-likelihoods on the 14 test datasets using the multilogit model (left) and the neural network model (right).

To further compare the out-of-sample fitting of the neural network model and the multilogit model, we split all the loans except those in the training data into 14 test datasets and compare the log-likelihoods of the two models on these test datasets. As shown by the boxplots in Figure 9.14, the neural network model has consistently higher log-likelihoods than the multilogit model, providing evidence that the neural network model fits the test datasets better than the multilogit model.

Chapter 10

Conclusion

We proposed a neural network model as an alternative to the multilogit regression model to estimate the conditional monthly prepayment and default probabilities. Both the multilogit model and the neural network model were applied to a large 2-28 adjustable rate subprime mortgage dataset originated from 2004 to 2006. The neural network model shows that the effects of the mortgage premium value on prepayment and default probabilities are nonlinear and interact with loan age. The effects of loan purpose, income documentation and occupancy status also change with loan age.

We used the neural network model and the multilogit model to predict the monthly prepayment and default rates on four pools of loans. The neural network model has smaller L_1 errors than the multilogit model. The conditional prepayment probability increases dramatically after 24 months since origination due to reset of mortgage rates and release from prepayment penalty. The pool originated in 2006 has higher default rates and lower prepayment rates compared to the pool originated in 2004. The neural network model can predict quite accurately the prepayment and default rates of the third pool under a situation of house price depreciation, although the model is trained over a period of house price appreciation.

Both the neural network model and the multilogit model estimate the hazards of

prepayment and default. We chose to estimate the hazards instead of the cumulative incidence functions, because the duration of the data is relatively short and in practice one is often interested in projecting the prepayment and default rates one or two months ahead. If the question to be addressed is how many loans will default or be prepaid within a given number of years, the methods in Part I of this thesis, which directly model the cumulative incidence functions, can also be applied.

The effects of the state of the economy on default and prepayment probabilities are characterized by two variables, the risk-free interest rate and the House Price Index (HPI). Although the risk-free interest rate and the HPI have major impacts on default and prepayment behavior, it is reasonable to assume that there are other influences of the state of the economy beyond these two variables. Future work will include modeling with time series methods the influence of the state of the economy unexplained by the interest rate and the HPI.

Bibliography

AALEN, O. O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *the Annals of Statistics*, **6** 534–545.

AALEN, O. O. and JOHANSEN, S. (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, **5** 141–150.

ANDERSEN, P. K., BORGAN, O., GILL, R. and KEIDING, N. (1992). *Statistical models based on counting processes.* Springer-Verlag, New York.

ANDERSEN, P. K., KLEIN, J. P. and ROSTHØ J, S. (2003). Generalized linear models for correlated pseudo-observations with applications to multi-state models. *Biometrika*, **90** 15–27.

CALHOUN, C. A. and DENG, Y. (2002). A dynamic analysis of fixed and adjustable rate mortgage terminations. *the Journal of Real Estate Finance and Economics*, **24** 9–33.

CODDINGTON, E. A. and LEVINSON, N. (1955). *Theory of Ordinary Differential Equations.* McGraw-Hill.

COLEMAN, T. and LI, Y. (1994). On the convergence of reflective newton methods

for large-scale nonlinear minimization subject to bounds. *Mathematical Programming*, **67** 189–224.

COX, D. (1959). The analysis of exponentially distributed lifetimes with two types of failure. *Journal of the Royal Statistical Society, Series B*, **21** 411–421.

COX, D. R. and SNELL, E. J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B*, **30** 248–275.

DENG, Y., QUIGLEY, J. M. and VAN ORDER, R. (1996). Mortgage default and low downpayment loans: the costs of public subsidy. *the Journal of Regional Science and Urban Economics*, **26** 263–285.

DENG, Y., QUIGLEY, J. M. and VAN ORDER, R. (2000). Mortgage termination, heterogeneity, and the exercise of mortgage options. *Econometrica*, **68** 275–307.

DUNN, K. B. and MCCONNELL, J. J. (1981a). A comparison of alternative models for pricing bnma mortgage-backed securities. *the Journal of Finance*, **36** 375–392.

DUNN, K. B. and MCCONNELL, J. J. (1981b). Valuation of gnma mortgage-backed securites. *the Journal of Finance*, **36** 599–617.

FINE, J. P. (2001). Regression modeling of competing crude failure probabilities. *Biostatistics*, **2** 85–97.

FINE, J. P. and GRAY, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94** 496–509.

GOOLEY, T., LEISENRING, W., CROWLEY, J. and STORER, B. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine*, **18** 695–706.

GRAW, F., GERDS, T. A. and SCHUMACHER, M. (2008). On pseudo-values for regression analysis in competing risks models. Tech. rep., Institute of Public Health, University of Copenhagen.

GRAY, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *the Annals of Statistics*, **16** 1141–1154.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

KAY, R. (1977). Proportional hazards regression models and the analysis of censored survival data. *Applied Statistics*, **26** 227–237.

KLEIN, J. P. and ANDERSEN, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics*, **61** 223–229.

KLEIN, J. P., GERSTER, M., ANDERSEN, P. K., TARIMA, S. and PERME, M. P. (2008). Sas and r functions to compute pseudo-values for censored data regression. *Computer Methods and Programs in Biomedicine*, **89** 289–300.

KUPIEC, P. and KAH, A. (1999). On the origin and interpretation of oas. *the Journal of Fixed Income* 82–92.

LAI, T. L. and SMALL, D. (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69** 79–99.

LAI, T. L. and YING, Z. (1988). Stochastic integrals of empirical-type processes with applications to censored regression. *Journal of Multivariate Analysis*, **27** 334 – 358.

LAI, T. L. and YING, Z. (1991). Large sample theory of a modified buckley-james estimator for regression analysis with censored data. *the Annals of Statistics*, **19** 1370–1402.

LAI, T. L. and YING, Z. (1994). A missing information principle and m-estimators in regression analysis with censored and truncated data. *The Annals of Statistics*, **22** 1222–1255.

LARSON, M. G. (1984). Covariate analysis of competing risks models with log-linear models. *Biometrics*, **40** 459–469.

LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **78** 13–22.

PEPE, M. S. (1991). Inference for events with dependent risks in multiple endpoints studies. *Journal of the American Statistical Association*, **86** 770–778.

PRENTICE, R. L., KALBFLEISCH, J. D., PETERSON, A. V., FLOURNOY, N., FAREWELL, V. T. and BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, **34** 541–554.

RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

ROBINS, J. M. and ROTNITZKY, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology-Methodological Issues* 24–33.

SCHEIKE, T. and ZHANG, M. J. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, **95** 205–220.

SCHWARTZ, E. S. and TOROUS, W. N. (1989). Prepayment adn the valuation of mortgage-backed securities. *the Journal of Finance*, **44** 375–392.

SCHWARTZ, E. S. and TOROUS, W. N. (1992). Prepayment, default and the valuation of mortgage pass-through securities. *the Journal of Business*, **65** 221–239.

SZYDLO, R., GOLDMAN, J. M., KLEIN, J. P., GALE, R., ASH, R., BACH, F., BRADLEY, B., CASPER, J., FLOMENBERG, N., GAJEWSKI, J., GLUCKMAN, E., HENSLEE-DOWNEY, P., HOWS, J., JACOBSEN, N., KOLB, H., LOWENBERG, B., MASAOKA, T., ROWLINGS, P., SONDEL, P., VAN BEKKUM, D., VAN ROOD, J., VOWELS, M., ZHANG, M. and HOROWITZ, M. (1997). Results of allogeneic bone marrow transplants for leukemia using donors other than HLA-identical siblings. *Journal of Clinical Oncology*, **15** 1767–1777.

TSIATIS, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, **72** 20–22.

VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer.