Likelihood-Based Statistical Inference of Proportional Hazards, Competing Risks Models with Mortgage Duration Data

MARK YUYING AN  
Vice President for Research  
Fannie Mae, Washington, DC, 20016  
E-mail: mark_y_an@fanniemae.com

ZHIKUN QI  
Senior Financial Economist  
Fannie Mae, Washington, DC, 20016  
E-mail: zhikun_qi@fanniemae.com

Summary This paper demonstrates two important results related to the estimation of a competing risks model under the proportional hazards assumption with grouped duration data, a model which has become the canonical model for the termination of mortgages with prepayment and default as two competing risks. First we show that the model with non-parametric baseline hazards is unidentifiable with only grouped mortgage duration data. Therefore assumption on the functional form of the baseline hazard is necessary for any meaningful inference. Secondly we demonstrate that under some parametric assumption such as piece-wise constant baseline hazards, the sample likelihood function has an explicit analytical form. Therefore there is no need for the approximation formula widely adopted in the previous literature. Both Monte Carlo simulations and actual mortgage data are used to demonstrate the adverse impact of the approximation.

Keywords: Competing Risks, Proportional Hazards, Grouped Duration Data, Mortgage prepayment and Default
1. INTRODUCTION

The competing risks model of duration data was first introduced in biometrics literature where a typical example is that an animal is exposed to multiple potential causes of death. Each cause is a competing risk in the sense that the smallest realized risk-specific duration makes the durations for other risks right-hand censored. For example, the death due to heart disease makes the potential time of death due to kidney failure unobserved.

Following the seminal work of Han and Hausman (1990), Sueyoshi (1992), and McCall (1996), competing risks modeling has become very popular in economic analysis of duration data when the duration of economic event has multiple causes of termination. For example, an unemployment spell may end either when the unemployed person accepts a job offer or when she drops out of labor force for good. A mortgage loan contract can terminate either when the borrower defaults on the loan by surrendering the collateral and walking away or when she prepays the loan before the end of the original term. In all these examples, it is customary to model each of the underlying duration associated with each of these risks.

Analysis of duration data typically starts with the specification of hazard function of the duration variable (as a function of potential observed and unobserved covariates) rather than its density function or conditional mean. Hazard function and density function of a random variable have a one-to-one relationship. One can always derive one from the
other. However, in many economic applications, hazard function is more natural than
density function, because (1) modeling of economic decision making results in direct
empirical implication on the form of the hazard function of event durations (Christensen
and Kiefer (2009)), and (2) hazard function is better suited to deal with time-varying
covariates.

The most convenient and popular functional form is the so-called proportional hazard. An (1995) studies economic decision on discrete choices and provides necessary and
sufficient conditions for proportional hazard specification. Recently the competing risks
model under proportional hazards assumption has been applied to analyzing loan
performance where a mortgage can terminate by either prepayment or default (Deng et al.,
decides when and in which way to stop the monthly payment of principal and interest.
She can either choose prepayment, which, viewed as a call option, refers to total payoff
of the outstanding loan, or choose default, which, viewed as a put option, refers to
surrendering the collateral to the lender and walking away.

A borrower would prepay for any (or a combination) of the following three reasons:

1) **Rate Refinance**. The market mortgage interest rate has dropped substantially so

---

1The net gain from Rate Refinance is the difference between the lifetime saving due to
the lower interest and the one-time refinance transaction costs. Thanks to the
advancement of underwriting technology and standardization of mortgage underwriting,
the transaction cost of mortgage origination has substantially reduced in recent years. As
a consequence, it used to be a rule of thumb that a 2-percentage drop in interest rate would
warrant a rate refinance. Now the threshold is only in the neighborhood of half a
percentage.
that there is a net financial gain for the borrower to apply for a new mortgage at the currently mortgage rate and pay back the existing mortgage.

2) **Cash-out Refinance.** When the homeowner has accumulated a substantial amount of home equity, which is the difference between the value of the house and the unpaid principal balance on the mortgage, she may want to borrow a larger amount of money relative to the current unpaid balance. By borrowing more from the new lender, the borrower can use the extra money for other purpose.

3) **Housing Turnover.** Due to the change in the economic and demographic conditions in the household (divorce or death in the family, more income, new job offer etc.), the homeowner decides to move out of the house (in order to buy a bigger or smaller house, to become a renter, to move to another area, etc.). Since each mortgage is tied to a particular collateral property, selling the house implies the total unpaid balance has to be repaid in total to the lender.

In contrast, mortgage default is more difficult to understand than mortgage prepayment. According to the ruthless default hypothesis, a borrower would default if the expected net benefit of default is positive. If the borrower thinks that she has a negative home equity due to a depreciation of the market value of the house, i.e. the market value of the house is less than the unpaid balance on the mortgage; she would just walk away and surrender the property to the lender. A borrower who walks away from an underwater mortgage even though she has the ability to pay is now called a **strategic defaulter**. In reality things are more complicated than that. This is because the decision to default depends on several uncertain factors.
1) Not every borrower knows with certainty what the house is actually worth.

2) The cost of defaulting includes non-monetary loss such as psychological discomfort, difficulty in borrowing in the future, etc. These non-monetary costs vary across people and are hard to observe and quantify.

3) Not every borrower with negative home equity would immediately default. One might as well wait hoping for potential future appreciation of house value.

Aside from the common belief that defaults are mainly caused by negative home equity, many believe defaults have to be triggered by a stressful event, such as a layoff from work, a divorce, etc. These events are hard to be observed by modeler.

Competing risks modeling often deals with grouped duration data. In grouped duration data, the realized durations are only known to fall into known intervals. For instance, unemployment spells are typically measured in weeks. Mortgage payments are usually observed in monthly intervals. And in the study of clinical trials, a subject's blood sample is only tested on certain scheduled time intervals, making the exact time of infection known only up to the interval between two consecutive tests.

With only grouped duration data, the modeler has two choices: (a) to treat the underlying durations as discrete valued variables, or (b) to treat the durations as continuous-time variables but to fit the model with grouped duration data. Both alternatives have limitations. This paper discusses the latter approach - fitting continuous-time competing
risks model with grouped duration data. It is a natural extension of single-risk duration analysis with grouped duration data (Prentice and Gloakler, 1978, Kiefer, 1988, Ryu 1994, An 2000). We are concerned with some methodological issues related to likelihood-based estimation of competing risks models under proportional hazards assumption with grouped duration data. Specifically, we establish the following two results:

1) First, we show that the model with non-parametric baseline hazards is unidentifiable with grouped duration data. This implies that any consistent estimation and meaningful inference have to hinge on and/or stem from assumption of the shape of the baseline hazards. This is quite contrary to the single-risk duration case, where consistent estimation model regression coefficients do not rely on parametric assumptions of the baseline hazard even with grouped duration data.

2) Second, we point out that under some parametric assumption, such as piece-wise constant baseline hazards, the sample likelihood function has an explicit analytical form. An immediate implication is that there is no need for approximation once one makes the assumption that the baseline hazards are piece-wise constant.

Section 2 of the paper introduces the basic notation and describes the framework of statistical inference with grouped duration data. Section 3 discusses the main non-identification result. In Section 4, we comment on the piece-wise constant case and derive the implied analytical likelihood function. Section 5 reports results from our Monte Carlo simulations as well as a real residential mortgage data set. These results illustrate different degrees of adverse impact of using the approximation formula on
parameter estimates. Concluding remarks are made in Section 6 with some extensions to the basic setting.

2. COMPETING RISKS MODEL UNDER PROPORTIONAL HAZARDS SPECIFICATION

To focus on the presentation of the main ideas, we restrict our attentions to situations when

1) all explanatory variables are time-invariant;

2) there are only two competing risks;

3) the duration variables are grouped into regular intervals bounded by positive integers;

4) the heterogeneity distribution either is degenerate, or has a known bivariate parametric distribution.

The leading example in this paper is mortgage loan termination, where two competing risks are prepayment and default. Let us introduce the model with that application in mind. For each Loan n in the sample, let $T_{1n}$ be the latent duration until prepayment and $T_{2n}$ be the latent duration until default. Instead of directly observe $T_{1n}$ and $T_{2n}$, the data only reveal $Y_n = \min\{T_{1n}, T_{2n}\}$ along with the information about the cause of the loan termination. If it is known that the loan is terminated due to prepayment, assign $R_n = 1$.

---

2Extensions to more general settings are briefly discussed in Section 6.
If it is known that the loan is terminated due to default, assign $R_n = 2$. However, if by the time the survey ends, the loan of age $c$ is still actively performing, then assign $R_n = 0$ and in this case, we say the both latent durations are right-hand censored at $c$.

Along with the above information on observed duration and the cause of termination for each loan, the econometrician is also equipped with a vector of weakly exogenous covariates. The key variables $W_{1n}$ that impact the hazard for prepayment are Loan Amount, Market Interest Rate, Note Spread, Loan-to-Value Ratio. The key variables $W_{2n}$ that impact the hazard for default are Credit Score, Loan-to-Value Ratio, Debt-to-Income Ratio, Owner Occupancy Status, etc. Let $X_n$ be the union of $W_{1n}$ and $W_{2n}$. And let $V_n = (V_{1n}, V_{2n})$ be two unobserved heterogeneity factors affecting prepayment risk and default risk, respectively.

A continuous-time competing risks model under proportional hazard specification has the following three components:

**Assumption 1 (Conditional Independence)** Conditional on the observed and unobserved heterogeneity, $(X_n, V_n)$, the two risk-specific durations $T_{1n}$ and $T_{2n}$ are

---

$^3$Although it is reasonable to assume that the two risk-specific durations may have difference sets of observable covariates, our notation here to use $X_n$ as the union of the two sets is nonetheless without loss of generality. Later we will assume that the effect of $X_n$ on Risk $j$ will be in the form of linear index $X_n \beta_j$, $j = 1, 2$. The above restriction can be accommodated by assigning certain element of the $\beta_j$ vector to be zero.
independent.

**Assumption 2 (Proportional Hazards)** Conditional on \((X_n, V_n) = (x, v)\), the hazard rates for \(T_{1n}\) and \(T_{2n}\) are, respectively,

\[
h_j(t | x, v) = \lambda_j(t) \exp\{x\beta_j + v_j\}, \quad j = 1, 2, \quad t > 0
\]

(1)

Where \(\lambda_j(t)\) is the baseline hazard, \(\exp\{x\beta_j + v_j\} = \exp\{x\beta_j\}\exp\{v_j\}\) is the loan specific effect.

**Assumption 3 (Heterogeneity Distribution)** The unobserved heterogeneity vector \(V_n = (V_{1n}, V_{2n})\) is independent from \(X_n\), and is distributed with a bivariate distribution function \(G(v_1, v_2)\), that satisfies either of the following conditions:

a. \(G(v_1, v_2)\) is degenerate, i.e., \(P(V_{1n} = 0, V_{2n} = 0) = 1\), or

b. \(G(v_1, v_2; \gamma)\) has a parametric form with parameter \(\gamma\) and satisfies normalization restrictions that \(E[V_{1n}] = 0\) and \(E[V_{2n}] = 0\).

For some applications the parameters of primary interest are the regression coefficients \(\beta_1\) and \(\beta_2\) together with possibly \(\gamma\) in the heterogeneity distribution. For example, robust estimation of \(\beta_1\) and \(\beta_2\) is vital in testing option theory of mortgage behavior.

According to this theory (see for example Deng et al (2000)), the borrower has essentially been given two options: prepayment as call option and default as a put option. The borrower can dynamically calculate whether these embedded options in mortgage
contract are ‘in the money’ and when is optimal to exercise these options. If \( X_n \) contains variables that are part of the borrower's calculation, then the estimated \( \beta_1 \) and \( \beta_2 \) can be used to test whether the option theory of mortgage contract is consistent with the data. In that setting it is now customary to leave the two baseline hazard functions \( \lambda_1(t) \) and \( \lambda_2(t) \) in (1) unspecified to enhance the robustness of estimating \( \beta_1 \) and \( \beta_2 \), following the tradition in single-risk setting due to the seminal work of Cox (1972).

However, in the next section, we will show why the attempt of leaving the baseline hazard functions non-parametrically specified is not fruitful when the model is to be fit with only grouped duration data.

In the context of competing risks model of mortgage performance, we will first assume, without loss of generality, that the duration variable is grouped into time intervals bounded by integers. Specifically,

**Assumption 4 (Data Grouping)** Every observation, Loan n, in the entire sample can be classified in one of the following three types of grouping associated with an integer \( K_n \):
Using the above notation, for each individual loan \( n \) in the sample we have the following triplets \((x_n, K_n, r_n)\). Notice that the information on the heterogeneity vector \((V_{1n}, V_{2n}) = (v_{1n}, v_{2n})\) is, by definition, unobserved to the econometrician.

### 3. NON-IDENTIFICATION

To simplify exposition, let

\[
\Lambda_j(s) = \int_0^s \lambda_j(t) \, dt, \quad j = 1, 2
\]

(2)

denote the risk-specific integrated baseline hazard; and let

\[
\phi_{jn} = \exp\{x_n \beta_j + v_{jn}\}, \quad j = 1, 2
\]

(3)

denote the effect of (observed as well as unobserved) individual-specific characteristics.

Under Assumptions 1 and 2, for \( s > 0 \), \( t > 0 \) and conditional on \((X_n, V_{1n}, V_{2n}) = (x_n, v_{1n}, v_{2n})\), the joint density function of \((T_{1n}, T_{2n})\) is:

\[
f(s, t \mid x_n, v_{1n}, v_{2n}) = h_1(s \mid x_n, v_{1n}) h_2(t \mid x_n, v_{2n}) \exp\{-\Lambda_1(s)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}
\]

\[
= (\lambda_1(s) \phi_{1n} * \lambda_2(t) \phi_{2n}) \exp\{-\Lambda_1(s)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}
\]

(4)
and the joint survivor function is:

\[
S(s, t \mid x_n, v_{1n}, v_{2n}) \equiv P(T_{1n} > s, T_{2n} > t \mid x_n, v_{1n}, v_{2n}) = \exp \{-\Lambda_1(s)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}.
\] (5)

Next we will derive the likelihood function based on the model and the data set up in the previous section. It is helpful to start with a Type C observation. Such an observation contributes to the sample likelihood function in the following conditional probability

\[
P(Y_n \geq K_n - 1 \mid x_n)
= \Pr(T_{1n} \geq K_n - 1, T_{2n} \geq K_n - 1 \mid x_n)
= \int [S(K_n - 1, K_n - 1 \mid x_n, v_{1n}, v_{2n})]dG(v_{1n}, v_{2n})
\] (6)

The contribution to the sample likelihood of a Type P observation can be derived with some algebra:

\[
\Pr(K_n - 1 < Y_n \leq K_n, R_n = 1 \mid x_n)
= \Pr(K_n - 1 < T_{1n} \leq K_n, T_{1n} \leq T_{2n} \mid x_n)
= \int \left[ \Pr(K_n - 1 < T_{1n} \leq K_n, T_{1n} \leq T_{2n} \mid X_n, v_{1n}, v_{2n}) \right]dG(v_{1n}, v_{2n})
= \int \left[ \int_{K_n - 1}^{K_n} \lambda_1(t)\phi_{1n} \int_{K_n - 1}^{K_n} \lambda_2(s)\phi_{2n} \exp\{-\Lambda_1(t)\phi_{1n} - \Lambda_2(s)\phi_{2n}\}dtds \right]dG(v_{1n}, v_{2n})
= \int \left[ \int_{K_n - 1}^{K_n} \lambda_1(t)\phi_{1n} \exp\{-\Lambda_1(t)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}dt \right]dG(v_{1n}, v_{2n})
\] (7)

Similarly for a Type D observation, its contribution to the sample likelihood is

\[
\Pr(K_n - 1 < Y_n \leq K_n, R_n = 2 \mid x_n)
= \int \left[ \int_{K_n - 1}^{K_n} \lambda_2(t)\phi_{2n} \exp\{-\Lambda_1(t)\phi_{1n} - \Lambda_2(t)\phi_{2n}\}dt \right]dG(v_{1n}, v_{2n})
\] (8)

To illustrate the fundamental non-identification, let us take the simplest case by making Assumption 3(a), when the unobserved heterogeneity is degenerate. In this case,

\[
\phi_{jn} = \exp\{x_n\beta_j\} \quad \text{for } j = 1, 2,
\]

and equation (7) simplifies to
\[
\Pr(K_n - 1 < Y_n \leq K_n, R_n = 1 | x_n) = \int_{K_{n-1}}^{K_n} \lambda_1(t) \phi_1(t) \exp\{-\Lambda_1(t)\phi_1(t) - \Lambda_2(t)\phi_2(t)\} dt
\]  

(9)

The above integral depends on the values of \( \lambda_1(t) \) and \( \lambda_2(t) \) for all \( t \) between the interval \((k-1, k]\), unless \( \lambda_1(t) \) and \( \lambda_2(t) \) are collinear, that is, there exists \( \rho > 0 \) such that \( \lambda_1(t) = \rho \lambda_2(t) \). First of all, this condition is unlikely to hold water. Second, this runs contrary to the non-parametric spirit. With that, we have arrived at the following result.

**Proposition 1.** If \( \lambda_1(t) \neq \rho \lambda_2(t) \) for any \( \rho > 0 \), without the parameterization of \( \lambda_1(t) \) and \( \lambda_2(t) \) and the competing risks model under proportional hazard specification is unidentified by grouped duration data.

Notice that the non-identification for the competing risk case is purely due to data grouping. The model is unidentified even without the presence of unobserved heterogeneity. The presence of unobserved heterogeneity was not the cause of non-identification.

We end this section with two remarks related to other well-known identification results in the statistics of duration models.

**Remark 1.** Proposition 1 asserts the non-identification under conditional independence. This non-identification is qualitatively different from the non-identification concept of Cox (1959) and Tsiatis (1975). In a \( J \) competing risks setting, suppose under i.i.d.
setting where every individual has the same joint distribution, $F(t)$, for the latent risk-specific duration vector $T = (T_1, T_2, \ldots, T_J)$. The analyst observe realizations of $Y = \min\{T_1, T_2, \ldots, T_J\}$ in continuous time (not grouped!), along with the risk $R = j$ for which $Y = T_j$. Tsiatis addresses the question that whether or not the analyst, equipped with data on $(Y, R)$, can non-parametrically identify the joint distribution of $T$ without assuming independence among the latent durations. Tsiatis shows that the answer is no. In fact he proves a theorem that for any distribution $H(y, r)$ of $(Y, R)$, there is always a version of $F^*(t)$ in which the risk-specific durations are mutually independent such that $F^*(t)$ and $H(y, r)$ are compatible.

Heckman and Honore (1989) show how identification of competing risks model (under proportional hazard assumption) can be achieved by the introduction of covariates. That identification of both the baseline hazard functions and the heterogeneity distribution are based on continuously observed durations. Recently, Canals-Cerda and Gurmu (2007) propose a semi-nonparametric specification of the heterogeneity distribution. Proposition 1 of this paper states that grouping in duration data prevent nonparametric identification of the baseline hazard functions.

**Remark 2.** This non-identification result is quite a contrast to the well-known situation in the single-risk setting. With a single risk, the duration variable is governed by only one hazard function $h(t | x_n, v_n)$. Under the proportional hazard setting, that is, $h(t | x_n, v_n) = \lambda(t) \exp\{x_n \beta + v_n\}$, it is well known that with grouped duration data, the likelihood function depends on the baseline hazard only through the discrete values of
integrated hazard function, that is,

$$\Pr(K_n - 1 < Y_n \leq K_n \mid x_n, v_n) = \int_{K_n - 1}^{K_n} \left[ \lambda(t)e^{x_{n}^{T}\beta + \nu_{n}} \exp\left\{ - \Lambda(t)e^{x_{n}^{T}\beta + \nu_{n}} \right\} \right] dt \quad (10)$$

The above integration does have an analytic expression, in fact,

$$\Pr(K_n - 1 < Y_n \leq K_n \mid x_n, v_n) = \exp\left\{ - \Lambda(K_n - 1)\phi_{n}^{1} \right\} - \exp\left\{ - \Lambda(K_n)\phi_{n}^{1} \right\} \quad (11)$$

Therefore provided the number of cut-off points in the entire sample of individuals is either fixed or grows slower to infinity than the sample size $N$, consistent estimation of the regression coefficients $\beta$ is achievable, even without the specification of the baseline hazard function. The existence of unobserved heterogeneity does not qualitatively alter the identification of the model, although the costly estimation procedure might be required.\(^4\)

### 4. EXACT LIKELIHOOD FUNCTION UNDER PIECE-WISE CONSTANT BASELINE HAZARDS

The direct implication of Proposition 1 is that it is rather a necessity to make functional form assumption about the baseline hazards $\lambda_{1}(t)$ and $\lambda_{2}(t)$. Any meaningful inference has to come from and hinge on that non-testable assumption.

One of the commonly used assumptions is the piece-wise constant assumption,

\(^4\)

For an intuition about the non-parametric identification in the single risk setting and how to fully exploit that feature for statistical inference purpose with and without the presence of unobserved heterogeneity, see, for example, An (2000).
popularized after Han and Hausman (1990). In this section we comment on the piece-wise constant baseline hazard and derive the exact likelihood function associated with this assumption.

Assumption 5 (Piece-wise Constant Baseline Hazards) For \( j = 1, 2 \), the baseline hazard function \( \lambda_j(t) \) is piece-wise constant, that is, there exist constants \( \{\alpha_{jk}\} \) such that

\[
\lambda_j(t) = \sum_{k=1}^{M} \alpha_{jk} \mathbb{1}_{[a(k-1), a(k)]}, \quad j = 1, 2, \tag{12}
\]

where \( M \) is the largest value in the set \( \{K_n, n=1, \ldots, N\} \). Notice that for different \( l \) and \( k \), \( \alpha_{jl} \) does not have to be different from \( \alpha_{jk} \).

Under Assumption 5, the two integrated baseline hazard functions are piece-wise linear with interval-specific slopes \( \alpha_{jk} \). We have,

\[
\Lambda_j(t) = \begin{cases} 
0, & \text{for } t = 0, \\
\alpha_{jl}, & \text{for } 0 < t < 1, \\
\alpha_{jl} + \alpha_{jl}(t-1), & \text{for } 1 \leq t < 2, \\
\sum_{l=1}^{k-1} \alpha_{jl} + \alpha_{jk}(t-k+1) & \text{for } k-1 \leq t < k
\end{cases} \tag{13}
\]

**Proposition 2.** Under Assumptions 1-5, the integral appearing in equations (7) and (8) has an analytical expression. To be specific, conditional on \((X_n, V_{1n}, V_{2n})\) = \((x_n, v_{1n}, v_{2n})\):

\[
\Pr(K_n - 1 < Y_n \leq K_n, R_n = 1 \mid x_n, v_{1n}, v_{2n})
\]

\[
= \int_{K_{n-1}}^{K_n} \hat{\lambda}_1(t) \phi_{1n} \exp\{-\Lambda_1(t)\phi_{1n} - \Lambda_2(t)\phi_{2n}\} dt
\]

\[
= \theta_n \exp\{-\Lambda_1(K_n - 1)\phi_{1n} - \Lambda_2(K_n - 1)\phi_{2n}\} [1 - \exp\{\alpha_{1K_n}\phi_{1n} - \alpha_{2K_n}\phi_{2n}\}] \tag{14}
\]
\[
\Pr(K_n - 1 < Y_n \leq K_n, R_n = 2 \mid x_n, V_{1n}, V_{2n}) \\
= \int_{K_n-1}^{K_n} \lambda(t) \phi_{1n} \exp\{-\Lambda_1(t)\phi_{1n} - \Lambda_2(t)\phi_{2n}\} dt \\
= (1 - \theta_n) \exp\{-\Lambda_1(K_n - 1)\phi_{1n} - \Lambda_2(K_n - 1)\phi_{2n}\}\left[1 - \exp\{-\alpha_{1K_n}\phi_{1n} - \alpha_{2K_n}\phi_{2n}\}\right]
\]

(15)

where \( \theta_n \) is:\n\[
\theta_n = \frac{\alpha_{1K_n}\phi_{1n}}{\alpha_{1K_n}\phi_{1n} + \alpha_{2K_n}\phi_{2n}}
\]

(16)

Remark 3. Refer to the following graph.

The total probability mass of the of the L-shaped shaded area \( (A, B, C) \), conditional on
\((X_n, V_{1n}, V_{2n}) = (x_n, v_{1n}, v_{2n})\) is:
\[
L = S(K_n - 1, K_n - 1 \mid x_n, v_{1n}, v_{2n}) - S(K_n, K_n \mid x_n, v_{1n}, v_{2n})
\]

Under Assumption 1, for \( j=1, 2 \)
\[
A_j(K_n) = A_j(K_n - 1) + \alpha_{jK_n}
\]

\(^{5}\)The result is proved by simple algebra in Appendix A.
Therefore,

\[
L = S(K_n - 1, K_n - 1 \mid x_n, v_{1n}, v_{2n}) \left[ 1 - \exp \{ -\alpha_{1x_n} - \alpha_{2x_n} \phi_{2n} \} \right]
\]

which is exactly the entire term of equation (13) without \( \theta_n \). Equations (14) to (16) make clear that Assumption 5 calls for a division of the probability mass according to the weights \( \theta_n \) and \( 1 - \theta_n \) respectively for \( R_n = 1 \) and for \( R_n = 2 \).

**Remark 4.** It is worth pointing out that under the assumption of piece-wise constant baseline hazard, Proposition 2 states that the division of the probability mass for a Type P observation

1) can be derived analytically as a function of model parameters, and

2) varies in \( X_n \) and \( K_n \).

In an influential paper, McCall (1996) proposes an approximation of the likelihood contribution of a Type P or Type D observation by essentially fixing the same division for all \( n \). The corresponding formula under McCall (1996) is

\[
\Pr(K_n - 1 < Y_n \leq K_n, R_n = 1 \mid x_n, v_{1n}, v_{2n}) \approx \Pr(ob(A) + 0.5 \cdot \Pr(ob(B))
\]

where A is the rectangular shaped area and B is the square shaped area. Notice that by fixing the 50-50 split, the division of probability mass in B does not depends on \( X_n \) or \( K_n \) and is not a function of the parameters. As such under the piece-wise linear baseline hazards (Assumption 5), using the 50-50 split approximation formula (17) constitutes a quasi-maximum likelihood method.

In recent papers on loan performance models, Deng et al (2000), Ambrose and
LaCour-Little 2001, and Ciochetti *et al.* 2002, for example, all adopt McCall's formula explicitly with their piece-wise constant assumption of the baseline hazards.

1) In mortgage termination models, compared with prepayments, loan default is an extremely rare event. It is well known that default hazard rate is only a tiny fraction (1/50, say) of the prepayment hazard rate. In this case, 50-50 split of the probability is inaccurate.

2) According to Proposition 2, the split ratio $\theta_n$ is individual specific, therefore cannot be fixed once for all for all observations.

Notice also that under Assumption 5, the joint survivor function, $S(K_n, K_n | x, v_{1n}, v_{2n})$, depends on the baseline hazards only through the 2M discrete values of the integrated baseline hazards. Define

$$
\rho_{jk} = \log[\Lambda_j(k) - (\Lambda_j(k - 1)]
$$

as the logarithm consecutive increments of $\Lambda_j$ from k-1 to k. With this parameterization, the full parameter vector is:

$$
\delta = (\beta_1, \beta_2, \rho_{11}, \rho_{12}, \ldots, \rho_{1M}, \rho_{21}, \rho_{22}, \ldots, \rho_{2M}, \gamma)
$$

Estimation of $\delta$ can be carried out by maximizing the sample log likelihood function. The optimization routine depends on how the heterogeneity distribution is specified. The most convenient way to specify the heterogeneity distribution is the two-dimensional discrete distribution. For example, on a 3x3 grids, there are 15 parameters,
In general, $V_1$ takes $L_1$ possible values $(a_1, a_2, a_3, ..., a_{L_1})$ and $V_2$ takes $L_2$ values $(b_1, b_2, b_3, ..., b_{L_2})$. And $P(V_1 = a_m, V_2 = b_n) = p_{mn}$. These parameters satisfy three constraints:

1. the probabilities sum to 1, i.e. $\sum_m \sum_n p_{mn} = 1$ ;
2. the mean of $V_1$ is zero, i.e. $\left( \sum_m a_m \left[ \sum_n p_{mn} \right] \right) = 0$ ;
3. the mean of $V_2$ is zero, i.e. $\left( \sum_n b_n \left[ \sum_m p_{mn} \right] \right) = 0$ .

With these restrictions, there would only be $L_1 \times L_2 + L_1 + L_2 - 3$ free parameters in the $\gamma$ vector. If past experience is our guide, then there is unlikely a need to increase the grid points. Typically a 2x2 grid with 8-3 = 5 free parameters should be enough (An (2000) and An et al (2004)).

5. EMPIRICAL APPLICATION

In this section we investigate the adverse effect of quasi maximum likelihood estimation by analyzing results based Monte Carlo simulation as well as results from a simple real mortgage data set.\(^6\)

\(^6\)All computing programs used for this paper, for data generation and for parameter
5.1 RESULTS BASED ON MONTE CARLO SIMULATIONS

The Monte Carlo experiment simulates loan activity data under continuous-time, competing risks, proportional hazard model. Specifically, we generate data sets that consist of a sample of loans with quarterly activities until loan terminates due to one of the two competing risks, i.e. default and prepayment risk, or until observation window ends. The data generating process is in continuous time, but the realized data are grouped into unit intervals between integers. Therefore, the output data observable to the analyst contains integer-valued duration and whether the duration is terminated due to prepayment or default, or it is right censored spell. The detailed description of the design of the simulation method is provided in Appendix C.

In the Monte Carlo experiment, we consider two scenarios with different relative probability of termination between the two competing risks. And in each scenario, we run 200 simulations with sample size of 1000, 2000, 5000 and 10000. First, we set the ratio of probability of termination due to prepayment risk (Risk P) to that of default risk (Risk D) to 50. The results are shown in Table 1. The true maximum likelihood method always performs better than the quasi-maximum likelihood method in terms of both standard deviation and mean square error. Specially, the mean square error in estimation estimation, are coded in the popular MatLab; and are available from the authors upon request.

This ratio is roughly in line with for primary residential mortgages in normal economic conditions. Mortgage default is a rare event compared to prepayments.
of the low probability risk parameters is roughly 20-30 times higher in the
quasi-maximum likelihood estimation method than in the true maximum likelihood
method we proposed. The reason is that the 50-50 split of the probability used in the
quasi-maximum likelihood method is quite inaccurate in this case.

In the second scenario, we set the ratio of probability of termination due to prepayment
risk (Risk P) to that of default risk (Risk D) to 1. The results are shown in Table 2.
Although the difference is smaller than in the first experiment, the true maximum
likelihood method still performs better than the quasi-maximum likelihood method in
terms of both standard deviation and mean square error. In this case, the 50-50 split
approximation of the probability mass in Area B as part of the quasi-maximum likelihood
method happens is less harmful, especially along the 45-degree line. But 50-50 split
formula is still independence the parameters, as it should be based on the exact likelihood
function given by Proposition 2, still negatively affects its performance.
Table 1: Monte Carlo Simulation Result

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Risk Type</th>
<th>True Parameter Mean</th>
<th>STD</th>
<th>MSE</th>
<th>Mean</th>
<th>STD</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Exact Likelihood Estimation</td>
<td>Quasi-Likelihood Estimation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>P</td>
<td>-0.5 -0.4999 0.0323 0.001</td>
<td>-0.5021 0.0363 0.0013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 0.9998 0.0312 0.001</td>
<td>1.0037 0.0362 0.0013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 -0.0009 0.0291 0.0008</td>
<td>-0.0019 0.0316 0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 0.9944 0.029 0.0009</td>
<td>0.9971 0.0317 0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-0.5 1.0585 0.345 0.1218</td>
<td>1.8507 2.2846 5.9169</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 1.9075 0.3364 0.1221</td>
<td>1.7685 2.3239 5.9639</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 0.0139 0.2287 0.0522</td>
<td>0.1567 1.3743 1.9038</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>P</td>
<td>-0.5 -0.499 0.022 0.0005</td>
<td>-0.5009 0.0238 0.0006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 0.9929 0.021 0.0002</td>
<td>1.0016 0.0254 0.0006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 0.0023 0.0201 0.0004</td>
<td>0.0019 0.0219 0.0005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 0.9998 0.0214 0.0005</td>
<td>1.0015 0.0242 0.0006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-0.5 1.067 0.2726 0.0784</td>
<td>1.3785 1.2172 1.6175</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 1.0463 0.2092 0.0457</td>
<td>1.2533 1.0414 1.1433</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 0.0021 0.1431 0.0204</td>
<td>0.0405 0.4031 0.1633</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>P</td>
<td>-0.5 -0.499 0.0136 0.0002</td>
<td>-0.4999 0.0147 0.0002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 1.0006 0.0132 0.0002</td>
<td>1.0017 0.0157 0.0002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 0.0007 0.0137 0.0002</td>
<td>0.0003 0.0143 0.0002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 0.9985 0.0144 0.0002</td>
<td>0.9995 0.0159 0.0003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-0.5 1.006 0.1446 0.0208</td>
<td>1.0568 0.3643 0.1353</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 0.0072 0.1327 0.0176</td>
<td>1.066 0.4331 0.191</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 -0.0009 0.0928 0.0086</td>
<td>0.0104 0.2505 0.0625</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>P</td>
<td>-0.5 -0.4988 0.0094 0.0001</td>
<td>-0.4997 0.0103 0.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 0.9988 0.0094 0.0001</td>
<td>0.9996 0.0104 0.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 -0.0006 0.0096 0.0001</td>
<td>-0.0009 0.0101 0.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 0.9987 0.0103 0.0001</td>
<td>0.9995 0.0114 0.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>-0.5 1.086 0.1022 0.0106</td>
<td>1.0202 0.2679 0.0718</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 0.9959 0.0931 0.0087</td>
<td>1.0085 0.203 0.0411</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 0.0012 0.0585 0.0034</td>
<td>-0.0022 0.1373 0.0188</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## 5.2 RESULTS BASED ON LOAN PERFORMANCE DATA

In this section, we use a small sample of real mortgage data to demonstrate the advantage of the true maximum likelihood method over the quasi-maximum likelihood method.
Specifically, we randomly select a 1000-loan sample, a 2000-loan sample and a 5000-loan sample from a population of subprime mortgage loans’ performance history provided by LoanPerformance.com (LP), now part of CoreLogic. Since our purpose is not to perform a thorough analysis of the data, but to illustrate the difference between two statistical methods, we select a relatively homogenous mortgage pool8 and use only three variables: loan age, FICO score, and Loan-to-Value ratio as regressors in the observed loan specific heterogeneity.

Table 3 shows the summary statistics. In the full 5000 loan sample, the average duration is 8 quarters and 60% of loans were still active at the end of observation period. There are roughly the same percentage (20%) of cumulative prepay rate and default rate between the origination and the end of 2009. As such, our selection of sample reflects to the latter and less damaging case in Monte Carlo simulation study covered in Section 5.1 – the case with equal magnitudes of prepayment and default risks. The high default rate for these loans is not surprising due to the severe declining in nationwide home prices since their peak in summer of 2006.

<table>
<thead>
<tr>
<th>Table 3. Summary Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration (Quarter)</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Std</td>
</tr>
<tr>
<td>Max</td>
</tr>
<tr>
<td>Min</td>
</tr>
</tbody>
</table>

Table 4 shows the estimation results using three different methods: The true-maximum

8 These samples contain loans that are thirty year fix-rate mortgages originated in 2006 and 2007, FICO>675, LTV<90,
likelihood method, the quasi-maximum likelihood method, and a third labeled as Cox bivariate proportional hazard. The third method, used as only a benchmark here, constitutes the estimations of two separate single-risk proportional hazard models, one for prepayment and the other for default. When estimating the single risk proportional hazard model for prepayment, observed default (Type D) is treated as a non-informative censoring the same as way as observed censoring (Type C). Same treatment is used in the estimation of single-risk proportional hazard model for default.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Risk</th>
<th>Variable</th>
<th>True-likelihood</th>
<th>Quasi-likelihood</th>
<th>Cox Bi-variate</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</td>
<td>Prepay</td>
<td>FICO</td>
<td>0.19 ( 0.09 )</td>
<td>-0.4 ( 0.22 )</td>
<td>0.19 ( 0.09 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LTV</td>
<td>-1.18 ( 0.18 )</td>
<td>-1.14 ( 0.30 )</td>
<td>-1.21 ( 0.18 )</td>
</tr>
<tr>
<td></td>
<td>Default</td>
<td>FICO</td>
<td>-0.71 ( 0.10 )</td>
<td>-0.03 ( 0.23 )</td>
<td>-0.74 ( 0.11 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LTV</td>
<td>3.83 ( 0.25 )</td>
<td>3.75 ( 0.46 )</td>
<td>4.07 ( 0.32 )</td>
</tr>
<tr>
<td>2000</td>
<td>Prepay</td>
<td>FICO</td>
<td>0.19 ( 0.14 )</td>
<td>-0.76 ( 0.32 )</td>
<td>0.36 ( 0.14 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LTV</td>
<td>-0.97 ( 0.29 )</td>
<td>-0.45 ( 0.51 )</td>
<td>-1.69 ( 0.28 )</td>
</tr>
<tr>
<td></td>
<td>Default</td>
<td>FICO</td>
<td>-0.82 ( 0.16 )</td>
<td>0.18 ( 0.29 )</td>
<td>-0.86 ( 0.17 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LTV</td>
<td>3.48 ( 0.38 )</td>
<td>2.68 ( 0.76 )</td>
<td>3.52 ( 0.46 )</td>
</tr>
<tr>
<td>1000</td>
<td>Prepay</td>
<td>FICO</td>
<td>0.03 ( 0.21 )</td>
<td>-0.78 ( 0.45 )</td>
<td>0.16 ( 0.20 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LTV</td>
<td>-1.01 ( 0.40 )</td>
<td>-0.22 ( 0.65 )</td>
<td>-1.65 ( 0.40 )</td>
</tr>
<tr>
<td></td>
<td>Default</td>
<td>FICO</td>
<td>-0.69 ( 0.22 )</td>
<td>0.07 ( 0.39 )</td>
<td>-0.61 ( 0.23 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LTV</td>
<td>3.38 ( 0.51 )</td>
<td>2.14 ( 0.92 )</td>
<td>3.75 ( 0.65 )</td>
</tr>
</tbody>
</table>

Standard errors are in parentheses.

The estimated parameters using the true-maximum likelihood method and the Cox bivariate proportional hazard model are of expected sign and magnitude, though the true-maximum likelihood method reports smaller standard errors. In contrast, the quasi-maximum likelihood method reports generally larger standard errors and some insignificant parameter estimates. In these three small samples, because of inaccurate approximation of the likelihood function, the signs on the estimated parameters of FICO are of the reverse sign and are insignificant. These inaccurate estimates that use the and the performance history is from the origination to 2009.
quasi-likelihood methods could lead to the wrong economic interpretation that FICO does not matter for loan performance in these samples, though the estimates from the more robust methods suggest the opposite. Therefore, researchers should take caution in using the quasi-likelihood estimation method. The true maximum likelihood method is a preferred method to study grouped duration data with competing risks.

6. CONCLUSION

The previous two sections delivered two main messages. First, the models with nonparametric baseline hazards are fundamentally unidentifiable with grouped duration data. When a competing risks model is fit with grouped duration data, any meaningful inference has to stem from and hinge on parametric assumption of the baseline hazard. Second, under parametric assumption such as the piece-wise linear baseline hazards, the sample likelihood function has explicit analytical functional form. Direct estimation using the full likelihood function is feasible and easy. Under this assumption, approximation of the likelihood function is no longer necessary. Specifically, when the two risks are very different in hazard rate, the folk approximation using a 50-50 split is inaccurate.

Throughout this short paper we have limited to the case where there are two competing risks, where all the observed covariates are time-invariant, where the data grouping is regular in the sense that the continuous duration variable falls into intervals bounded by whole integers. Generalization to more than two competing risks only involves notational
complication. Treatment of time-varying covariates can be typically accommodated by making assumptions that the time trajectories of the X's are also piece-wise constant whose value changes are conforming to the interval of the duration variable. Non-regular data grouping can also be easily handled without difficulty, just as in the case of single-risk models (An, 2000).

**TECHNICAL APPENDICES**

**Appendix A. Identification for collinear \( \lambda_1(t) \) and \( \lambda_2(t) \).**

Proposition 1 of the paper states that the non-parametric baseline hazards functions are not identifiable with only grouped duration data unless the two base line hazards are collinear. We now show that if the two are collinear (albeit a very unlikely and unreasonable assumption), identification is assured.

If \( \lambda_1(t) = \rho \lambda_2(t) \), for some \( \rho > 0 \), then

\[
\Lambda_1(t) = \int_0^t \lambda_1(s) ds = \int_0^t \rho \lambda_2(s) ds = \rho \int_0^t \lambda_2(s) ds = \rho \Lambda_2(t)
\]

\[
Pr(K_n - 1 < Y_n < K_n, R_n = 1 | x_n) = \int_{K_n-1}^{K_n} \lambda_1(t) \phi_{1n} \exp\left\{ - \Lambda_1(t) \phi_{1n} - \Lambda_2(t) \phi_{2n} \right\} dt \\
= \int_{K_n-1}^{K_n} \lambda_1(t) \phi_{1n} \exp\left\{ - \phi_{1n} \int_0^t \lambda_1(s) ds - \phi_{2n} / \rho \right\} \int_0^t \lambda_1(s) ds \right\} dt \\
= \int_{K_n-1}^{K_n} \lambda_1(t) \phi_{1n} \exp\left\{ - (\phi_{1n} + \phi_{2n} / \rho) \int_0^t \lambda_1(s) ds \right\} dt \\
= \frac{\phi_{1n}}{\phi_{1n} + \phi_{2n} / \rho} \left[ \exp\{ - \Lambda_1(K_n - 1) (\phi_{1n} + \phi_{2n} / \rho) \} - \exp\{ - \Lambda_1(K_n) (\phi_{1n} + \phi_{2n} / \rho) \} \right]
\]

**Appendix B. Proof of Proposition 2**

The central step is to calculate the integral inside the square bracket in equation (7).
Because the two baseline hazards are piece-wise constant and all the covariates are time-invariant, the risk-specific hazard function is constant in each interval. For example, 
\( \alpha_{jk} \phi_{jn} \) is the constant hazard for risk j in the interval \((k-1, k)\).

Let \( \psi_{jn} = \alpha_{jk} \phi_{jn} \) for \( j=1, 2 \)

\[
\int_{K_{n-1}}^{K_n} \lambda_j(t) \phi_{jn} \exp\left\{-\Lambda_1(t)\phi_{jn} - \Lambda_2(t)\phi_{2jn}\right\} dt
\]

\[
= \left[ K_n \cdot \exp\left\{-\Lambda_1(K_n - 1)\phi_{jn} - \psi_{1n}(t - (K_n - 1)) - \Lambda_2(K_n - 1)\phi_{2jn}\psi_{2jn}(t - (K_n - 1))\right\} dt
\]

\[
= \psi_{1n} \cdot S(K_n - 1, K_n - 1 \mid x_n, \nu_{1n}, \nu_{2jn}) \cdot \exp\left\{\psi_{1n} + \psi_{2jn}\right\} \int_{K_{n-1}}^{K_n} \exp\left\{-\psi_{1n} - \psi_{2jn}\right\} dt
\]

Notice the above integrand is just an exponential function of \( t \). Therefore

\[
\int_{K_{n-1}}^{K_n} \lambda_j(t) \phi_{jn} \exp\left\{-\Lambda_1(t)\phi_{jn} - \Lambda_2(t)\phi_{2jn}\right\} dt
\]

\[
= \frac{\psi_{1n}}{\psi_{1n} + \psi_{2jn}} \cdot S(K_n - 1, K_n - 1 \mid x_n, \nu_{1n}, \nu_{2jn}) \cdot \left[1 - \exp\left\{-\psi_{1n} - \psi_{2jn}\right\}\right]
\]

Appendix C. Simulation of Loan Activity Data under Continuous Time, Competing Risks, Proportional Hazard Model

C.1 Purpose

To generate a data set that consists of N loans with quarterly activities until loan termination due to one of the two competing risks, or until observation window ends.

Observation window is from \((0, M)\). Loans originate uniformly in interval \([0, A)\) with \( A < M \).

The data generating process is in continuous time. But the realized data are grouped into
unit intervals between integers. Output: \( Y \in \{0, 1, M\} \) integer-valued duration; \( R \in \{1, 2, 0\} \) whether the duration is a termination due to prepayment or default, or it is right-hand censored spell.

**C.2 Notation:**

- \( N \): Sample Size (number of loans)
- \( X_n \): Vector of Time-invariant covariates
- \( S_n \): Origination time
- \( T_{1n} \): Latent Duration until Prepayment
- \( T_{2n} \): Latent Duration until Default
- \( Y_n \): = min\{\( T_{1n} \), \( T_{2n} \)\}, duration until loan termination

**C.3 Assumptions:**

**Assumption 1** \( T_{1n} \) and \( T_{2n} \) are conditionally independent given the observed covariates.

**Assumption 2** The risk specific hazard function is of the proportional hazard type:

\[
h_j(t \mid X_n, Z_n) = \lambda_j(t) \exp \{x_j \beta_j\}, j = 1, 2.
\]

where \( t \) is continuous loan age. The baseline hazard function \( \lambda_j(t) \) is piece-wise constant.

**C.4 Two Useful Results**

**Lemma A (Probability Integral Transform)**

Let \( F(y) \) be a CDF for a continuous random variable. Let \( S(y) = 1 - F(y) \). Let \( U \) be a uniform random variable on \((0, 1)\). Then \( S^{-1}(U) \) is distributed as \( F(y) \).
Lemma B Proportional Hazard

Let $W$ be a positive valued random variable, whose distribution is characterized by its hazard function $h(t)$. Let $H(t) = \int_0^t h(s) \, ds$. Then the survivor of $W$ is $S(y) = \exp\{-H(y)\}$.

The immediate implication of Lemmas A and B is that if we want to draw random variable $Y$ with hazard $h(t)$, we would draw $U$ from uniform and solve $W$ from $\exp\{-H(W)\} = U$.

C.5 Simulation of $T_{i\pi}$ or $T_{2\pi}$ with Proportional Hazard:

The two hazard functions in (1) are piece-wise constant. Define

$$\theta_{jnk} = h_j(k \, | \, x_n), \text{ for } j = 1, 2.$$ 

The integrated hazard function for risk $j$ is piece-wise linear:

$$H_j(t \, | \, x_n) = \sum_{i=1}^{k-1} \theta_{jns} + \theta_{jnk} [t - (k - 1)], \text{ for } t \in (k - 1, k]$$

To simulate a random variable $T_{i\pi}$ with piece-wise constant hazard function $\theta_{i\pi}$, provoking the two lemmas, we would simply do the following:

**Step 1.** Draw $U_n$ form the uniform $(0,1)$.

**Step 2.** Solving $T_{i\pi}$ from $U_n = \exp\{-H_i(T_{i\pi} \mid X_n)\}$, that amounts to:

Let $V_n = -\log(U_n)$

If $V_n < \theta_{i1}$ then $T_{i1} = V_i / \theta_{i1}$;

If $\theta_{i1} \leq V_n < \theta_{i1} + \theta_{i2}$ then $T_{i\pi} = 1 + V_n / \theta_{i2}$;
C.6 Simulation of $Y_n$ and Data Grouping and Censoring

Step 1. Generate integer $S_n$ from $[0, A]$

Step 2. Input $A, M, X_n$, and all the parameters.

Step 3. Calculate and store all $\theta_{jnk}$ for $k = S_n, S_n + 1, ..., M$.

Step 4. Draw $T_{1n}$ and $T_{2n}$ according to Section 5.

Step 5. $Y_n = \min\{T_{1n}, T_{2n}\}$. $R_n = 1$ if $T_{1n}$ is smaller; $R_n = 2$ if $T_{2n}$ is smaller.

Step 6. If $Y_n < M - S_n$ then $K_n = \text{integer}(Y_n)$ else $K_n = M - S_n$ and $R_n = 0$.

REFERENCES


**ACKNOWLEDGEMENTS**

This is an expansion of an earlier version circulated as An (2004). Both authors are with Fannie Mae, formally known as Federal National Mortgage Association. The opinion expressed in this paper is solely of the authors and does not necessary represent that of Fannie Mae. For correspondence send email to mark_y_an@fanniemae.com or ordinary mail to Dr. Mark An, Fannie Mae, Mail Stop 1H1S-03, 3900 Wisconsin Ave., NW, Washington, DC, 20016. The authors thank the editor and three referees for their constructive suggestions that improved the paper.