# A Reduced-Form Approach to Mortgage Valuation

by

Alexey A. Smurov

(Under the direction of Donald C. Keenan)

## Abstract

In recent years, the reduced-form approach to valuation has become widely used in asset pricing. Unlike earlier structural models, reduced-form models do not require that data on the underlying asset be available, which makes them a convenient tool for empirical applications.

I develop a simple reduced-form model of mortgage pricing where both default and prepayment are exogenous, and empirically estimate it using historical data on about one million fixed-rate residential mortgages. The empirical analysis proceeds in several stages. First, individual mortgage histories are used to estimate effects of exogenous variables, those being loan-specific characteristics, on hazards of termination. In the second stage, particle filtering is employed to estimate stochastic hazard processes. Once the parameters of hazard processes, as well as those of the conventional risk-free terms structure, are obtained, I use calibration to convert physical processes into risk-neutral ones. After that, tests of pricing performance are conducted.

Index words: Asset Pricing, Mortgage, Reduced-Form Pricing, Particle Filter, Term Structure

# A Reduced-Form Approach to Mortgage Valuation

by

Alexey A. Smurov

B.A., Tomsk State University, 1998

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2004

# A Reduced-Form Approach to Mortgage Valuation

by

Alexey A. Smurov

Approved:

Major Professor: Donald C. Keenan

Committee: James B. Kau
Scott E. Atkinson

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2004

# Acknowledgments

I would like to thank my major professor, Dr. Donald Keenan, who has devoted a great amount of time and effort to helping me complete this project. I am also thankful to my dissertation committee members, Professors Kau and Atkinson, for their constructive comments and valuable advice.

Special thanks goes to Garland Durham, for providing the code originally used for estimation of default and prepayment processes; Robert Bliss, for his term structure data; Brian McCall and Insan Tunali, for their help with the proportional hazard model.

## Table of Contents

Appendix

# Chapter 1

# Introduction

Until very recently, the structural approach to option pricing has been the preferred method for building a theoretically sound and empirically testable model of mortgage valuation. The logic behind the structural approach goes back to Merton's [107] work on corporate debt, describing default as a call option on the firm's assets. By exercising this option, the equity holders can exchange the firm's assets for its outstanding debt.

In the case of a mortgage, the underlying asset takes the form of a house serving as a collateral on the loan. Mortgage termination (via either default or prepayment) in the structural framework is determined endogenously, by rational decisions of borrowers maximizing the value of their portfolios, so the structural form of modelling is often termed the endogenous approach. Building a structural model involves an explicit model of the evolution of the value of the house, which constitutes a considerable problem in applications, due to limited availability of data on housing prices[1] and resulting difficulties in estimation.[2] The problem is even more severe for corporate debt, where the underlying asset is virtually never observable and the structure of debt contracts is often extremely complex.

Even when data on the value of the underlying asset are available, another problem is that too much or too little default and prepayment are usually predicted

[1] In particular, the volatility of the individual house price becomes crucial for empirical estimation.

[2] For examples of structural models of mortgage valuation, see Titman and Torous [131], Kau et al. [91], or the survey of Kau and Keenan [86].

by straightforward applications of the Black-Scholes model. This problem has led to more complicated models, ones that distinguish between endogenous ("optimal") and exogenous ("suboptimal") termination (Dunn and McConnell [59]), or ones that introduce unobservable as well as observable transaction costs of default (e.g. Deng et al. [45]).

Reduced-form models, also originated in the corporate debt literature, avoid most of the aforementioned problems. In models of this type, default and prepayment are not determined endogenously, but are imposed as exogenous events, described by their own arrival processes, so that the actual event of default or prepayment always comes as a surprise. As a result, there is no problem of distinguishing between "optimal" and "suboptimal" default and prepayment: borrowers may or may not be behaving optimally in the structural sense.[3] The general concept thus resembles models of the term structure, e.g. the one of Cox, Ingersoll and Ross [38], but with the discount factor augmented by the hazard of default, more than it resembles the usual Black-Scholes model [12].

The more flexible nature of the reduced-form approach also allows it to incorporate various exogenous variables into the pricing model, thus making it more amenable to empirical analysis compared to the tighter structural approach. However, this goal is achieved without loss of consistency with the general theory of asset pricing: in the end, the price of a contingent claim is still determined by the risk-neutral expectation of the present value of its future cash flows.

The primary goal of the present dissertation is to introduce a reduced-form model into mortgage valuation, similar to models previously applied to corporate debt, as

[3] Although basic structural models do not permit a reduced form in the usual sense, it is possible to use the concept of asymmetric information to build a structural model with a well-defined reduced form. See, for example, Duffie and Lando [54], Giesecke [67].

in Madan and Unal [103], Duffie and Singleton [56] or Lando [98][99].[4] This model includes two apparently distinct types of mortgage termination: default and prepayment, with particular emphasis on default. Although it is a rare event, mortgage default can be analyzed with a relatively high degree of accuracy given the large amount of data available.

Once the model is built, I proceed to estimate it empirically. The data used for estimation contain payment histories of approximately one million mortgages originated over the period 1970-2001 and observed over the period 1990-2001. At any month during the observation period, it is known whether the borrower continues to make mortgage payments, prepays, or defaults on the loan. Several loan-specific characteristics at the time of origination are also available. The additional external information I am using in estimation is the data on the evolution of the standard Treasury bond and Treasury bill series during the observation period, used for estimation of the risk-free term structure, and the general housing price index provided by Freddie Mac.[5]

The dissertation is organized as follows. Chapter 2 summarizes existing theoretical and empirical literature on mortgage pricing. Chapter 3 describes the theoretical pricing model used for empirical estimation. Chapters 4 to 7 analyze the empirical analysis involved. In Chapter 4, I estimate a mortgage duration model using the discrete stratified partial likelihood approach. Individual mortgage duration data are used to obtain estimates of the effects of exogenous variables, such as original loan-to-value ratio, loan size, etc., on loan termination. In Chapter 5, empirical analysis

[4] For a thorough discussion of such models, see Duffie [53] or the monographs by Bielecki and Rutkowski [10], Duffie and Singleton [57].

[5] Ability to estimate virtually all parameters required for mortgage pricing using just the data on mortgage history and risk-free interest rates is yet another attractive feature of the reduced-form approach. The only parameters that I have to set to some known values are the tax rates. Empirical implementation of structural models, on the other hand, usually requires most parameter values to be imposed using external information, e.g. as in Titman and Torous [131].

of loan termination processes via particle filtering is performed. Chapter 6 describes the term structure estimation, which is a traditional problem in bond pricing literature. Chapter 7 develops a calibration model used to estimate risk adjustment parameters and recovery on default. Some pricing exercises are also performed in this chapter. Finally, Chapter 8 concludes, and the Appendix contains the pseudo-code of the particle filter algorithm.

# Chapter 2

# Mortgage Pricing: A Literature Review

In this chapter, I review existing literature related to mortgage pricing, with specific emphasis on papers relevant to either theoretical or empirical aspects of the model employed in this dissertation. The first section of the chapter analyzes theoretical literature related to the structural (option-theoretic) approach to mortgage pricing, the second section describes relevant theoretical reduced-form literature, and the third one contains a discussion of empirical mortgage duration literature. Additional short reviews of empirical research relevant to different steps of the estimation process are included in corresponding chapters.

## 2.1 The Structural Approach

The basic idea behind the structural approach to mortgage valuation is based on the seminal work of Black and Scholes [12] and Merton [107].

For a fixed-rate mortgage, which is the focus of this dissertation,[1] the value of the mortgage at any point in time depends on the borrower's right to default or prepay the loan. From the borrower's perspective, the right to default, i.e. terminate mortgage payments in return to giving up possession of the house, represents a put option. More precisely, it is an European compound option, since default becomes

[1] For adjustable-rate mortgages, the second biggest category of mortgage loans, usual references include Buser, Hendershott and Sanders [20], Cox, Ingersoll and Ross [37], Kau et al. [87][90][92], Schwartz and Torous [125].

obvious only on payment dates,[2] and these dates usually occur more than once during the term of the loan. Prepayment, on the other hand, may be considered an American-style call option, whereby the borrower has the right to obtain full ownership of the house at any time by paying off the loan.

The joint right to prepay or default at any time $t$ can be considered a complex termination option, $J(r, H, t)$, where $r$ represents the risk-free spot rate and $H$ is the value of the house. The value of this option has to be taken into account when one is performing mortgage valuation.

Like any derivative asset, a mortgage with maturity at time $T$ can be valued as a present value of its future payoff, i.e.

$$X(r, H, t) = E^{\mathbb{Q}}[X(r(T), H(T), T)\mathbf{e}^{-\int_t^T r(s)ds}]$$

which is the expectation under the market's risk-neutral measure $\mathbb{Q}$.

The price of the underlying asset (the house price) and the instantaneous interest rate are assumed to be following the risk-adjusted diffusions

$$dH = (\mu_H - \lambda_H \sigma_H)dt + \sigma_H dz_H$$

and:

$$dr = (\mu_r - \lambda_r \sigma_r)dt + \sigma_r dz_r$$

respectively, with $dz_H dz_r = \rho(r, H, t)dt$.

Using the local expectation hypothesis, $\lambda_r = 0$, as well as the argument that the expected risk-adjusted return on a house is simply the risk-free rate, i.e.

$$(\mu_H - \lambda_H \sigma_H)/H + s = r$$

[2] In addition, a rational borrower will always choose to default on payment dates, to avoid rent payments.

where $s(r, H, t)$ is the house rental rate, one can obtain the final forms for the diffusions, as

$$dH = (r - s(r, H, t))Hdt + \sigma_H dz_H$$

and

$$dr = \mu_r dt + \sigma_r dz_r$$

Note that any dependence of pricing on the mean rate of house depreciation has disappeared, in a manner consistent with that on other option-theoretic models of asset pricing.

Given the above information, one can now build a fundamental pricing partial differential equation (PDE):

$$\frac{1}{2}\sigma_r^2\frac{\partial^2 X}{\partial r^2} + \rho\sigma r\sigma H\frac{\partial^2 X}{\partial r\partial H} + \frac{1}{2}\sigma_H^2\frac{\partial^2 X}{\partial H^2} + \mu_r\frac{\partial X}{\partial r} + (r - s)H\frac{\partial X}{\partial H} + \frac{\partial X}{\partial t} - rX = 0$$

which is usually solved in backward time using appropriate boundary and terminal conditions.

Earlier structural models were built on the assumption of prepayment but no default, and were applied to mortgage-backed securities, usually issued by a government-sponsored agency, and with a guarantee against default. Examples of such models with only a call option present include Dunn and McConnell [58][59], Buser and Hendershott [19] and Brennan and Schwartz [16]. In mid-eighties, the first attempts to model the put option were made. Originally, option-theoretic models of default had no stochastic term structure involved (e.g. Cunningham and Hendershott [40]), which is a rather unrealistic assumption for such long-lived securities. Later, a stochastic interest rate was included as a state variable, but default remained the only option in the model, as in Epperson et al. [63]. Eventually, models with a joint option to terminate a mortgage via default or prepayment were developed by Kau et al. [88], Leung and Sirmans [100], Kau et al. [91]. The remainder of the section discusses some influential option-theoretic models of mortgage pricing in greater detail.

Dunn and McConnell [59] use pass-through securities issued by the Government National Mortgage Association (GNMA or Ginnie Mae) to build the first structural model of mortgage pricing, but with only a call option present. The model is specified using the instantaneous interest rate as a state variable responsible for the prepayment decisions of mortgagors. The authors indicate two types of prepayment: “optimal,” i.e. completely determined by financial reasons, and “suboptimal,” i.e. determined by factors other than the interest rate. Dunn and McConnell [58] further analyze their original model and compare its predictions to those of older, more traditional certainty models.

Quite a few similar, more detailed models were developed in mid-eighties. Buser and Hendershott [19] introduce taxes and refinancing costs into a prepayment-only model. Brennan and Schwartz [16] allow for prepayment penalties. Other examples include Hall [73], Pozdena and Iben [114], among others. Schwartz and Torous [124] also use the option pricing approach to value GNMA mortgage-backed securities. Their model is based on the term structure model of Brennan and Schwartz and employs a deterministic prepayment function with a logistic trend,[3] with the interest rate entering as a covariate.

Cunningham and Hendershott [40] develop the first structural-form model of mortgage default. On the other hand, no interest-rate uncertainty is included in the model, which makes it somewhat peculiar.

[3] Estimation of hazard functions of prepayment/default is usually a major objective of empirical mortgage duration literature. This is why I also include this paper in the survey of empirical duration literature, where I describe the empirical technique employed in greater detail.

Epperson et al. [63] introduce the idea of looking at mortgage default as a compound European put option in a stochastic interest rate environment. While the borrower's option to prepay is ignored, specific attention is paid to pricing default insurance. Kau et al. [88] use the same approach to build a pricing model that is applied to Freddie Mac's multifamily participation certificates backed by commercial mortgages. The same study is extended by Kau et al. [89], who provide more detailed numerical results of the same model.

Titman and Torous [131] build a model of commercial "bullet" mortgages, with prepayment prohibited. The model is based on two correlated state variables, a square-root interest rate process and a lognormal house price process. The interest rate process is estimated, while the house price process parameters are exogenously set by the authors, due to a lack of data.

Schwartz and Torous [126] extend their earlier model in [124] to allow for the possibility of default. Even though "default-free"[4] GNMA securities are used, the authors argue that, since default occurs under different conditions than prepayment, its effect cannot be ignored by the investor.

Leung and Sirmans [100] model default and prepayment options in a discrete-time framework using a lattice approach. Overall, their results are similar to those of Kau et al. [91].

Kau et al. [91] develop what can be considered a complete option-theoretic model of residential mortgage pricing, by assuming that the borrower has a joint option to prepay or default. In addition to financial termination, i.e. the one fully explained by the option pricing theory, the authors also have "nonfinancial termination" (the same as the "suboptimal" termination of Dunn and McConnell). Kau et al. [93]

[4] In case of default, GNMA pays the investor the full amount of the outstanding balance, thus making the payoff in the case of default identical to that in case of prepayment.

further develop the joint option model of financial termination, but without any "suboptimal" prepayment or default.

Kau and Keenan [86] provide a thorough discussion of theoretical work on mortgage pricing. To this day, this is probably the most complete survey of option-theoretical models of mortgage pricing. Describing the then current state of the art, rather then individual papers, the authors analyze a considerable body of theoretical literature developing the option-theoretical approach to mortgage pricing, as well as possible extensions of the general framework.

It appears that the theoretical literature on structural-form mortgage pricing has not developed much since 1995. Some theoretical work has been done on heterogenous transaction costs (e.g. Stanton [127]), as well as on extending the term structure of interest rates to include more state variables (e.g. Brunson et al. [18]). An interesting study by Chatterjee et al. [25] compares different option-pricing models. The authors provide empirical results comparing the pricing performance of models involving one interest rate state variable, two interest rate variables, an interest rate and a building variable, and combinations thereof. All models exhibit satisfactory results, with the "spot-value" (the spot rate plus the house price) model being the superior one. Most research, however, has been done on empirical applications of the theory, in particular, testing the extent to which the option-pricing model explains the behavior of historical mortgage default and prepayment over time. Some of this empirical work is discussed in Section 3 of this chapter.

## 2.2 The Reduced-Form Approach

This section would logically provide a quick review of reduced-form models of mortgage pricing. Unfortunately, there are no such models; at least, none of them that have been published.[5] There exists, however, an extensive literature on reduced-form models of corporate debt, which, with some work, can be extended to mortgage pricing. In this review, I will concentrate on the most influential papers on reduced-form modelling, as well as on some specific models directly related to the one used in this dissertation.

Reduced-form models analyze default probability and recovery as stochastic processes varying over time. These stochastic processes determine the price of credit risk, i.e. the spread between the defaultable and the default-free bond. Although the probability of default and the amount of recovery do not have to be formally linked to the value of the underlying asset, such a link can be easily incorporated in a reduced-form model, under some mild assumptions regarding the information available to the bondholder and the equity holder.

In a general reduced-form model, the valuation equation for a defaultable zero coupon bond that promises to pay $1 at time $T$ can be expressed as

$$X(r,t) = E^{\mathbb{Q}}[(\mathbb{1}_{\{\tau>T\}} + W(\tau)\mathbb{1}_{\{\tau\leq T\}})\mathbf{e}^{-\int_t^T r(s)ds}]$$

where $\tau$ is the stopping (default) time, $W(\tau)$ is the recovery on default, and $\mathbb{1}$ is the usual indicator function. In the event of default, the bondholder gets the amount $W$, which may be random variable or a constant. If default does not occur, the bondholder get the promised amount of one dollar.

Under different assumptions, the model may be reexpressed in terms of the hazard, the intensity, or the compensator of the default process. Most reduced-form literature employs the intensity-based approach, for example, Duffie and Singleton [55] [56], Lando [98] [99], or Duffee [49]. Compensator-based modelling, a more recent

[5] There are, however, several working papers. See, for example, De Giorgi [42].

development, is still a rather uncommon approach (see Giesecke [67] [68]). Hazard-based models are also discussed in Schönbucher [123] and Bélanger et al. [8].

Another important difference is based on the way recovery on default is modelled. There are three major approaches to this problem: recovery of face value (e.g. Duffie [51], Duffee [49], Driessen [47]), recovery of market value (e.g. Duffie and Singleton [56]), or recovery of an otherwise identical default-free bond (Jarrow and Turnbull [82]), the latter being the least common. The remainder of this section contains a more detailed discussion of major theoretical reduced-form models and their empirical applications.

Jarrow and Turnbull [82] introduce one of the earliest reduced-form models of option pricing, along the lines developed for exchange rate models (making an analogy between default and exchange devaluation), under the assumption that the interest rate process and the bankruptcy process are independent. Jarrow et al. [80] refine this model, by letting the bankruptcy process depend on firms' credit ratings, while still maintaining the assumption that the processes of default-free term structure and default are independent.

Madan and Unal [103] separate the arrival risk of default, i.e. the probability of default's occurring, from the magnitude risk of default, i.e. the severity of loss in case of default. This is an extension of the Jarrow-Turnbull model, with severity modelled as a stochastic process, while the arrival rate of default is modelled as a function of the ratio of the firm's value to the money market account, but still independent of the interest rate process (the same assumption holds for recovery). The theoretical model is then implemented using monthly rates on certificates of deposit of thrift institutions, where the presence of insured and uninsured certificates of deposit for the same thrift allows the authors to identify both an arrival risk and a magnitude risk.

Duffie, Schroder and Skiadas [50] provide a theoretical foundation for most continuous-time reduced-form models, based on mathematical tools described, for example, in Brémaud [15]. Lando [99] also provides a thorough discussion of the role of doubly stochastic Poisson processes (also known as Cox processes) in the bond-pricing literature, and develops a general model similar to the model of Jarrow et al. [80], but without the assumption of independence between the intensity of default and other state variables (e.g. term structure of risk-free rates, credit ratings, etc.).

Another discussion of intensity-based models can be found in Duffie [51], who analyzes two general classes of models, one based on affine diffusions of state variables, and the other one built in a general Heath-Jarrow-Morton setting. Both models are based on the assumption of the recovery of face value (RFV). A similar setup is used in Duffie [52], where the author analyzes a model allowing for multiple termination processes.

Duffie and Singleton [56] develop reduced-form models in both discrete and continuous time under the convenient recovery or market value (RMV) assumption. The risky discount rate is modelled as a function of the risk-free rate process, the default hazard process, the recovery process, and the liquidity process.

Lando [98] provides a review of literature related to valuation of defaultable bonds, with an explicit emphasis on intensity-based reduced-form models that were in existence at the time the review was written (it was published in 1997). The author emphasizes the fact that there is a close relationship between structural models and reduced-form ones, e.g., the value of the underlying asset, the most important variable in a structural model, can easily be incorporated as a variable affecting default in a reduced-form setting. The author provides a thorough discussion of several existing intensity-based models, classifying them as those that impose the independence between the interest rates and the default intensity (e.g. Jarrow-Turnbull [82], Jarrow-Lando-Turnbull [80]) and those that don't (e.g. Duffie-Singleton [56], Duffie-Schroder-Skiadas [50]).

Since 1997, quite a few studies have employed reduced-form modelling. The two papers I consider particularly interesting are Duffie and Lando [54] and Giesecke [67].

Duffie and Lando [54] discuss a model that includes both stochastic evolution of the value of the underlying asset and a default-arrival intensity process. The link between the traditional structural approach and the reduced-form one is built using the assumption of incomplete (or noisy) accounting information available to the bondholders. The authors prove that this feature of the model allows one to explain the empirical fact that credit spreads for corporate debt are always positive, even for bonds with near-zero maturity.

Giesecke [67] further examines the link between reduced-form and structural models of default. The new feature of Giesecke's work is the use of the compensator (as discussed in more detail in Chapter 4) of the default process to analyze defaultable security prices. The approach is more general than the intensity approach, since intensity does not necessarily exist in reduced-form models. Specific attention is paid to information asymmetries between bondholders and shareholders, regarding not only the evolution of firm's assets, but also the stochastic default threshold, i.e. the level of asset values that triggers the shareholders' decision to default. Several examples are considered, including the one of Duffie and Lando [54].

Although theoretical literature of reduced-form modelling is quite extensive, only a few empirical studies have been published. Duffie and Singleton [55] develop a reduced-form model for pricing plain-vanilla fixed-for-floating swaps, based on theoretical models discussed in [56]. The authors employ the extended independent two-factor Cox-Ingersoll-Ross term structure model of Pearson and Sun [111]. After conducting some impulse-response analysis, the authors point out the importance of the liquidity effect in explaining the spread between Treasuries and swaps, but emphasize that the credit risk is also of great importance. Unfortunately, since market data are used for the analysis, the authors cannot proceed any further, i.e. they cannot identify the riskless yield, the liquidity convenience yield, or the credit risk adjustment separately.

Duffee [49] develops another empirical model, similar to those described by Duffie and Singleton [56][55], and estimates it using a sample of corporate bonds. The risk-free term structure is modelled as a "translated" (i.e. extended) two-factor CIR process, the risky rate is modelled as a single-factor CIR process correlated with the risk-free rate,[6] while recovery is assumed to be a constant fraction of the promised payment (the recovery of face value assumption). Duffee also includes the effects of liquidity, as well as state taxes, in the model, and finds presence of a substantial credit risk component, as well as a liquidity premium component, in the spread between corporate bonds and Treasuries, in a manner consistent with the findings of Duffie and Singleton [55].

Driessen [47] provides an application of intensity-based modelling in the general pricing framework of Duffie and Singleton [56]. Following Duffee [49], Driessen builds a model based on the RFV assumption, and includes a two-factor risk-free term structure, as well as firm-specific effects, as determinants of credit spread. The model is estimated by the extended Kalman filter (QML), using data on Treasuries and corporate bond yields. Unlike Duffee's work, two factors common for all firms are included as determinants of default. The empirical results point out that these common factors are priced by the market, while the individual firm factors of Duffee are not. In the final part of the paper, the author investigates the empirical validity

[6] The most controversial fact about Duffie's choice of parametrization is that it allows negative default rates, a technical impossibility. In fact, some estimated values of default hazard in the paper are negative.

of the conditional diversification hypothesis[7] for corporate bonds of Jarrow, Lando and Yu [81]. The analysis is performed comparing historical and model-predicted hazard rates. The empirical results provide evidence against the conditional diversification hypothesis, with the jump risk adjustment parameter value estimated to be around six. When the model is later extended to allow for a state tax effect and a liquidity premium, the effect of the jump risk premium still remains significant.

## 2.3 Empirical Mortgage Duration Literature

In this section, I provide a relatively non-rigorous discussion of empirical studies that employ various econometric techniques to analyze mortgage termination via default and/or prepayment. This review includes major applied studies that estimate conditional probabilities (hazard rates) of termination, as well as the various factors affecting those probabilities. More emphasis will be put on more recent developments in the field; however, some influential early studies will be discussed as well.

Estimation of the underlying hazard of mortgage termination[8] is usually performed in the general framework of duration analysis,[9] which has been a very widely applied empirical tool in mortgage-related literature during the past twenty years.

Since this dissertation has an empirical emphasis, I will base the classification of papers discussed on the empirical technique used. This straightforward classification produces five basic classes of models: (1) fully parametric models for continuously

[7] According to the conditional diversification hypothesis, under certain assumptions, one of which requires a countably infinite number of bond-issuing firms to be present in the economy, the default intensities under the true probability measure and the equivalent martingale measure are approximately equal. Intuitively, this means that default jumps can be diversified away and are, therefore, not priced by the market.

[8] If there are no covariates present, consistent estimation of termination hazard can be performed non-parametrically, e.g. via the Kaplan-Meier method. Since non-parametric methods are rarely used as a primary method of estimation in mortgage studies, I will not concentrate on them in this review.

[9] See Keifer [94] for a through, but slightly dated, review of a variety of duration models used in general economics literature.

observed data, (2) semiparametric proportional hazard/partial likelihood models for continuously observed data,[10] (3) proportional hazard models for grouped data, (4) multinomial logit models for grouped data, and (5) the semiparametric approach (SPA) for grouped data. Studies that use several empirical methods are classified on the basis of the method indicated as the preferred one by the author(s). In the following subsections, I describe the logic and the basic assumptions behind each of the models, and provide examples of papers employing each one to analyze mortgage termination.

### 2.3.1 Fully Parametric Duration Models for Continuously Observed Data

When one is willing to assume a continuous distribution for duration time $t$, specifying a likelihood function becomes a relatively straightforward exercise. Popular distributional choices produce the Weibull hazard model, the lognormal hazard model, and the log-logistic hazard model. For a more detailed classification of fully parametric models, see Lancaster [97].

Random censoring, unobserved heterogeneity and competing risks of spell termination are easily incorporated into fully parametric models. However, though they have become very popular in biometrics and other natural science literature, the use of parametric methods in economics, and in mortgage studies in particular, has been limited. In mortgage duration literature, more flexible semi-parametric methods of estimation are usually preferred to fully parametric models, since they that do not require as many restrictive assumptions about the shape of the baseline hazard.

[10] Proportional hazard models are a part of a more general class of mixed proportional hazard models, the major difference between the two being the presence of a multiplicative unobserved heterogeneity factor in the mixed class. For a thorough survey of mixed hazard models and their applications in economics, see Van den Berg [132].

Schwartz and Torous [124] study prepayment using a continuous-time version of the proportional hazard model, with an aggregated sample of insured 30-year single-family mortgage pools. The baseline hazard is modelled as a log-logistic function, i.e. $\ln(t)$ is assumed to have a logistic distribution.[11] The only covariates included in the model are the interest rate spread and a variable used to model the "burnout effect".[12] Due to high degree of data aggregation, any detailed empirical analysis is impossible, and the authors proceed to using the estimated coefficients in a mortgage pricing model.

### 2.3.2 Partial Likelihood Models for Continuously Observed Data

In 1972, David R. Cox proposed an estimation technique that allows one to estimate covariate effects without specifying the form of the baseline hazard function (Cox [36]). In the following years, this method, promptly labelled Cox partial likelihood [35], became one of the most widely used semiparametric methods of estimation in duration analysis, with mortgage literature being no exception.[13]

The semiparametric nature of Cox partial likelihood estimation, along with its ability to incorporate censoring, time-varying covariates and competing risks, make it a convenient tool for the analysis of mortgage data. Another big advantage of the

[11] Note that this is not what is usually referred to as a log-logistic model, but rather a proportional hazard model with a log-logistic baseline.

[12] The "burnout effect" means that as the size of a pool decreases, borrowers who are more likely to prepay have already done so.

[13] One of the most serious problems of Cox proportional likelihood is the presence of ties, i.e. situations in which several individual durations terminate exactly at the same time, which contradicts the continuous nature of the model. Another name for this phenomenon is interval censoring. Ties can be handled with a number of numeric methods, such as the Breslow [17] and the Efron [62] methods. The exact likelihood can also be used; however, it usually is relatively complicated and hard to compute (see Kalbfleisch and Prentice [84] for a discussion). Given estimates of covariate effects $\beta$, one can obtain non-parametric estimates of the baseline hazard, using the Kaplan-Meier method.

Cox proportional hazard model is its great flexibility, allowing one to obtain nonparametric estimates of the baseline hazard for any period in a mortgage's life. This is especially important since mortgages are long-lived securities, and the amount of mortgage data available for research is usually quite extensive, often measured in hundreds of thousands of loans. The incorporation of unobserved individual-specific heterogeneity is a little more difficult. Group-specific heterogeneity, on the other hand, is easily modelled. For all these reasons, the partial likelihood estimation has been widely employed in various studies analyzing mortgage default and prepayment.

Partial likelihood models of prepayment can be found in Green and Shoven [72], Quigley and Van Order [118], as well as Fu et al. [65]. Examples of partial likelihood models of default include Harmon [75], Vandell et al. [133], as well as Quigley and Van Order [119]. In a recent paper, Pavlov [110] uses the partial likelihood approach to empirically analyze three separate risks of mortgage termination: prepayment due to refinancing, prepayment due to moving to another residence, and default.[14]

### 2.3.3 Proportional Hazard Models for Grouped Duration Data

The duration data available to an econometrician are usually grouped, which means that individual durations are only observed to fall some time within a known time interval. The data format in this case closely resembles that of a panel dataset: for each individual covariate value, the duration data are represented as a binary response variable, 0 if the individual remains in the original state, and 1 otherwise. In the case of multiple exit destinations, there will be a number of such variables, each representing an exit into a particular state.

The original proportional hazard duration model designed to explicitly deal with such data was proposed by Prentice and Gloecker [115], and later extended by Meyer

[14] The model of default is estimated using probit analysis, without any time-related variables. This choice of the estimation technique is probably due to the small number of defaults in Pavlov's sample (49 defaults overall).

[108], Narendranathan and Stewart [109]. It was further developed and applied to competing risks by Han and Hausman [74], Sueyoshi [129][130], and McCall [105][106].

Left and right censoring, truncation, competing risks and, most importantly, unobserved heterogeneity can be relatively easily included in the model, which makes it very useful in the mortgage-related duration literature.

Follain et al. [64] perform the first application of Meyer's model for grouped data to test the extent to which the option-theoretic model explains the prepayment rate. Most applications of this model, however, concentrate on estimating the competing risks of default and prepayment simultaneously. Deng et al. [45] provide an empirical test of how well the option-theoretic model explains mortgage default and prepayment, in the presence of transaction costs, trigger events and borrower heterogeneity. The authors include both observed and unobserved heterogeneity, the latter modelled using the Heckman-Singer [76] approach, allowing for up to three distinct, but observed types of individuals to be present in the sample. Very similar empirical model are used by Ciochetti et al. [30][31] and Huang and Ondrich [77].

### 2.3.4 Multinomial Logit Models for Grouped Duration Data

Nothing in the world of duration models prevents the hazard function from being of non-proportional form. A widely used example of a model with a non-proportional hazard function is logit, binomial for the cases where only one reason for mortgage termination is studied, and multinomial for competing risks of loan termination.[15]

[15] See Sueyoshi [130] for a derivation of a logit specification in the continuous-time context.

Greater flexibility of the covariate effects in the logit model,[16] together with its other attractive features, such as a non-parametric baseline hazard and easy implementation using existing commercial software, have made it a popular choice in mortgage-related literature.[17] On the other hand, greater flexibility comes at the cost of less intuitive interpretation. While covariate estimates from proportional hazard models have straightforward interpretation as semi-elasticities for an individual risk, those from the logit model have to be interpreted in terms of risk ratios.

Logit was the first empirical duration model ever applied to mortgage termination data, and it still remains relatively popular, mostly due to its flexibility and relatively easy implementation. Archer et al. [6] use logit to study mortgage prepayment. Vandell and Thibodeau [134] apply logit in a study of the risk of default. Several studies employ a logit specification to study both risks of mortgage termination. Examples include Campbell and Dietrich [23], Cunningham and Capone [39], Clapp et al. [32][18], Calhoun and Deng [21], along with Ambrose and Sanders [3], among others.

### 2.3.5 The Semiparametric Approach (SPE)

Quigley and Van Order developed what they named a semiparametric approach (SPE) to estimate a proportional hazard model with competing risks. The technique is based on grouping observed data according to certain characteristics, estimating the total hazard of termination using non-parametric "local smoothing," and then

[16] It is possible to include unobserved heterogeneity into the logit model in a manner similar to that in the proportional hazard models; however, to my knowledge, there have been no empirical applications of such models in mortgage literature.

[17] As Cox [36] points out, the logit model is just as a discrete-time representation of the continuous-time Cox proportional hazard model. If there are no ties in the data, the two likelihood functions converge to the same expression and produce identical covariate estimates.

[18] Clapp et al. employ a competing risks model that includes prepayment for financial reasons and prepayment due to relocation as separate risks, at the same time allowing for default as yet another competing risk, in a manner similar to that in Pavlov [110].

employing the non-parametric estimates as dependent variables in a system of seemingly unrelated regressions.

This rather unconventional technique based on restrictive assumptions about the data has not become widely used in mortgage literature, producing only a couple of empirical applications. They can be found in Deng et al. [44] and Deng [43].

### 2.3.6 Concluding Remarks

In conclusion, I would like to make several general remarks about empirical duration models in mortgage literature.

First, it is clear that no matter which specification is chosen for the variables modelling the values of the call and the put options, they are always found to be the most important, in terms of both the economic and the statistical significance of the estimates obtained. Therefore, whenever possible, one should include some indicators of the book value of a mortgage, the market value of a mortgage, and the market value of the collateral in any empirical model of mortgage termination.

Second, it is also true that observed borrower heterogeneity does matter, though not as a much as loan-to-value ratios or interest rates. There exists a significant amount of observed borrower heterogeneity determined by trigger events, regional variation, as well as variation in income and social status of individual borrowers.

Third, unobserved borrower heterogeneity, as determined, for example, by different borrower risk-taking preferences or different general economic conditions, is also potentially important. Studies that include an unobserved component into termination models, e.g. Deng et al. [45], Follain et al. [64], provide evidence that estimated coefficients are adversely affected if unobserved heterogeneity is omitted from the model.

Fourth, despite their conceptual differences, different empirical models often produce similar estimates, which becomes especially obvious when a single study uses more than one empirical technique, as in Clapp et al. [32], Harmon [75], Ciochetti et al. [31]. This result is not unexpected, given that the likelihood for the logit model, for example, converges to the likelihood of the proportional hazard model for grouped data as the total hazard approaches zero. It also becomes identical to the Cox partial likelihood if there are no ties in data. This convergence is especially obvious in case of default, which usually has a very low risk and so a small number of durations terminating in each period.

In general, proportional hazard models still seem to be preferred in empirical studies, mostly due to their easy interpretation. Models that allow for unobserved heterogeneity are also preferred over the models that do not. However, models with unobserved heterogeneity are hard to implement and often produce no or very little indication of such heterogeneity, especially in the full likelihood framework with competing risks, as demonstrated, for example, by Ciochetti et al. [31][30]. In the present dissertation, I employ the stratified partial likelihood developed by Ridder and Tunali [121], which allows for unobserved group-specific heterogeneity, while at the same time allowing the baseline hazard to be estimated non-parametrically and separately from covariate estimates, therefore leading to greater flexibility and less computational difficulty in estimation.

# Chapter 3

## The Pricing Model

Assume a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and a given information filtration $\{\mathcal{F}_t : t \geq 0\}$ of $\sigma$-algebras of $\mathcal{F}$ satisfying the usual conditions.[1] Also assume that the interest rate $r$ follows a $\mathcal{F}_t$-measurable process, so that the savings account is well defined. These assumptions will be used throughout the dissertation, unless otherwise noted.

The general objective of this chapter is to describe valuation of a defaultable security, in particular, a fixed-rate mortgage with a possibility of prepayment, that pays $\mathbb{1}_{\{\tau<t(i)\}}M$ at a given time $t(i) > 0$,[2] where $\tau_d$ is the time of default, $\tau_p$ is the time of prepayment, $\tau = \tau_d \wedge \tau_p \equiv \min(\tau_p, \tau_d)$, $\mathbb{1}_{\tau>t(i)}$=1 if $\tau > t(i)$, $\mathbb{1}_{\tau>t(i)}$=0 if $\tau \leq t(i)$. The two random variables representing the stopping time, $\tau_d$ and $\tau_p$, are also assumed to be non-negative, bounded and $\mathcal{F}_t$-measurable.

The evolution of default and prepayment processes is described by two "latent" stochastic *baseline hazard rate processes* $\lambda_0{}^d(t)$ and $\lambda_0{}^p(t)$, which propagate through time according to the Cox-Ingersoll-Ross processes with time-varying trends:

$$d\lambda_0^\ell(t) = \kappa_\ell(\theta_\ell(t) - \lambda_0^\ell(t))dt + \sigma_\ell\sqrt{\lambda_0^\ell(t)}dz_\ell^{\mathbb{P}} \qquad \ell = d, p \tag{3.1}$$

[1] A rather intuitive discussion of these conditions can be found in Duffie [53]. For a more rigorous exposition, see, for example, Brémaud [15].

[2] For a typical fixed-rate mortgage, this is usually a sequence of payments, though this is irrelevant for the purposes of this discussion.

where the two white noise processes $z_d^{\mathbb{P}}(t)$ and $z_p^{\mathbb{P}}(t)$ are assumed to be independent.[3] The exact forms of the trends will be discussed in Chapter 5. At each payment date, these processes yield the "realized" *discrete hazard rate processes*

$$\lambda_i^\ell = \mathbf{e}^{\beta_\ell' \mathbf{x}(t(i))} \lambda_0^\ell(t(i)) \qquad \ell = d, p \tag{3.2}$$

where $\mathbf{x}(t(i))$ represents the vector of exogenous variables affecting termination (covariates), which are discussed further in Chapter 4, and where $\beta$ is the vector of their coefficients to be estimated.[4] For the purposes of valuation, the most important covariate is the interest rate spread, approximating the difference between the coupon rate on the mortgage and the rate at which the loan can be refinanced. The spread is time-varying, in that it depends on the also stochastic term structure, through the then current yield on a 10-year Treasury bond. Another variable of great interest is the loan-to-value ratio (LTV), which is an indicator of the cost of default.

Since termination in this model occurs only at payment dates, the "intensity" of termination, as usually employed in reduced-form models working in continuous time, is not well defined. From a formal point of view, it is better to think, instead, in terms of the aggregated *martingale hazard process* $\Lambda_\ell^{\mathbb{P}}(t) = \sum_{t(i) \leq t} \lambda_i^\ell$.[5] It is called this since, if one considers a *stopping time* $\tau$ such that $\mathbb{P}_{\{\tau > t\}} = \prod_{t(i) \leq t}(1 - \lambda_i)$ and the *jump process* $H(t) = \mathbb{1}_{\{\tau \leq t\}}$, then $\Lambda(t \wedge \tau)$ is the *compensator* of the jump process, in the sense that $H(t) - \Lambda(t \wedge \tau)$ is then a martingale (i.e., trendless). Schönbucher [123] and Bélanger et al. [8] develop the theory of reduced-form default in terms of hazard processes rather than intensities, while Giesecke [68] has introduced the

[3] Given how small the intensity can be, particularly for default, and that the observed default rate is often zero, it was judged particularly important to choose a stochastic hazard form not permitting negative values.

[4] The term $t(i)$ means the calendar time of the $i$th payment date.

[5] Being piecewise constant, the hazard process has no density (intensity) in the usual sense.

terminological distinction between *intensity-based* and *compensator-based* default modelling.[6]

Assuming the absence of arbitrage, the real probability measure $\mathbb{P}$ driving the model can be transformed into an equivalent martingale measure $\mathbb{Q}$. The Girsanov theorem of Jacod and Shiryaev [79] for random measures is general enough for this purpose and assures that the risk adjustments take the forms

$$\begin{aligned} dz_\ell^{\mathbb{Q}} &= dz_\ell^{\mathbb{P}} - \nu_t^\ell dt && (3.3) \\ d\Lambda_\ell^{\mathbb{Q}} &= \mu_t^\ell d\Lambda_\ell^{\mathbb{P}} \qquad \ell = d, p && (3.4) \end{aligned}$$

where the $\nu_t^\ell$ are the usual additive drift adjustments and the $\mu_t^\ell$ are multiplicative adjustments for jump risk. Note that, in this case, $d\Lambda^{\mathbb{P}}(t)$ signifies $\lambda_i$ on payment date $t(i)$, but zero on nonpayment dates. Following convention, I take $\nu_t$ to be of the form $\nu\sqrt{\lambda_0(t)}/\sigma$, leading to the usual drift adjustment $-\nu\lambda_0(t)dt$ for $d\lambda_0(t)$.[7] Also following convention, I take $\mu$ to simply be constant, so that $\Lambda^{\mathbb{Q}} = \Sigma_i(\mu\lambda_i) = \mu\Sigma_i\lambda_i = \mu\Lambda^{\mathbb{P}}$.[8]

The term structure is taken to be the “extended” [111] or “translated” [56] form of an independent two-factor time-invariant Cox, Ingersoll & Ross (CIR) [38] model, so that in risk-neutralized form

$$dy_\imath(t) = (\kappa_\imath(\theta_\imath - y_\imath(t)) - \nu_\imath y_\imath(t))dt + \sigma_\imath\sqrt{y_\imath(t)}dz_\imath^{\mathbb{Q}} \qquad \imath = 1, 2 \tag{3.5}$$

where the instantaneous interest rate is just the sum of these two factors and a constant, so that $r(t) = y_1(t) + y_2(t) + \bar{y}$. Since the interests of this disseratation are

[6] Also see Giesecke [67], who develops the compensator method to unify the structural and reduced-form approaches, along the lines of Duffie and Lando [54].

[7] As indicated below, one must simultaneously introduce drift adjustments, assumed to be of the same form, on the state variables of the stochastic term structure. These do not, however, require any corresponding jump risk adjustment.

[8] See Bjork et al. [11] and Schönbucher [123], who use the cited Girsanov result, though the former works within an intensity framework.

not primarily in modelling interest rates, I have chosen as straightforward a term structure as possible, in the same affine class as the termination processes.[9]

Now, the time of termination is $\tau = \tau_d \wedge \tau_p$, with the associated martingale hazard process $\Lambda(t) = \Lambda_d(t) + \Lambda_p(t)) = \sum_{t(i)\leq t}(\lambda_i^d + \lambda_i^p)$, in the manner of what, in the literature, is called the "first-to-default" (Duffie [52]) . Taking this elaboration into account, the value of a mortgage $V(t(0))$ may be expressed as

$$V(t(0)) = E^{\mathbb{Q}}_{t(0)}\Bigg[\sum_{i=1}^{I} \mathrm{e}^{-\int_{t(0)}^{t(i)}((1-\tau_F)r(s)+\ell)ds}\Bigg(\prod_{j=1}^{i-1}(1-\lambda_j^d-\lambda_j^p)\Bigg) \\ \Bigg(\lambda_i^d W(i) + \lambda_i^p A(i) + (1-\lambda_i^d-\lambda_i^p)M\Bigg)\Bigg] \quad I = 360 \tag{3.6}$$

where $W(i)$ is the *recovery value* upon default, $A(i)$ is the tax-adjusted unpaid balance due on prepayment, $M$ is the tax-adjusted mortgage payment, indicating that the contract is to be continued, $\ell$ is a liquidity premium, and finally, $\tau_F$ is the Federal tax rate (all further described in Chapter 7).[10]

In the following chapters, I discuss the steps of the estimation procedure in greater detail, presenting results at each step. Emphasis is primarily on estimation of the real termination processes, and secondarily on risk adjustments and valuation, with less interest in estimating the covariates and the term structure.

[9] While the termination processes are time-varying and the term structure processes are not, the former are only varying in the age of the mortgage; none of the processes' functional forms depend on calendar time.

[10] Notice the asymmetry between the interest rate discounting and the risk discounting. One can always, though, take the natural logarithm of the latter and express everything in terms of exponentiated sums and integrals, whereas a complete synthesis of the two can be achieved using the notion of product integration (Andersen et al. [5]). Also note that I use the presumed conditional independence of the various processes (Kijima [95], Bielecki & Rutkowski [10]).

# Chapter 4

# Estimation of Covariate Effects Using a Discrete Stratified Proportional Hazard Model

In this chapter, I concentrate on estimation of effects of exogenous variables (covariates) on default and prepayment hazards. The proportional hazard model is a very convenient tool, since, by employing partial likelihood maximization, it imposes no structure on the underlying baseline hazards[1], while allowing one to estimate covariate effects given by the convenient exponential functional specification.

While partial likelihood is a rather common estimation procedure, several relatively recent developments are included in the model. First, the model employed in this dissertation specifies hazards of termination as discrete processers, which avoids the problem of ties, as discussed later. The model is due to Prentice and Kalbfleisch [116], who show that in this discrete-hazard setting, the conventional Breslow [17] estimator of covariate effects is still consistent, though the estimation of standard errors changes.[2] Second, the stratification device is used, allowing the termination processes to be stochastic, by not imposing any parametric structure on baseline hazards. Third, time-dependent covariates are included, in the form of the

[1] Some authors use this fact to estimate the parameters of a deterministic baseline hazard function at the second stage using linear regression or parametric techniques, e.g. Ridder and Tunali [121], Fu et al. [65]. Unlike those models, mine uses a stochastic termination process, the baseline of each individual stratum representing a random trajectory of the same stochastic process.

[2] Prentice and Kalbfleisch also show that the asymptotic normality of the estimators continues to hold. Since the model employed in this paper involves stratification, some further adjustments to the covariance matrix are necessary.

time-varying interest rate spread.[3] Fourth, left truncation, as well as right censoring and uninformative censoring due to loans sold, are present in the model, though not affecting the consistency of estimates. Finally, default and prepayment are modelled as competing risks, with the conditional independence assumption allowing one to estimate the two models separately.

The following sections of this chapter are structured as follows. In the first section, I describe the empirical model. The second section contains a discussion of data and empirical results.

## 4.1 The Mixed Discrete and Continuous Hazard Model

### The Basic Model

Assume the complete probability space described in Chapter 3. Let $N_i(t)$ be a right-continuous (*càdlag*) counting process that takes value zero at $t = 0$ and jumps by one at an observed failure time for an individual $i$, and denote by $\mathcal{N}$ the associated filtration $\mathcal{N}_t = \{N_i(s) : s \leq t, i = 1, ..., n\}$.

Assume that the covariate process $\mathbf{x}_i(t)$ and the censoring process $Y_i(t)$ , where $Y_i(t)$ takes value of one if an individual $i$ is at risk at time $t^-$, and zero otherwise,[4] also generate a certain filtration $\mathcal{Y}$, with $\mathcal{Y}_t = \{Y_i(s), \mathbf{x}_i(s), s \leq t : i = 1, ..., n\}$. Finally, define $\mathcal{G} = \mathcal{N} \vee \mathcal{Y}$. For each $t$, the $\sigma$-field $\mathcal{G}_t$ is assumed to represent all information available at that time.

[3] Strictly speaking, there is no such thing as a proportional hazard model when time-varying covariates are present. Time-dependent covariates will change at different rates for different individuals, hence introducing non-proportionality into the model. However, I will follow common terminology and refer to this model as a proportional hazard model with time-dependent (time-varying) covariates.

[4] Censoring is assumed to be an independent process. For details, see Andersen at al. [5].

Following Prentice and Kalbfleisch [116], I specify the multiplicative intensity model[5] by defining the baseline hazard as

$$\Lambda_i(dt) = \Lambda_0(dt)\mathbf{e}^{\beta'\mathbf{x}_i(t)} \tag{4.1}$$

where $\Lambda_0(dt) = \Lambda_0(t) - \Lambda_0(t^-)$ if $t$ is a mass point of the failure distribution (*discrete case*), while $\Lambda_0(dt) = \{d\Lambda_0(t)/dt\}dt$ at a continuity point of the failure distribution (*continuous case*), with the cumulative intensity process defined as:

$$\Lambda_i(t) = \int_0^t Y_i(s)\mathbf{e}^{\beta'\mathbf{x}_i(s)}\Lambda_0(ds) \tag{4.2}$$

For estimation the the parameter vector $\beta$, consider the Doob-Meyer decomposition of the counting process $N_i(t)$

$$N_i(t) = \Lambda_i(t) + M_i(t), \quad i = 1, ..., n \tag{4.3}$$

where $\Lambda_i(t)$ is the *compensator* of the counting process $N_i(t)$ with respect to the filtration $\mathcal{G}_t$, and $M_i(t)$ is the *counting process martingale* corresponding to $N_i(t)$. When sample paths of both $\mathbf{x}_i$ and $Y_i$ are *càdlag* processes, the martingale $M_i = N_i - \Lambda_i$ is square integrable with respect to the filtration $\mathcal{G}_t$. As Prentice and Kalbfleisch [116] demonstrate, an estimating function for $\beta$ (which reduces to the partial likelihood score process if failure time is absolutely continuous) is expressed as

$$U(\beta, t) = \int_0^t \sum_{i=1}^{n} \left\{ Y_i(s)\mathbf{x}_i(s) - \sum_{l=1}^{n} \mathbf{x}_l(s)p_l(s) \right\} N_i(ds) \tag{4.4}$$

where:

[5] The original multiplicative intensity model was developed by Aalen [2], based on the proportional hazard model of Cox [36].

$$p_l(s) = \frac{Y_l(s)\mathbf{e}^{\beta'\mathbf{x}_l(s)}}{\sum_{j=1}^{N} Y_j(s)\mathbf{e}^{\beta'\mathbf{x}_l(s)}} \tag{4.5}$$

Substitution of (4.5) into (4.4) yields

$$U(\beta, t) = \int_0^t \sum_{i=1}^{n} \left\{ Y_i(s)\mathbf{x}_i(s) - \sum_{l=1}^{n} \mathbf{x}_l(s) p_l(s) \right\} M_i(ds) \tag{4.6}$$

which is a stochastic integral of a predictable process with respect to a square integrable martingale. Hence, $U$ itself is a square integrable martingale with respect to $\mathcal{G}_t$, and can be employed to obtain consistent and asymptotically normal estimates of $\hat{\beta}$.

As Prentice and Kalbfleisch also point out, the overall score $U(\beta, \infty)$ is equal to $\partial \ln L(\beta)/\partial\beta$, where $L(\beta)$ is the Breslow [17] approximate partial likelihood estimator[6]

$$L(\beta) = \prod_{i=1}^{n} \frac{\mathbf{e}^{\beta'\mathbf{s_i}(t)}}{\left[\sum_{l \in R_i(t)} \mathbf{e}^{\beta'\mathbf{x}_l(t)}\right]^{d_i}} \tag{4.7}$$

and where, for each $i$, $R(t)$ includes all individual durations that are at risk at time $t$ (the risk set), $D(t)$ includes all individual durations that terminate at time $t$, $d$ is the number of durations in $D(t)$, and $\mathbf{s}(t) = \sum_{i \in D(t)} \mathbf{x}_i(t)$.[7]

The presence of competing risks can be incorporated in the model in the usual manner. Denote by $\mathcal{G}^k$, the filtration produced by the counting process, the covariate process and the censoring process of the $k$-th risk. Then, $\mathcal{G} = \mathcal{G}^1 \vee \mathcal{G}^2 \vee ... \vee \mathcal{G}^K$. For

[6] While parameter estimates obtained using the Breslow method are consistent and asymptotically normal, the same cannot be said about estimates of the covariance matrix. Prentice and Kalbfleisch develop a consistent covariance estimate in the martingale framework.

[7] Given $\hat{\beta}$, a consistent estimate of $\Lambda_0(t)$ can be obtained using the Nelson-Aalen estimator

$$\hat{\Lambda}(\beta, t) = \int_0^t \frac{\sum_{i=1}^{n} N_i(ds)}{\sum_{l=1}^{n} Y_l(s)\mathbf{e}^{\hat{\beta}'\mathbf{x}_l(s)}} \tag{4.8}$$

where ratios 0/0 are defined to have the value of 0.

estimation purposes, I am assuming that, conditional on $\mathcal{G}_t$, the risks are independent, though they may be correlated through the corresponding covariate processes.

### Introducing Baseline Stochasticity via Stratification

In the model described by (4.2), all variation in $\Lambda_i(t)$ is explained by the covariate process and the censoring process. The baseline hazard is deterministic, i.e. given the appropriate covariate values, one can always predict the conditional probability of termination.

This feature of the estimation technique is not consistent with the *doubly stochastic* character of the pricing model. By double stochasticity, I mean that not only is it uncertain whether an individual duration will terminate at a particular time, but that the probability with which this occurs is also uncertain beforehand.[8] Thus, not only the actual terminations, but the baseline hazards of these terminations, are both modelled as stochastic processes, propagating through time.

A potential solution to the dilemma of how to treat termination probabilities as a random process, while still employing a proportional hazard framework in estimation, is to stratify the data by time of origination; that is, to assume that all mortgages in the same quarterly stratum share the same baseline hazard function, but that a different baseline hazard is drawn quarterly from the common process.[9] While the market is assumed to know this process, there is no reason for it to know any actual draw, except in retrospect. While admittedly ad hoc, this convenient device is entirely consistent with the doubly stochastic reduced-form framework, which in

[8] A narrower meaning of doubly stochastic is that it be a Cox process, an extension of the classical Poisson process to incorporate a stochastic intensity (Duffie [53]). My process in not doubly stochastic in this narrower sense, since termination occurs only at payment dates.

[9] A thorough discussion of stratified partial likelihood can be found in Ridder and Tunali [121].

its essence, addresses only a single mortgage and not the relationship among mortgages. Note that this device takes good advantage of the large data set, since there are still enough mortgages within each strata that the rare event of default can be adequately observed. While, in the reduced-form spirit, one can remain agnostic as to the precise interpretation of each termination probability's stochastic component, they may be thought to involve changes in general economic conditions, beyond those represented by the term structure, which are affecting the probability of that termination, and which are then indeed shared by all mortgages of a given cohort.[10]

If data are divided into $J$ strata, with $n_j$ observations in each one, the Breslow estimator can be easily adapted to take the form of:

$$L(\beta) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} \frac{\mathbf{e}^{\beta' \mathbf{s_{ij}}(t)}}{\left[\sum_{l \in R_{ij}(t)} \mathbf{e}^{\beta' \mathbf{x}_{lj}(t)}\right]^{d_{ij}}} \tag{4.9}$$

Introducing competing risks into the stratified model does not produce any major complications. As in the unstratified case, the two equations can be used separately, since all correlation is assumed to be captured either by covariates, or by the unobservable stratum-specific effects.

### Testing for Stratum-Specific Baselines

The test for the presence of stratum-specific effects is performed with the clustering test statistic (Ridder and Tunali [121]), which uses the information from both the Stratified Partial Likelihood Estimation (SPLE) and the Unstratified Partial Likelihood Estimation (UPLE). The null hypothesis is that there are no strata-specific

[10] Under the interpretation of the stochastic component as representing external conditions, the time-varying trend of the stochastic hazard rate, as described below, will then account for patterns of termination entirely internal to a mortgage, which should then be mostly governed by the maturity of the loan, and so, similar among all mortgages. Note that both external and internal forces are allowed to interact in forming the actual hazard rate.

effects present, i.e. that $\Lambda_{0j}(dt) = \Lambda_0(dt)$, $j = 1, ..., J$. The test statistic then takes the form

$$C = (\hat{\beta}_s - \hat{\beta}_u)' V(\hat{\beta}_s - \hat{\beta}_u)^{-1}, (\hat{\beta}_s - \hat{\beta}_u) \tag{4.10}$$

where $\hat{\beta}_s$ denotes the vector of coefficient estimates from the stratified model, and $\hat{\beta}_u$ is the vector from the unstratified one.

Under the regularity conditions found in Andersen and Gill [4], the difference $(\hat{\beta}_s - \hat{\beta}_u)$ is asymptotically normal with the variance given by

$$V(\hat{\beta}_s - \hat{\beta}_u) = V(\hat{\beta}_s) - V(\hat{\beta}_u) \tag{4.11}$$

$V(\hat{\beta}_s)$ and $V(\hat{\beta}_u)$ being inverted Hessians for the respective log-likelihoods evaluated at their maxima.[11]

The variance matrix has the same form as the Hausman test, and, as Ridder and Tunali argue, for large $J$, the test statistic $C$ has a chi-square distribution with $p$ degrees of freedom, $p$ being the number of covariates in the model.

## 4.2 Data and Results

### 4.2.1 Data Structure and Sources

The original data set is obtained from the Bank of America, and contains information about single-family fixed-rate residential mortgages originated during the period from December 1967 to February 2002. The original number of loans in the database is 1,972,116. The variables available in the data set include the day of origination, the

[11] In a discrete-time setting, since inverted Hessians are not consistent estimates of the covariance of the coefficients, one should use the covariance estimates developed by Prentice and Kalbfleisch [116].

day of the first observation, the *month* of termination (if terminated),[12] the reason for termination—defaulted, prepaid, sold, or censored—and a number of loan-specific characteristics observable at the time of origination. The characteristics available are the original dollar loan amount, the original loan-to-value ratio, the contract rate at origination, and the amount of points paid at origination.

There are obvious recording errors with the data prior to 1970, and therefore 1970 is chosen to be the first year of the sample used for estimation. The year 2001 is chosen to be the last year, first, because of a negligibly small number of defaults among the loans originated in 2002, and second, because the interest rate series that are used to construct time-varying covariates are not available for 2002. This makes the time horizon of the data set equal to thirty-two years, two years beyond the length of the thirty-year mortgages under consideration. Left truncation, which is a rather common problem in mortgage loan data, is also present in this data set. The loans originated in the seventies and the eighties are not observed until at least 1990, or, in some cases, until 1993. Fortunately, the problem is relatively easily solved by the proportional hazard procedure employed for estimation.

After all observations with missing or apparently erroneous values were deleted, the sample was restricted to 30-year conforming mortgages. The remaining dataset contains 917,703 observations.[13]

For the purposes of estimation, the data have to be converted into a counting style format, with each observation representing a month during which a loan is under observation. This effectively increases the total number of loan-period observations

[12] This feature of the loan data (as well as risk-free interest rate data), along with the theoretical considerations described in Chapter 2, motivates the use of monthly frequency throughout the dissertation.

[13] The possibility of selection bias relative to the entire stock of U.S. mortgages certainly also exists, since there is peripheral evidence that the lending institution has sought to retain only "higher-quality" loans, which would presumably result in lower than usual default rates.

to 25,780,497,[14] which makes any data manipulations and the estimation itself rather cumbersome and time-consuming.

The interest rate data for US Treasury Bonds are obtained from Robert Bliss. The data are discussed in more detail in Chapter 6.

The housing price data used for adjusting the original loan balance are obtained using Freddie Mac's Conventional Mortgage Home Price Index (CMHPI), available at www.freddiemac.com/finance/cmhpi. Since the price data are reported quarterly, and I need monthly data for discounting loan balances, I interpolate the price index data using a cubic spline method.

### 4.2.2 Choice of Variables and Specification

The decisions about the choice of variables included in estimation, as well as their functional form, are based on the following two general observations.

First, there exists a large body of literature (as described in Section 2 of this chapter) that has identified several major determinants of mortgage default and prepayment. Starting from early nineties, option-based theories of mortgage termination have been used to describe how the value of the call option, as determined by the spread between the market value of a loan and its book value, and the value of the put option, as determined by the contemporaneous loan-to-value ratio, are the most important factors affecting borrowers' behavior. In later work, transaction costs of default and prepayment, as well as general borrower heterogeneity and a variety of macroeconomics factors, were shown to be important factors affecting the mortgage termination.

On the other hand, since the results on the proportional hazard estimation will later be used in a pricing model, it is crucial to restrict the set of covariates to

[14] The procedure employed for estimation allows me to drop observations for periods during which the older left-truncated loans are not observed, e.g. 1970-1990, reducing the number of observations required for estimation from 39,098,353 to a mere 25,780,497.

those directly observable in the mortgage data. Other variables, especially those related to the general state of the economy, should not be included in the model, unless one is later going explicitly model these variables as stochastic processes describing the state of the world. This is exactly the same problem as the one observed by Schwartz and Torous [124] in their mortgage pricing model. As a result, the only time-varying and not borrower-specific or loan-specific variable included in the proportional hazard model is the long-term interest rate, which is relatively easy to include, as the earlier-mentioned CIR model of the term structure.

To maintain sufficient flexibility of the model for pricing purposes, while maintaining relative parsimony, I use quadratic terms for all covariates, along with the usual linear ones. The test for the joint significance rejects the null hypothesis of a linear model, as reported in Panel B of Table 4.4. For simplicity of interpretation, all covariates are measured in percentage terms.

To preserve consistency between the estimated model and the theoretical pricing model, the data are stratified by the quarter of origination, each quarterly stratum viewed as a draw from the same underlying stochastic process. This produces 128 strata. The quarterly stratification is then tested against its alternatives, i.e. the models with no stratification, annual stratification, and a semiannual one,[15] using the clustering C-statistic given in 4.10. The results of the tests provide strong support for quarterly stratification, as reported in Panel C of Table 4.4.

The remainder of this section provides a short discussion of covariate specification. Since the present dissertation is not explicitly based on an option-theoretical pricing model, but rather on a more empirical reduced-form one, the choice of covariates is also more intuitive than theoretical; however, it is not inconsistent with previous work.

[15] It has proven to be virtually impossible to obtain any meaningful results with stratification of higher than quarterly frequency.

Some measure of the difference between the current contract rate and the rate at which a loan can be refinanced, i.e. the *interest rate spread*, is almost always included in models of mortgage termination. As the market interest rate decreases, it makes the loan relatively more expensive, and so gives the borrower an incentive to prepay the mortgage and refinance (i.e. exercise the call option). Exactly the opposite is true for an increase in the market rate. The risk of default is also expected to be affected by the spread; however, it is not always clear in which direction. One argument is that lower interest rates make it more likely that a borrower can default on a loan, move to another residence, and obtain a new loan with a lower rate. For this to be a rational decision, the transaction costs of default have to be low, and the contemporaneous LTV ratio has to be relatively high. Another argument working in the same direction is that if the current rate goes up, a borrower has an incentive to either keep the current loan with a below-market rate [134], or, if the loan is assumable, sell the house and include the current financing in the sale price of the house [23], making default less likely. On the other hand, a lower current interest rate makes the future option to prepay more valuable, potentially leading to a higher probability of default in any given month. Which of these two effects dominates the other remains to be tested empirically. [16]

The functional form chosen for the interest rate spread varies from one study to another. Some authors calculate the present value of remaining mortgage payments

[16] This is an example of what I will further refer to as the direct effect versus the indirect effect. While the direct effect is almost always rather straightforward—e.g. a decrease in the interest rate will cause an increase in the risk of default since the present value of remaining mortgage payments becomes higher—the indirect effect is more subtle. The indirect effect is based on the fact that when considering whether to default or prepay a loan, a rational borrower considers future possibilities of default and prepayment. By making either the default or the prepayment decision, the borrower gives up a valuable right to make the other decision in the future (see Kau and Keenan [86] for a discussion). As a result of the indirect effect, a decrease in the interest rate can be expected to cause an indirect decrease in the risk of default, since the value of future prepayment becomes higher.

at the contract rate and at the current interest rate, representing the book value and the market value of a loan, respectively, and then compute either the normalized difference or the ratio of the two (e.g. Follain et al. [64], Ciochetti et al. [31], Deng et al.[45], among others). Others simply specify a functional dependence between the two interest rates, such as the ratio (Richard and Roll [120], Pavlov [110]), or the difference (Schwartz and Torous [124]). Often, the difference is normalized either by the contract rate (Calhoun and Deng [21], Cunningham and Capone [39]), or by the current rate (Campbell and Dietrich [23], Ambrose and Sanders [3]). Closed-form formulae for the mortgage rate, like the one proposed by Collin-Dufresne and Harding [33], have also been used (Clapp et al.[32]). A variety of additional non-linear functional forms for the spread have also been employed in the mortgage literature. Some examples include a cubed term (Schwartz and Torous [124]), a spline function with one (Huang and Ondrich [77]) or several (Follain et al. [64]) knots, as well as dummy variables (Calhoun and Deng [21]). Recently, a quadratic term became widely used (Deng et al. [45], Ciochetti et al. [31][30]).

For reasons described in the beginning of this section, using either a closed-form formula or a discounted cash flow model does not seem to be appropriate for my purposes. Therefore, I have chosen to use a function of the two rates, specified as the contract rate $r_c$ less the current rate $r_t$ over the contract rate, i.e. $(r_c - r_t)/r_c \times 100$, to represent the relative interest rate spread.[17] Following Schwartz and Torous [124], I use a *lagged* value of the average yield on 10-year US Treasury Bonds as a proxy for the long-term interest rate. The optimal length of the lag is two months, as

[17] The upper boundary of the spread variable is fixed at 100%, since interest rates are always positive. The lower boundary is usually greater that zero, zero value of the spread meaning that the contract rate is equal to the Treasury bill yield, which is not very likely, given the risk premium built into the mortgage rate and the general downward-sloping trend in interest rates during the observed period. In fact, the spread is negative for only about 1% of loan-period specific observations in our sample.

indicated by the Akaike Information Criterion and the Schwartz Bayesian Criterion. The results of the tests are reported in Panel A of Table 4.4.

To test for the additional effect of the contract rate on the risk of default, I include the *contract rate* as a covariate. A priori, I expect the effect to be positive for both risks of termination.

Another major variable of interest is the *loan-to-value ratio* of a mortgage. Earlier arguments in the mortgage-related literature appealed to a high LTV as an indicator of the borrower's having potential cash flow problems [134][75]. Later, option-pricing theory was employed to conclude that higher contemporaneous LTV should increase the risk of default, due to a high possibility of negative equity, indicating that the put option is likely to go in-the-money (see Vandell et al. [133], Clapp et al. [32], Ciochetti et al. [31], among others). On the other hand, high loan-to-value ratio at origination could have a negative effect on the risk of default, due to adverse selection at origination, e.g. higher underwriting standards for loans with high loan-to-ratio, or, equivalently, high down payment requirements for risky loans [23][3]. To distinguish the two effects, contemporaneous loan-to-value ratio of a loan has to be observed together with the LTV ratio at origination. Since I have no information about the evolution of housing prices, and I am unwilling to make any assumptions about their dynamic behavior,[18] I use LTV at origination as a proxy for contemporaneous LTV. Only the dominant of the two effects will therefore be detected in estimation. For prepayment, higher LTV also has at least two effects working; however, they are in the same direction: the direct one making refinancing more costly and generally less attractive, the indirect one increasing the value of future default. Both effects are expected to decrease the probability of prepayment.

[18] See Deng [43], Deng et al. [45] for examples and a discussion of typical assumptions and data necessary to construct contemporaneous LTV series. In the pricing model, having a contemporaneous LTV covariate is equivalent to having another state variable.

The *original loan balance* is also included as a covariate. Prior specifications present in the literature include the dollar amount (e.g Clapp et al. [32]) or a set of dummies, such as small/medium/large (e.g. Ciochetti et al. [31][30], Deng and Calhoun [21]). I choose another popular specification, the natural logarithm of the *adjusted*[19] original balance [77], since it allows easy interpretation of the estimated coefficient, and also avoids the arbitrary assumptions required to construct dummy variables. As for the expected sign of the coefficient, it is often argued that larger loans give a larger incentive to prepay and refinance, due to a greater dollar benefit of refinancing [32] and lower refinancing costs caused by a fixed component in the costs of prepayment [77], but it is not clear what the effect on default should be. It has been argued that it should be positive, due to greater reward or lower costs of default [32][77]. However, if the loan size is used as a proxy for borrower's other wealth or total income, we should expect wealthier borrowers to be less likely to default, which would lead to a negative sign of the coefficient estimate. The empirical evidence on this issue has been mixed, mostly producing an insignificant estimate.

One major advantage of the data at hand is that they include information about *points paid at origination*, as a percentage of the original loan size. Such information is rarely available in mortgage data, and often has to be extracted from mortgage contract rate data and other loan characteristics [110], leading to unavoidable errors

[19] The original amount of each loan is adjusted to match the general house price level of December 2001. The main reason for detrending original loan size is so that we can make consistency claims for our estimation. One could go on to make a good argument that, in predicting a borrower's behavior, a loan should continue to be adjusted through time by the then current index, making loan size a time-varying covariate. Indeed, one could argue that the loan's size in the LTV ratio should be similarly deflated through time, in which case it is particularly clear that one wants to be using a house price index, rather than, say, the CPI. The main reason I do not do any of this is that, when valuing a mortgage, I do not want to be faced with the problem of projecting the future course of an index. On a similar note, one could argue that "loan" should be everywhere replaced by "unpaid principal." However, the relation between the two is entirely predetermined, and common to all mortgages, so that it can be considered as being absorbed into the trend of the baseline.

in the process of estimation. Loans with low origination points are expected to be prepaid more quickly, since for a given contract rate, higher points lead to higher effective rate at the time of prepayment, and the borrowers who are expecting to prepay the loan choose their point-rate combinations accordingly [128]. For default, the situation is slightly more complex, since, on the one hand, there is a direct effect of higher points meaning higher effective costs of default, thus leading to a negative expected coefficient; while, on the other hand, it is also possible that there is an indirect positive effect caused by the negative relation between points and prepayment.

Apparently, not all covariates described in this section are expected to be of equal importance for default and prepayment.[20] As a result, some researchers choose to make the sets of covariates different across risks,[21] with some covariates being present in one of the models, but not in the other. Rather then making such *a priori* assumptions, I estimate both models with the full set of covariates, which enables me to test assumptions about their importance, instead of imposing them.

### 4.2.3 Descriptive Statistics

Descriptive statistics for the entire data set are reported in Panel A of Table 4.1. The average original balance of a loan in the sample is $120,638, which translates into about a $111 billion value for the entire sample. The smallest original loan balance is only $1,400 and the biggest one is $528,700. After I adjust the original balances to match the average housing prices of December 2001, the average loan balance increases to $150,689, indicating the general level of inflation in the housing market

[20] It can be argued, for example, that the interest rate spread, the loan size, and the points are more important for the risk of prepayment, while the loan-to-value ratio and the contract rates are more important in determining the risk of default.

[21] Pavlov [110], for example, includes only six covariates in the default model estimation, while the prepayment model is estimated using the entire set of ten covariates. Most likely, data problems were the reason for this choice of specification.

over time. The average loan-to-value ratio in the sample is 78.5%, and its range is from 0.7% to 100%.[22] The average contract rate at origination is 7.62%, with a minimum of 0.74% and a maximum of 17.75%. Points paid at origination have an average value of 0.02%, and range from -19%, which means that the lender pays the borrower 19% of the original loan amount at the time of loan origination, to 24%.

Out of 917,703 loans in the sample, 345,997 (37.7%) were prepaid during the period under observation, and 4,564 (0.5%) defaulted during that period. Descriptive statistics for defaulted and prepaid loans are given in Panels A & B of Table 4.1, respectively. All variables seem to be roughly the same across the categories, with the exception of the original LTV ratio. Defaulted loans apparently have higher average LTV when compared to the total sample or to loans that have been prepaid.[23]

Panel A of Table 4.2 displays the number of loans originated each year, along with the percentage of each cohort that defaulted or were prepaid during the period of observation. The total number of mortgages originated ranges from 240 in 1970 to 208,327 in 1999. The average percentage of mortgages prepaid is approximately 60% across years of origination, and is relatively stable, with an obvious tendency to decrease for less seasoned loans. Mortgages originated in the early eighties seem to have slightly lower levels of prepayment. The loans originated in the later years of the sample tend to have overall lower levels of prepayment, since they have been active and observed for only a short period. Percentage defaulted shows a different pattern over time. It increases up to the early nineties, which is most likely due to left truncation of earlier mortgages, i.e. to the fact that some mortgages of the same cohort were prepaid or defaulted prior to the beginning of the observation period.

[22] Several loans with LTV over 100% were deleted from the sample when it was restricted to include only conforming loans.

[23] The contract rate also seems to be slightly higher; however, the spread between the contract rate and the risk-free bond rate at the time of origination is roughly the same.

By 1990, the unconditional probability of default reaches the maximum of almost 5%, and falls rapidly in the following years.

Panels B & C of the table represent loans observed, defaulted and prepaid, by the year of termination. The total number of observed loans increases from 33 in 1970 to 736,982 in 2001. Prepayment usually constitutes the major reason for termination, accounting for at least 80% of the total number of loans terminated in any particular year, with the exception being 2001, when the observation period stopped, i.e. the data were right-censored.

Table 4.3 presents Pearson correlation coefficients for the variables included in the estimated model. One can easily observe the high degree of correlation between linear and quadratic terms, particularly for the LTV ratio, the original loan balance, and the contract rate, i.e. those variables restricted to always be positive. The close correlation is likely to lead to low statistical significance of the estimated coefficients.

Figure 4.1 displays the CMHPI price index used for constructing the adjusted loan balances, normalized to make the price level in December 2001 equal to one. The index clearly displays the upward trend in house prices during the observation period, indicating that overall prices increased by a factor of more than five. Different interest rates related to the data set are displayed in Figure 4.2: the conventional mortgage rate reported by the Federal Reserve Bank of St. Louis, the 10-year Treasury Bond Rate, and the sample average of the contract rate by the month of origination.

For illustrative purposes, I also provide non-parametric Kaplan-Meier [85] estimates of both default and prepayment hazard in each month of mortgage life (Figures 4.3 & 4.4). The hazard of default seems to have a tendency to increase monotonically over the first ten years of mortgage life, and decrease afterwards. For the hazard of prepayment, is is hard to say whether there is any trend, other than a tendency to increase in the very last months of mortgage life, when the remaining balance becomes increasingly small.

### 4.2.4 Empirical Results

This subsection presents empirical results of estimating a model with two conditionally independent risks of termination, employing the estimation function of Prentice and Kalbfleisch [116], which is equivalent to maximization of the likelihood function using the conventional Breslow [17] estimator.[24] The model uses a quadratic specification of the covariates (results of the test for the joint significance are presented in Panel B and Table 4.4) and quarterly stratification (results of the test of quarterly stratification against its alternatives are given in Panel A of Table 4.4). The covariates included in the model are *LTV, Contract Rate, Log Loan Size, Points* and *Spread*, as well as their squared terms.

The results of estimation for both default and prepayment models are presented in Table 4.4. The results include the estimated values of coefficients, the standard error obtained by inverting the estimated Hessian at the maximum of the log-likelihood function, and the corrected standard error, developed along the lines of the Prentice and Kalbfleisch covariance estimator [116]. However, the difference between the conventional estimates of the standard error and the corrected ones is negligibly small for almost all covariates, possibly due to the large size of the data set I am using.

Almost all coefficients in the models are statistically significant, with the exception of *LTV* and *LTV Squared* in the prepayment model, and *Contract Rate Squared*, *Points Squared* and *Spread* in the default model, though their joint significance is very high, as indicated in Panel B of Table 4.4. Apparently, this lack of statistical significance is explained for the most part by a high degree of correlation between linear and quadratic terms in the model.

[24] There are, in fact, two likelihood functions, maximized separately, based on the independence assumption. The total log-likelihood is just the sum of the two.

Since the total effect of each covariate is the sum of the linear and the quadratic effects, the total effect is not always clearly observed by simply looking at the individual coefficients. For this purpose, in Figure 4.5 I also provide graphs of the joint effect of each linear and squared term together, $\mathbf{e}^{\beta_k^{linear}\mathbf{x}_k+\beta_k^{quadratic}\mathbf{x}_k}$ for each of the ten covariates. The expression has the value of one if the corresponding covariate is set to zero. By looking at the graphs, one can easily observe that, for some variables, the total effect is non-monotonic, since coefficient estimates for the linear and the quadratic term do not have the same sign. This non-monotonicity is particulary obvious for the effect of the loan size on the hazard of default and the effect of the contract rate on the hazard of prepayment. The rest of the section discusses each pair of coefficient estimates in more detail.

The original mortgage loan-to-value ratio is found to have a positive and non-linear effect on the probability of default, which supports the hypothesis that higher original LTV serves as an indicator of possible future negative equity. Interestingly, a linear effect is significant only at 10%, but the quadratic term is very significant, which is due to close correlation between the two terms, and consistent with previous research that used a similar functional specification (e.g. Ciochetti et al. [31]). The adverse selection effect, even though it may be present in the data, is dominated by the negative equity effect. The effect of LTV of the hazard of default is given in Panel A of Figure 4.5. It demonstrates that, ceteris paribus, a loan with the original LTV of 100% has the hazard of default 7 times that on a loan with the original LTV of 75%. The effect of original LTV of the hazard of prepayment is statistically insignificant.

The contract rate on a loan is found to have a significant positive effect on the hazard of default (even though the squared term is not significant). Panel B of Figure 4.5 shows that the hazard of default for a loan with the contract rate of 10% is 8.4 times higher than that of a loan with contract rate of 7%. The contract rate also has a significant, though not monotonic, effect on the hazard of prepayment. This effect is also much smaller in magnitude than for default: a loan with a 10% rate has about the same hazard of prepayment as a similar loan with the contract rate of 7%.

The adjusted loan balance at origination has a positive and significant effect on the risk of prepayment: a loan with the original balance of $380,000 has the hazard of prepayment about 1.5 times that of a loan with the original balance of $140,000. The sign of the coefficient estimate is identical to the results obtained by Clapp [32], Huang and Ondrich [77], Pavlov [110], among others, and supports the idea that borrowers who face a greater reward for prepayment are more likely to prepay the loan. For default, the effect is non-monotonic. Very small loans tend to have the highest tendency to default, possibly, because borrowers taking smaller loans (under $100,000), i.e. having a less expensive collateral, are potentially more income-constrained, and more likely to default on the mortgage payments. For loans with the original balance over $100,000, the relationship between the original loan balance and default is nearly non-existent.

The coefficient estimate for points paid at origination is negative and significant in both models (the squared term is insignificant for default). A borrower who has paid 4 points at origination is only 80% as likely to default and 60% as likely to prepay the loan as a borrower who has paid no points at all.

The relative spread between the contract rate and the contemporaneous interest rate is found to be positive and highly significant for both default and prepayment. The hazard of prepayment is 14.3 times higher when the spread is 50% of the contract rate than when it is zero. This result is consistent with the option pricing theory of prepayment, and is strongly supported by prior research, including papers that used different functional forms for the spread specification. For default, the effect of relative spread is similar to the one for prepayment, yet smaller in magnitude. The hazard of default increases by a mere 60% when the spread increases from zero to 50%. This result supports the original assumption about the negative relationship between current interest rates on the probability of default. Prior empirical research also indicates some evidence of such an effect, at least in linear specifications (see Campbell and Dietrich [23], Vandell and Thibodeau [134], Cunningham and Capone [39], among others).

Overall, the estimation produces plausible coefficient estimates for all covariates included in the model, with the minor exception of several statistically insignificant effects. Also, direct effects seem to dominate the indirect substitution effects in the majority of cases.

Figures 4.7 and 4.9 display predicted hazard rates aggregated over 128 strata, with covariates set at the sample average level. Although these rates are not used in the following computations of hazard processes, they should give a relatively good idea about the general pattern of loan termination. The cumulative default and prepayment levels, as displayed as Figures 4.8 and 4.10, indicate that, for an average loan, the cumulative hazard of default over the entire 30-year period is 1.5%, and the hazard of prepayment over the same period is nearly 100%, i.e. almost all loans in the sample are prepaid before their maturity. Figures 4.11 and 4.12 display cumulative termination rates for different values of major covariates, with all other variables set to the mean. For the hazard of default, the major determinant is the original LTV ratio. The results indicate that, for a loan with the original LTV of 95%, the cumulative hazard of default can reach 4.2%, while it is only 2.8% for a loan with 90% LTV, 1.3% for a loan with 80% LTV, and 0.7% for a loan with 70% LTV. The hazard of prepayment, on the other hand, is heavily dependent on the interest rate spread, i.e. on the relative difference between the contract rate and the current 10-year Treasury Bond rate. The results indicate that, even though almost all loans are predicted to be prepaid by their maturity, the speed of prepayment is different. Given the average contract rate of 7.6%, 48% of loans will be prepaid by the end of year 15 if the Treasury rate remains at 7%, 73% will be prepaid by the same time if the rate remains at 6%, 93% will be prepaid if the rate remains at 5%, and nearly 100% will be prepaid if the rate remains at 4%.

**TABLE 4.1**

**Panel A**

**Summary Statistics: All Loans**

| Variable | Mean | Median | Std.Dev. | Min. | Max. |
|---|---:|---:|---:|---:|---:|
| *Original Loan Size ($)* | 120,638.47 | 112,000.00 | 57,317.80 | 1,400.00 | 528,700.00 |
| *Adjusted Original Loan Size ($)* | 150,688.89 | 141,207.93 | 65,330.39 | 1,711.72 | 878,886.08 |
| *Original LTV (%)* | 78.50 | 80.00 | 17.15 | 0.07 | 100.00 |
| *Contract Rate (%)* | 7.62 | 7.50 | 0.80 | 0.74 | 17.75 |
| *Bond Rate at Origination (%)* | 6.06 | 5.92 | 0.89 | 4.67 | 14.92 |
| *Points (%)* | 0.02 | 0.00 | 0.80 | -19.00 | 24.00 |

**Panel B**

**Summary Statistics: Defaulted Loans**

| Variable | Mean | Median | Std.Dev. | Min. | Max. |
|---|---:|---:|---:|---:|---:|
| *Original Loan Size ($)* | 96,169.93 | 86,750.00 | 47,293.58 | 11,210.00 | 306,400.00 |
| *Adjusted Original Loan Size ($)* | 143,786.89 | 128,966.10 | 68,775.42 | 17,190.07 | 469,986.16 |
| *Original LTV (%)* | 89.12 | 93.65 | 11.04 | 1.00 | 100.00 |
| *Contract Rate (%)* | 8.46 | 8.38 | 1.14 | 3.75 | 15.50 |
| *Bond Rate at Origination (%)* | 7.00 | 6.95 | 1.14 | 4.70 | 13.78 |
| *Points (%)* | -0.03 | 0.00 | 0.73 | -11.50 | 9.00 |

**Panel C**

**Summary Statistics: Prepaid Loans**

| Variable | Mean | Median | Std.Dev. | Min. | Max. |
|---|---:|---:|---:|---:|---:|
| *Original Loan Size ($)* | 118,285.25 | 112,000.00 | 55,196.00 | 1,400.00 | 528,700.00 |
| *Adjusted Original Loan Size ($)* | 159,152.39 | 151,517.93 | 64,685.40 | 4529.04 | 878,886.08 |
| *Original LTV (%)* | 78.22 | 80.00 | 16.67 | 0.78 | 100.00 |
| *Contract Rate (%)* | 7.98 | 7.88 | 0.79 | 0.75 | 17.75 |
| *Bond Rate at Origination (%)* | 6.42 | 6.22 | 0.93 | 4.67 | 14.92 |
| *Points (%)* | 0.00 | 0.00 | 0.70 | -19.00 | 5.63 |

Descriptive statistics presented in Panel A are constructed using the entire sample of 917,703 loans. Descriptive statistics presented in Panels B and C are constructed using the sub-samples of 4,564 loans that defaulted during the period of 1991-2001 and 345,997 loans that were prepaid during the same period, respectively. All variables were obtained from the original mortgage database, except for the Bond Rate Origination (10-Year Treasury Bond yield), which was computed by the Nelson-Siegel-Bliss method using FORTRAN code provided by Robert Bliss.

## TABLE 4.2

### Panel A

### Loan Origination, Default and Prepayment by Year of Origination

| Year | # Originated | # Defaulted | % Defaulted | # Prepaid | % Prepaid |
|---|---|---|---|---|---|
| 1970 | 240 | 0 | 0.00 | 201 | 83.75 |
| 1971 | 1,165 | 0 | 0.00 | 989 | 84.89 |
| 1972 | 1,906 | 0 | 0.00 | 1,526 | 80.06 |
| 1973 | 1,993 | 2 | 0.10 | 1,369 | 69.69 |
| 1974 | 1,097 | 4 | 0.36 | 704 | 64.18 |
| 1975 | 1,691 | 4 | 0.24 | 1,056 | 62.45 |
| 1976 | 2,514 | 4 | 0.16 | 1,581 | 62.89 |
| 1977 | 4,012 | 12 | 0.30 | 2,398 | 59.77 |
| 1978 | 3,254 | 14 | 0.43 | 1,983 | 60.94 |
| 1979 | 2,171 | 27 | 1.24 | 1,325 | 61.04 |
| 1980 | 1,096 | 11 | 1.00 | 628 | 57.30 |
| 1981 | 506 | 6 | 1.19 | 304 | 60.08 |
| 1982 | 387 | 6 | 1.55 | 177 | 45.74 |
| 1983 | 904 | 16 | 1.77 | 479 | 52.99 |
| 1984 | 838 | 10 | 1.19 | 486 | 58.00 |
| 1985 | 827 | 23 | 2.78 | 458 | 55.38 |
| 1986 | 3,748 | 97 | 2.59 | 2,028 | 54.11 |
| 1987 | 5,086 | 115 | 2.26 | 2,912 | 57.27 |
| 1988 | 2,913 | 92 | 3.16 | 1,797 | 61.69 |
| 1989 | 2,983 | 134 | 4.49 | 1,779 | 59.64 |
| 1990 | 4,642 | 224 | 4.83 | 2,928 | 63.08 |
| 1991 | 14,343 | 395 | 2.75 | 9,960 | 69.44 |
| 1992 | 44,460 | 690 | 1.55 | 30,488 | 68.42 |
| 1993 | 83,973 | 569 | 0.68 | 44,120 | 52.54 |
| 1994 | 39,859 | 470 | 1.18 | 23,714 | 59.49 |
| 1995 | 42,529 | 502 | 1.18 | 25,793 | 60.65 |
| 1996 | 23,539 | 189 | 0.80 | 12,280 | 52.17 |
| 1997 | 21,429 | 85 | 0.40 | 11,527 | 53.79 |
| 1998 | 64,837 | 154 | 0.24 | 21,823 | 33.66 |
| 1999 | 208,336 | 564 | 0.27 | 62,108 | 29.81 |
| 2000 | 146,241 | 143 | 0.10 | 68,899 | 47.11 |
| 2001 | 184,084 | 2 | 0.00 | 8,177 | 4.44 |
| Total | 917,703 | 4,564 | 0.50 | 345,997 | 37.70 |

**Panel B**

**Loan Termination, Default and Prepayment by Year of Termination**

| Year | # Terminated | # Defaulted | % Defaulted | # Prepaid | % Prepaid |
|---|---:|---:|---:|---:|---:|
| 1990 | 0 | 0 | 0.00 | 0 | 0.00 |
| 1991 | 19 | 0 | 0.00 | 0 | 0.00 |
| 1992 | 180 | 0 | 0.00 | 0 | 0.00 |
| 1993 | 3,370 | 0 | 0.00 | 2,900 | 86.05 |
| 1994 | 2,703 | 7 | 0.26 | 2,426 | 89.75 |
| 1995 | 2,211 | 45 | 2.04 | 2,069 | 93.58 |
| 1996 | 15,742 | 400 | 2.54 | 14,467 | 91.90 |
| 1997 | 20,742 | 723 | 3.49 | 19,949 | 96.18 |
| 1998 | 63,530 | 767 | 1.21 | 49,915 | 78.57 |
| 1999 | 40,364 | 695 | 1.72 | 39,147 | 96.98 |
| 2000 | 31,950 | 675 | 2.11 | 30,846 | 96.54 |
| 2001 | 736,892 | 1,252 | 0.17 | 184,278 | 25.01 |

**Panel C**

**Active Loans Observed, Default and Prepayment by Year of Observation**

| Year | # Observed | # Defaulted | % Defaulted | # Prepaid | % Prepaid |
|---|---:|---:|---:|---:|---:|
| 1990 | 33 | 0 | 0.00 | 0 | 0.00 |
| 1991 | 5,098 | 0 | 0.00 | 0 | 0.00 |
| 1992 | 14,870 | 0 | 0.00 | 0 | 0.00 |
| 1993 | 32,304 | 0 | 0.00 | 2,900 | 8.90 |
| 1994 | 34,919 | 7 | 0.02 | 2,426 | 6.94 |
| 1995 | 133,599 | 45 | 0.03 | 2,069 | 1.55 |
| 1996 | 226,158 | 400 | 0.18 | 14,467 | 6.40 |
| 1997 | 253,182 | 723 | 0.29 | 19,949 | 7.88 |
| 1998 | 319,984 | 767 | 0.24 | 49,915 | 15.60 |
| 1999 | 452,783 | 695 | 0.15 | 39,147 | 8.65 |
| 2000 | 560,304 | 675 | 0.12 | 30,846 | 5.50 |
| 2001 | 736,892 | 1,252 | 0.17 | 184,278 | 25.00 |

The Table represents the number and percentage ($\frac{\#\ of\ defaults(prepayments)\ in\ year\ i}{\#\ of\ loans\ originated(terminated/active)\ in\ year\ i}$) of defaults and prepayments by the year of loan origination and termination (Panels A and B), as well as the number of loans observed each year (Panel C). Panel A also presents the total number of defaults and prepayments in the sample. The first loans are included in the observation plan in August 1990, but there are no recorded defaults or prepayments prior to 1991. Note that the sum of defaulted and prepaid loans in any particular year is not equal to the total number of terminations in that year, since loans can be sold, excluded from the observation plan (which explains the extremely high number of terminations is 2001), or paid off during the observation period, the latter obviously being relevant only for loans originated in 1970-71.

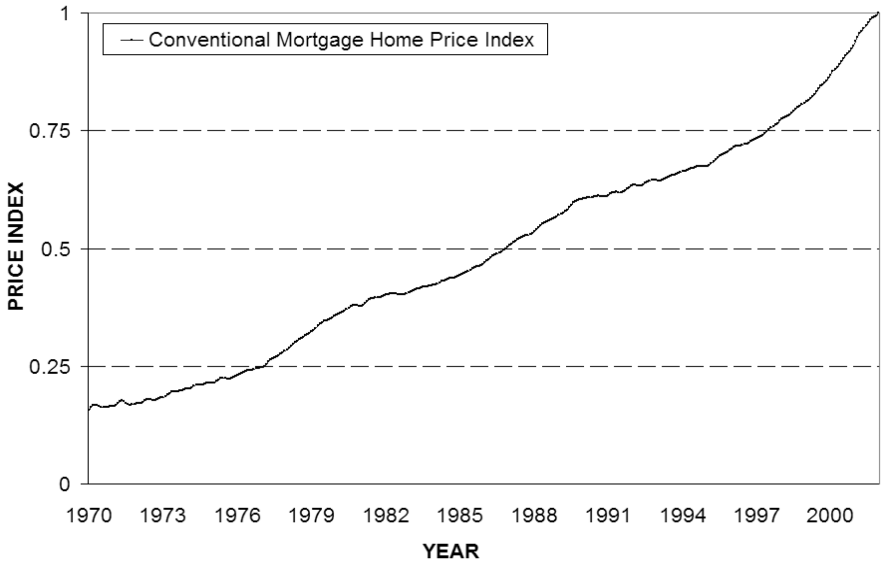
## TABLE 4.3

**Pearson Correlation Coefficients**

Sample Size, N = 25,780,497

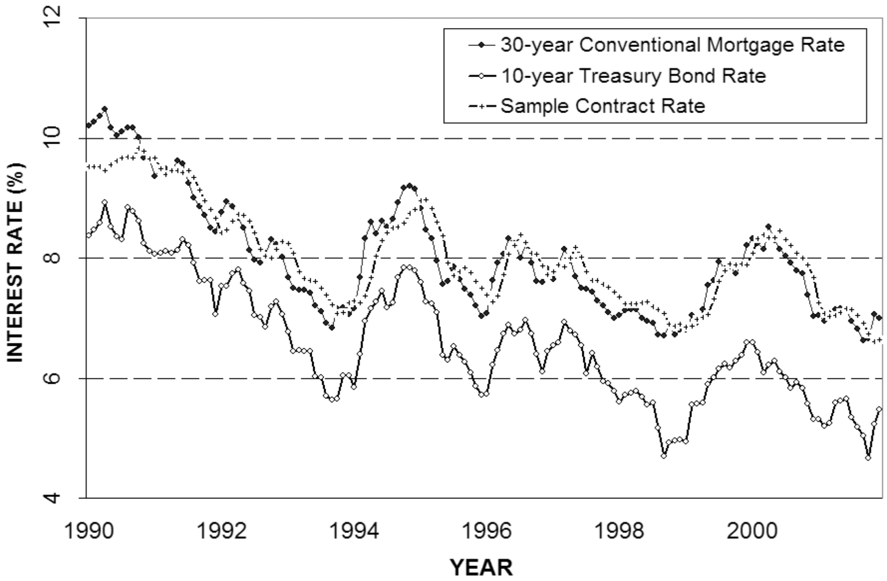
| **Variable** | *LTV* | *LTV Sq.* | *Log Loan Size* | *Log Loan Size Sq.* | *Cont. Rate* | *Cont. Rate Sq.* | *Points* | *Points Sq.* | *Spread* | *Spread Sq.* |
|---|---|---|---|---|---|---|---|---|---|---|
| *LTV* | 1.000 | 0.985 | 0.106 | 0.103 | 0.052 | 0.053 | -0.150 | -0.039 | 0.061 | 0.010 |
| *LTV Squared* | . | 1.000 | 0.067 | 0.064 | 0.060 | 0.060 | -0.150 | -0.031 | 0.063 | 0.011 |
| *Log Loan Size* | . | . | 1.000 | 0.999 | -0.135 | -0.134 | 0.018 | -0.050 | -0.084 | -0.022 |
| *Log Loan Size Squared* | . | . | . | 1.000 | -0.134 | -0.134 | 0.017 | -0.048 | -0.084 | -0.021 |
| *Contract Rate* | . | . | . | . | 1.000 | 0.994 | -0.013 | -0.029 | 0.606 | 0.076 |
| *Contract Rate Squared* | . | . | . | . | . | 1.000 | -0.006 | -0.024 | 0.583 | 0.099 |
| *Points* | . | . | . | . | . | . | 1.000 | 0.097 | -0.101 | -0.015 |
| *Points Squared* | . | . | . | . | . | . | . | 1.000 | -0.052 | 0.011 |
| *Spread* | . | . | . | . | . | . | . | . | 1.000 | -0.180 |
| *Spread Squared* | . | . | . | . | . | . | . | . | . | 1.000 |

Figure 4.1: **Home Price Index, 1970-2001**



The Conventional Mortgage Home Price Index (CMHPI) is based on conventional mortgages that were purchased or securitized by Freddie Mac or Fannie Mae. It can be obtained from Freddie Mac at `www.freddiemac.com/finance/cmhpi/` . To produce series of monthly frequency, I interpolate quarterly data using the cubic spline method. The resulting data series are normalized to make the price level in December 2001 equal to one.

Figure 4.2: **Interest Rates During the Observation Period, 1990-2001**



Interest rate data come from three different sources. The 10-year Treasury Bond Rate is computed using the Nelson-Siegel-Bliss method. The Conventional Mortgage Rate is obtained from the database FRED II maintained by the Federal Reserve Bank of St. Louis and available at `/research.stlouisfed.org/fred2/`. Finally, the Sample Mortgage Rate is the average for all loans in the mortgage database by the month of origination.

Figure 4.3: **Kaplan-Meier Estimates of Default Hazard**

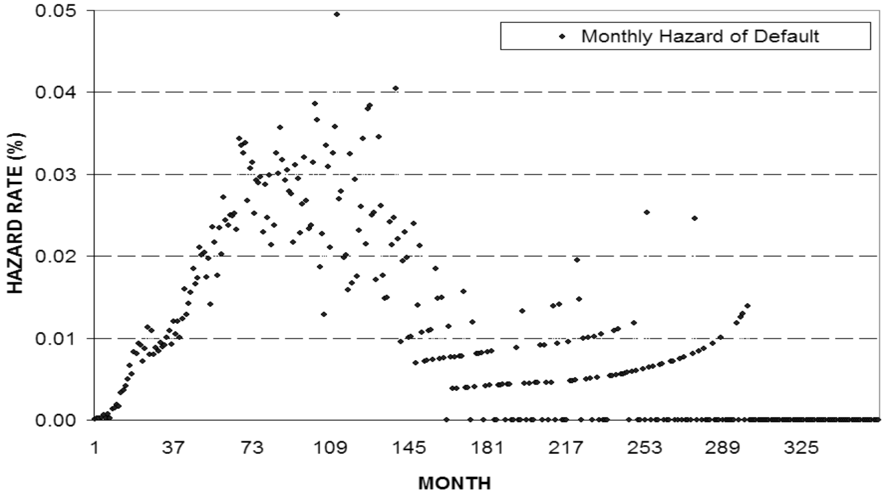
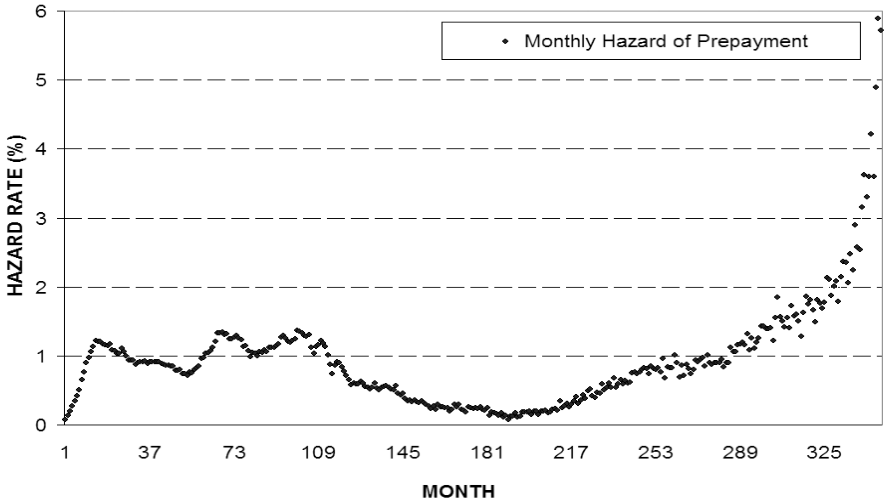


Figure 4.4: **Kaplan-Meier Estimates of Prepayment Hazard**



The non-parametric Kaplan-Meier estimator (product limit estimator) is a strictly empirical approach to hazard function estimation. The estimator of the hazard of default(prepayment) is $\frac{\#\ of\ defaults(prepayments)\ at\ time\ i}{\#\ of\ observed\ loansat\ risk\ at\ time\ i}$, where $i = 1, 360$ is the month in mortgage life. For presentation purposes, the last 12 observations in the prepayment figure are truncated due to the very low number of surviving loans, resulting in extremely high hazard estimates.

**TABLE 4.4**

**Panel A**

**Tests for Optimal Lag Length on Interest Rate Spread**

Sample Size, N=917,703

| **Lag Length** | **AIC** | **SBC** |
|---|---:|---:|
| 0 | 6,384,912.8 | 6,385,084.5 |
| 1 | 6,383,579.3 | 6,383,751.1 |
| 2 | 6,381,216.9 | 6,381,388.5 |
| 3 | 6,382,805.7 | 6,382,977.4 |
| 4 | 6,385,029.5 | 6,385,201.2 |
| 5 | 6,385,041.7 | 6,385,213.3 |
| 6 | 6,385,035.4 | 6,385,207.2 |

**Panel B**

**Tests for Joint Significance**

Sample Size, N=917,703

Degrees of Freedom, Q=10

| **Variables** | **Wald Criterion** | **P-value** |
|---|---:|---:|
| *All Squared Terms* | 9660.43 | 0.000 |
| | Degrees of Freedom, Q=5 | |
| *LTV, LTV Squared,*<br>*Contract Rate Sq.,*<br>*Points Sq., Spread* | 55.97 | 0.000 |

**Panel C**

**Tests for the Presence of Quarterly Stratification**

Sample Size, N=917,703

Degrees of Freedom, Q=20

| **Stratification** | **C-Statistic** | **P-Value** |
|---|---:|---:|
| *Quarterly vs. None* | 8804.69 | 0.000 |
| *Quarterly vs. Annual* | 1734.03 | 0.000 |
| *Quarterly vs. Semiannual* | 1297.03 | 0.000 |

$AIC = -2\ln(L) + 2K$ is Akaike's Information Criterion. $SBC = -2\ln(L) + \ln(n)K$ is Schwarz's Bayesian Criterion. The AIC and SBC are used to compare competing models' fit to the same data. The model with the smaller information criteria is said to fit the data better. C-statistic is the clustering statistic of Ridder and Tunali [121]. Under the null hypothesis of the lower level of stratification, the statistic has a Chi-squared distribution with K degrees of freedom, where K is the number of covariates in both models.

**TABLE 4.5**

**Stratified Proportional Hazard Estimates for**

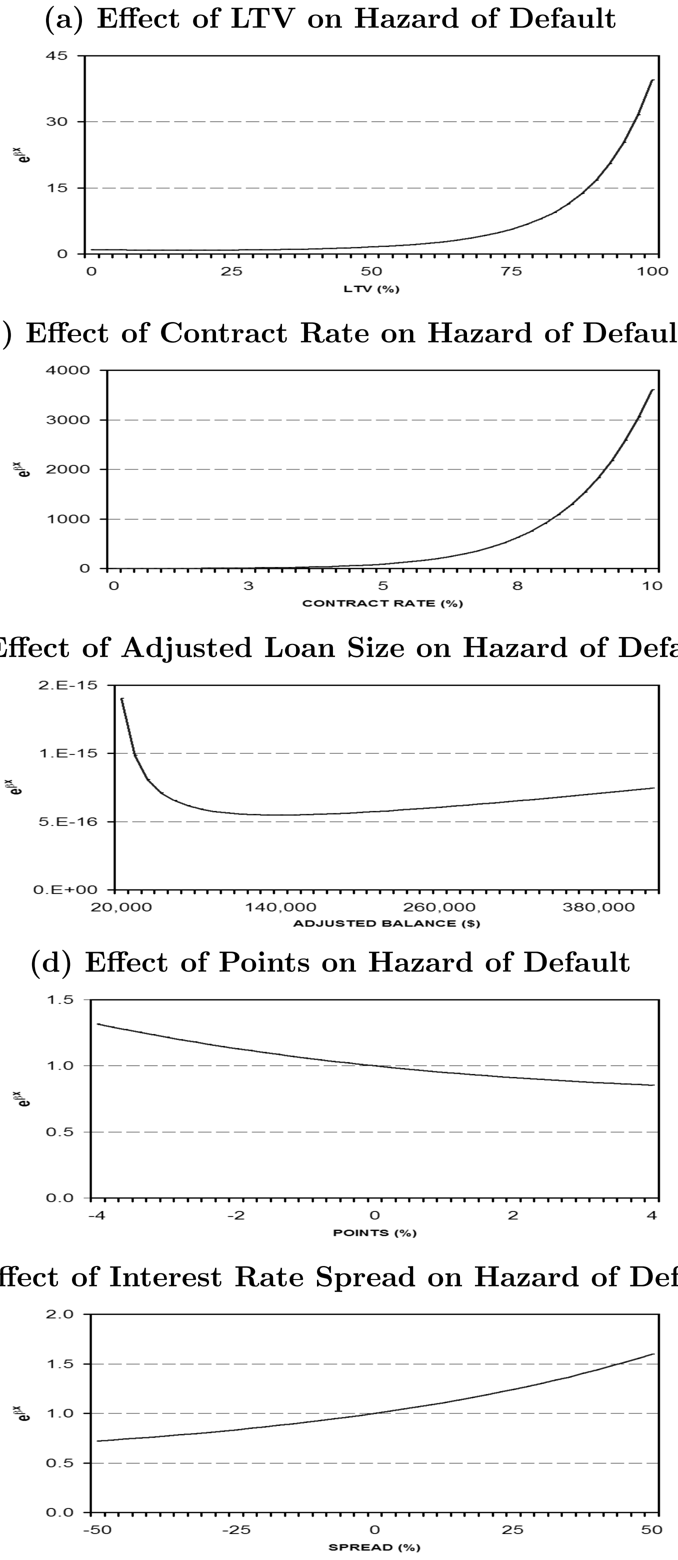
**Competing Risks of Default and Prepayment**

Sample Size: N=917,703

| | **Default Model** | **Prepayment Model** |
|---|---:|---:|
| **Variable** | **Estimate** | **Estimate** |
| | **(Standard Error)** | **(Standard Error)** |
| | **(Corrected Standard Error)** | **(Corrected Standard Error)** |
| *Original LTV* | -0.01810 | -0.00032 |
| | (0.01039) | (0.00063) |
| | (0.01039) | (0.00062) |
| *Original LTV Squared* | 0.00055 | $-3.12440 \times 10^{-6}$ |
| | (0.00007) | ($4.48770 \times 10^{-6}$) |
| | (0.00007) | ($4.42600 \times 10^{-6}$) |
| *Contract Rate* | 0.97596 | 1.89780 |
| | (0.25444) | (0.04139) |
| | (0.25436) | (0.04096) |
| *Contract Rate Squared* | -0.01569 | -0.11102 |
| | (0.01205) | (0.00229) |
| | (0.01204) | (0.00227) |
| *Log Loan Size* | -5.93580 | 2.80487 |
| | (1.09216) | (0.15338) |
| | (1.09132) | (0.15137) |
| *Log Loan Size Squared* | 0.25067 | -0.09786 |
| | (0.04676) | (0.00650) |
| | (0.04673) | (0.00642) |
| *Points* | -0.05412 | -0.07395 |
| | (0.02661) | (0.00265) |
| | (0.02659) | (0.00261) |
| *Points Squared* | 0.00368 | -0.01375 |
| | (0.00297) | (0.00117) |
| | (0.00294) | (0.00116) |
| *Spread* | 0.00794 | 0.04907 |
| | (0.00762) | (0.00082) |
| | (0.00762) | (0.00081) |
| *Spread Squared* | 0.00003 | 0.00008 |
| | (0.00001) | ($9.78000 \times 10^{-7}$) |
| | (0.00001) | ($9.78000 \times 10^{-7}$) |

Note that the Log Loan Size variable was formed by inflating a loan's size to match the Freddie Mac 2001 Conventional Mortgage Home Price Index (CMHPI) (available at `www.freddiemac.com/finance/cmhpi/` ), before being taken into logs. The spread variable is the difference between the loan's contract rate and the current ten-year Treasury rate, as a percent of the contract rate. It is then a time-varying covariate. The first listed standard error is the conventional one for proportional hazard models, whereas the second includes a correction required by the Prentice and Kalbfleisch modification of the Cox proportional hazard model.

Figure 4.5: **Effects of Covariates on Hazard of Default**



**(a) Effect of LTV on Hazard of Default**

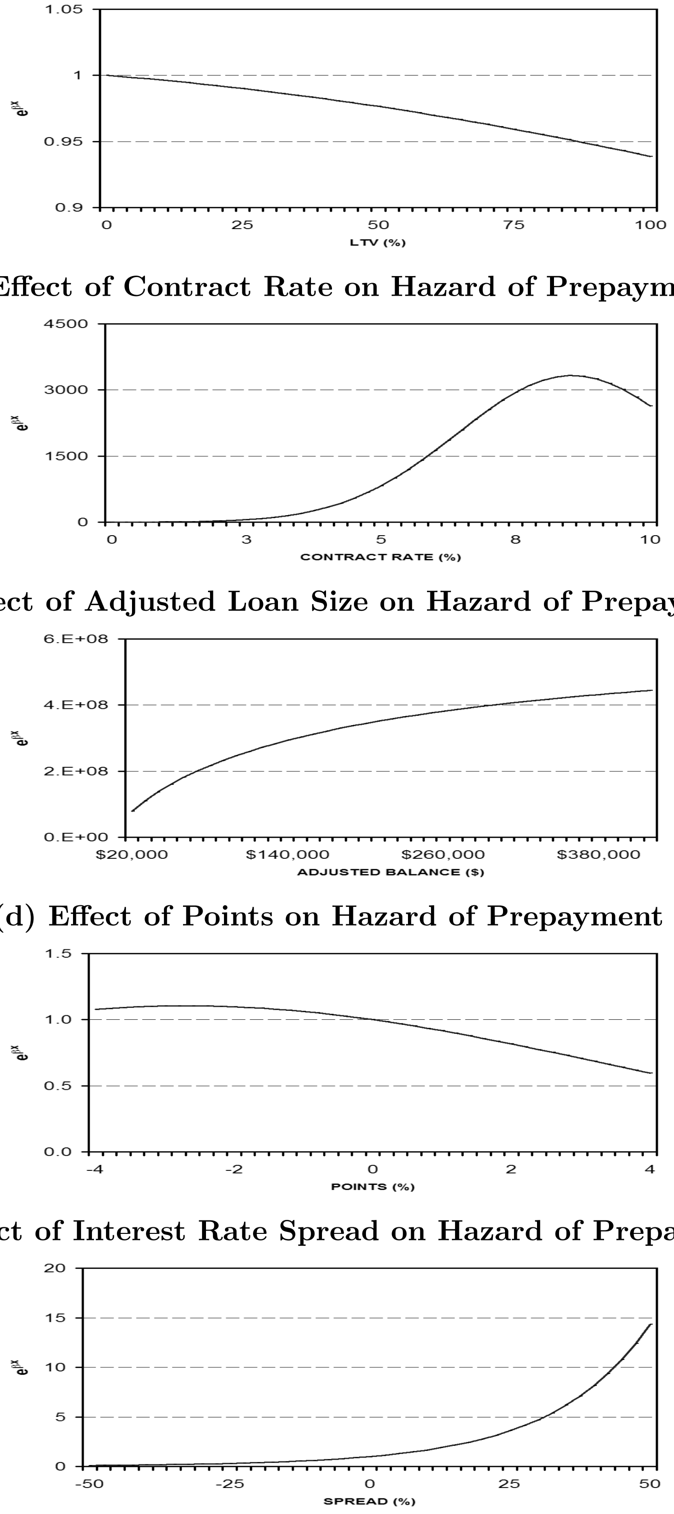
**(b) Effect of Contract Rate on Hazard of Default**

**(c) Effect of Adjusted Loan Size on Hazard of Default**

**(d) Effect of Points on Hazard of Default**

**(e) Effect of Interest Rate Spread on Hazard of Default**

Figure 4.6: **Effects of Covariates on Hazard of Prepayment**



**(a) Effect of LTV on Hazard of Prepayment**

**(b) Effect of Contract Rate on Hazard of Prepayment**

**(c) Effect of Adjusted Loan Size on Hazard of Prepayment**

**(d) Effect of Points on Hazard of Prepayment**

**(e) Effect of Interest Rate Spread on Hazard of Prepayment**

Figure 4.7: **Predicted Hazard of Default with Covariates at the Mean**

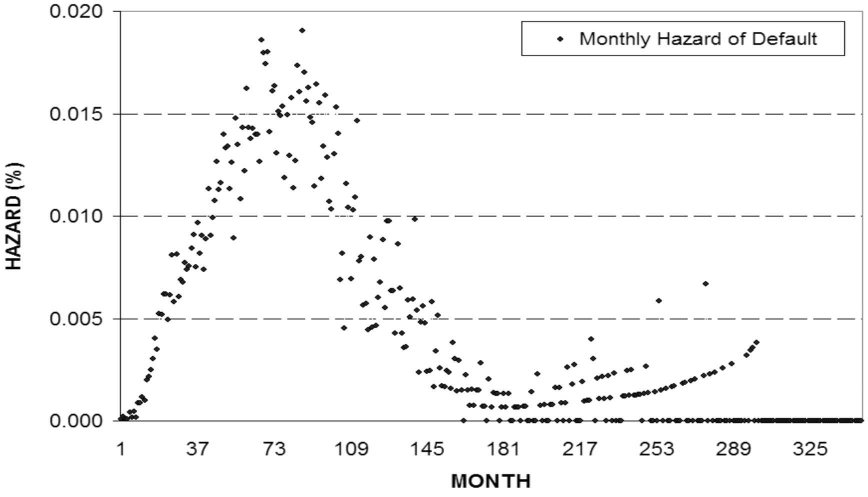
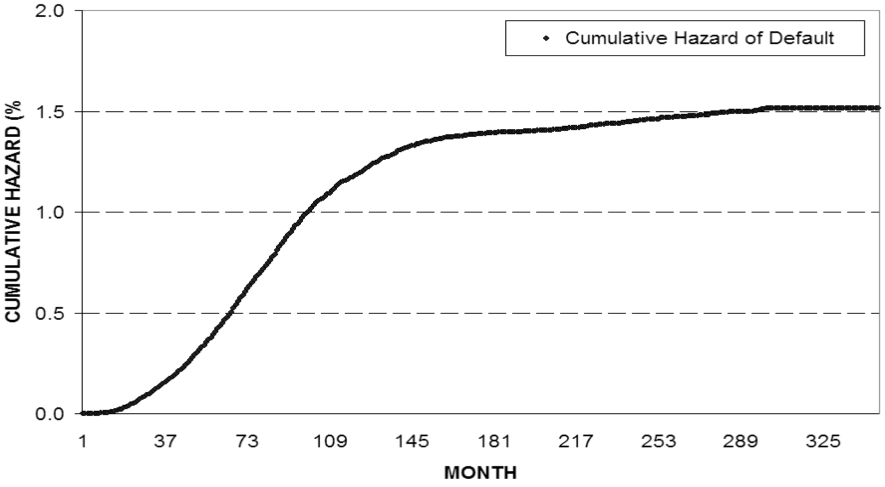


Figure 4.8: **Cumulative Predicted Hazard of Default with Covariates at the Mean**



The predicted hazard of default at time $t(i)$ is obtained as $\hat{\lambda}(\hat{\beta}, t(i)) = \frac{1}{J} \sum_{j=1}^{J} \hat{\lambda}_j(\hat{\beta}, t(i)) = \frac{1}{J} \sum_{j=1}^{J} \frac{\sum_{h=1}^{n(j)} N_{hj}(t(i))}{\sum_{k=1}^{n(j)} Y_{kj}(t(i)) e^{\hat{\beta}' \mathbf{x}_{kj}(t(i))}} e^{\hat{\beta}' \bar{\mathbf{x}}(t(i))}$, where $N_{hj}(t(i))$ indicates default of the $h$th mortgage in strata $j$ at time $t(i)$, and $Y_{kj}(t(i))$ is an indicator variable, equal to one if the $k$th mortgage in strata $j$ was under observation at time $t(i)$, and zero otherwise. All covariates are set to their sample means, as reported in **Table 4.1**. The predicted cumulative hazard is obtained in the usual way, as $1 - S(t(i))$, where the survival function $S(t(i))$ is defined by $S(t(i)) = \prod_{i=1}^{t(i)} (1 - \lambda(t(i)))$. Dependence of $S(t(i))$ and $\lambda(t(i))$ on $\hat{\beta}$ is ignored for the purpose of convenience.

Figure 4.9: **Predicted Hazard of Prepayment with Covariates at the Mean**

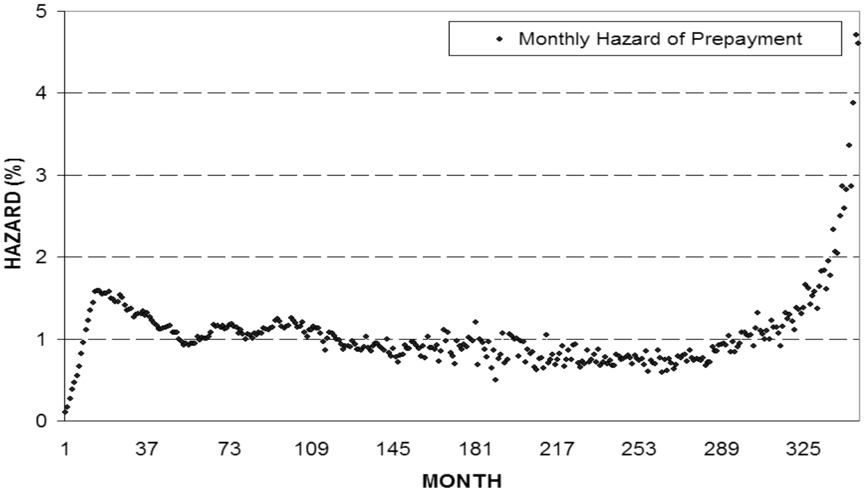
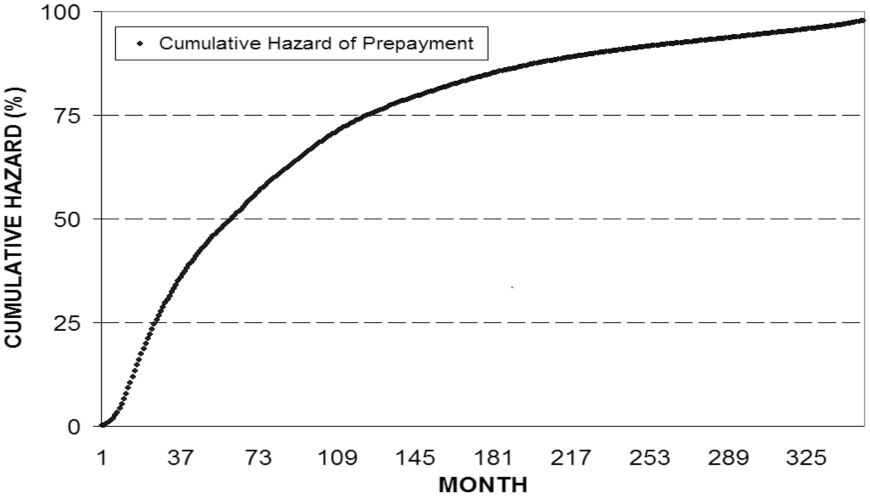


Figure 4.10: **Predicted Cumulative Hazard of Prepayment with Covariates at the Mean**



The predicted hazard of prepayment at time $t(i)$ is obtained as $\hat{\lambda}(\hat{\beta}, t(i)) = \frac{1}{J}\sum_{j=1}^{J}\hat{\lambda}_j(\hat{\beta}, t(i)) = \frac{1}{J}\sum_{j=1}^{J}\frac{\sum_{h=1}^{n(j)} N_{hj}(t(i))}{\sum_{k=1}^{n(j)} Y_{kj}(t(i))e^{\hat{\beta}'\mathbf{x}_{kj}(t(i))}}e^{\hat{\beta}'\bar{\mathbf{x}}(t(i))}$, where $N_{hj}(t(i))$ indicates prepayment of the $h$th mortgage in strata $j$ at time $t(i)$, and $Y_{kj}(t(i))$ is an indicator variable, equal to one if the $k$th mortgage in strata $j$ was under observation at time $t(i)$, and zero otherwise. All covariates are set to their sample means, as reported in **Table 4.1**. The predicted cumulative hazard is obtained in the usual way, as $1 - S(t(i))$, where the survival function $S(t(i))$ is defined by $S(t(i)) = \prod_{i=1}^{t(i)}(1 - \lambda(t(i)))$. Dependence of $S(t(i))$ and $\lambda(t(i))$ on $\hat{\beta}$ is ignored for the purpose of convenience.

Figure 4.11: **Cumulative Hazard Rate of Default by LTV ratio**

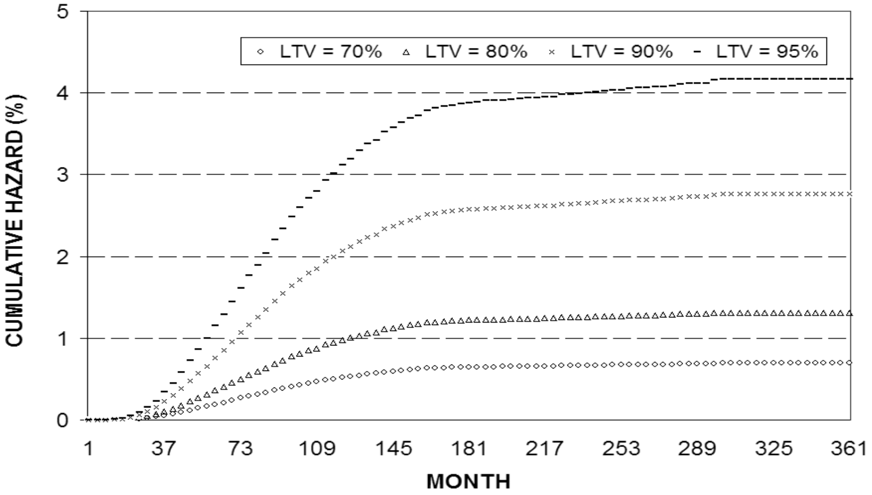
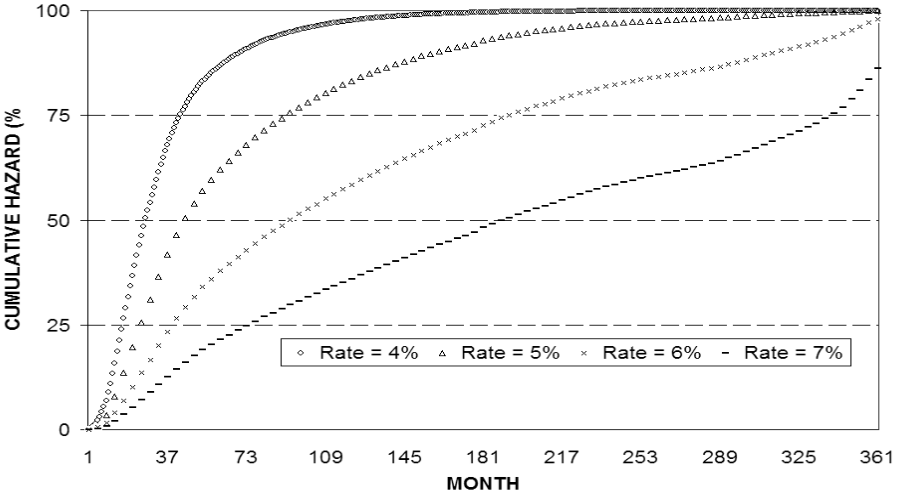


Figure 4.12: **Cumulative Hazard Rate of Prepayment by Interest Rate at Origination**



The predicted cumulative hazard is obtained in the usual way, as $1 - S(t(i), \hat{\beta}, \tilde{\mathbf{x}})$, where the survival function $S(t(i), \hat{\beta}, \tilde{\mathbf{x}})$ is defined by $S(t(i), \hat{\beta}, \tilde{\mathbf{x}}) = \prod_{i=1}^{t(i)} (1 - \lambda(t(i), \hat{\beta}, \tilde{\mathbf{x}}))$. All covariates in $\tilde{\mathbf{x}}$ are set to their sample means, except for ones indicated in the titles of respective figures. In particular, the contract rate is held constant.

# Chapter 5

# Estimation of Termination Processes Using Particle Filters

To maintain the doubly stochastic nature of the pricing model, I estimate the two termination processes within the state-space framework, allowing for two layers of uncertainty, the first one representing the uncertainty of individual mortgage termination, given the realization of the hazard process, and the second one modelling the stochastic diffusion of the hazard process itself. The chapter has four sections. The first one describes the basic ideas behind particle filtering, as well as some of the most common types of filters. The second section contains a description of my empirical filtering model specification. Since particle filters are still a relatively novel technique, in the third section I provide some implementation notes on programming a particle filter using FORTRAN (the pseudo-code for the algorithm is given in the Appendix). The fourth section discusses data and results.

## 5.1 Particle Filters in the Context of Bayesian Tracking

In this section, I describe general methods of estimating the state of a system with noisy observations via particle filtering. This relatively new Bayesian approach allows one to obtain estimates of the state as well as parameters of the underlying densities, without relying either on the linearity of the model, or on the normality assumption, or even on any approximations to the Gaussian distribution. Moreover, even though I employ a discrete-time setting in this section, continuous-time models can easily be analyzed within the particle filtering framework (as explained in the next section), thus making the approach more appealing for the problem at hand than the more conventional Kalman filter and the extended Kalman filter methods.

A general state space model is specified as

$$\begin{aligned} y(t(i)) &\sim f(y(t(i))|\lambda(t(i))) \qquad &&- \text{Observation equation} \\ \lambda(t(i))|\lambda(t(i-1)) &\sim f(\lambda(t(i))|\lambda(t(i-1))) \qquad &&- \text{Transition equation} \end{aligned} \tag{5.1}$$

where $t(i)$ is the time of $i$'th observation.

The observations in the sequence $\{y(t(i)), i = 1, ..., n\}$ are assumed to be independent, and the states $\lambda(t(i))$ are assumed to be first-order Markovian. It is assumed that the *measurement density* $f(y(t(i))|\lambda(t(i)))$ is known, and one can simulate from the *transition density* $f(\lambda(t(i))|\lambda(t(i-1)))$, with the latter initialized by a known value $f(\lambda(t(0)))$.

Filtering occurs in two stages. At first stage (*propagation*), the current density must be propagated forward in time, using the transition equation to produce the prediction density (Chapman-Kolmogorov equation):

$$f(\lambda(t(i))|y(t(1:i-1)) = \int f(\lambda(t(i))|\lambda(t(i-1))f(\lambda(t(i-1))|y(t(1:i-1)))d\lambda(t(i-1)) \tag{5.2}$$

At time $t(i)$, a new measurement arrives, and is used to update the prior (*update* stage) via Bayes' rule to obtain the filtering density:

$$f(\lambda(t(i))|y(t(1:i)) = \frac{f(y(t(i))|\lambda(t(i)))f(\lambda(t(i))|y(t(1:i-1)))}{f(y(t(i))|y(t(1:i-1)))} \tag{5.3}$$

$$f(y(t(i))|y(t(1:k-1))) = \int f(y(t(i))|\lambda(t(i)))f(\lambda(t(i)|y(t(1:i-1))d\lambda(t(i)) \tag{5.4}$$

Equations (5.2), (5.3) and (5.4) form the foundation for the optimal Bayesian solution. However, with the exception of linear Gaussian models, the recursive time propagation of the posterior density (5.4) is only a conceptual solution to the problem, and requires some degree of approximation. One possible solution is to use the Taylor series expansion of non-linear functions, producing the extended Kalman Filter (still approximating the filtering density to be normal[1]). Another way to address this issue is to use simulation-based particle filters.

The body of literature on simulation-based filters is becoming increasingly extensive. The term "particle filter" is due to Carpenter et al. [24], however, the same method is often used under different names, such as, for example, bootstrap filtering (Gordon et al. [71]), the sequential Monte Carlo approach, on-line Bayesian tracking (Arulampalam et al. [7]), or the condensation algorithm (Izard and Blake [78]). Most applications of particle filtering have been concentrated in physics and related fields. Berzuini et al. [9] applied particle filters to sequential analysis of medical patients, Izard and Blake [78] used them for robust tracking of motion in a visual clutter. Gordon et al. [71][70], Pitt and Shephard [112], among others, used Bayesian tracking on the bearings-only navigation model. A large collection of different applications is presented in Doucet et al. [46]. Some progress, however, is being made in economic and financial applications as well, like the estimation of stochastic volatility model of exchange rates of the US dollar vs. the UK pound in Pitt and Shephard [112], modelling term structure of interest rates in Durham and Gallant [61], or the disequilibrium model analysis in Lui and Chen [102]. A variety of different simulated models applicable to econometric time series can also be found in Pitt [113].

Particle filters recursively approximate density (5.4) by drawing a large sample of particles $\{\lambda(t(i))^1, ..., \lambda(t(i)^M)^M\}$ with discrete probability mass (weights) of

[1] A transformation of the extended Kalman filter, the unscented Kalman filter, deterministically selects points form the Gaussian approximation, and uses them for propagation and re-estimation of the model. Although this method gives better results in a non-linear setting than the extended Kalman filter, it is also based on the Gaussian approximation.

$\pi(t(i))^1, ..., \pi(t(i))^M$, and computing the estimates based on these particles and weights. This allows one to produce an approximation to the prediction density (5.2) by using the discrete support of the particles:

$$\hat{f}(\lambda(t(i+1))|y(t(1:i))) = \sum_{m=1}^{M} f(\lambda(t(i+1))|\lambda^m(t(i)))\pi^m(t(i)) \tag{5.5}$$

This expression is later combined with the measurement density to produce the empirical filtering density:

$$\hat{f}(\lambda(t(i))|y(t(1:i))) \propto f(y(t(i))|\lambda(t(i)))\hat{f}(\lambda(t(i))|y(t(1:i-1))) \tag{5.6}$$

For each following observation, the particle filter samples new particles from the corresponding empirical filtering density to produce particles $\lambda(t(i+1)^1), ..., \lambda(t(i+1)^M)$ with corresponding weights, and the process is propagated through time using the transition equation. As the number of particles increases, this process produces an accurate description of the posterior at each point in time.

This general method of propagating particles and weights is called the Sequential Importance Sampling Algorithm (SIS). A common problem with SIS filters is degeneracy, i.e. the situation when after a few propagation steps only a few particles will have relatively high weights, the others having weights close to zero (due to the increase in the variance of weights as $i$ increases). The "unimportant" particles, however, are still propagated through the filter, consuming valuable computational resources. One proposed solution to this problem is to use the Sampling Importance Resampling Algorithm (SIR) of Rubin [122] and Gordon et al. [71]. This is the method I use to estimate diffusion parameters for default and prepayment.

To implement the algorithm, one samples $R$ particles at time $t(i)$ from a set of $M$ particles with equal weights, and propagates them via the transition density in (5.1). After that, one computes their unnormalized weights, given by

$$w^r(t(i+1)) = f(y(t(i+1))|\lambda^r(t(i+1)))$$

and the normalized weights given by:

$$\pi^r(t(i+1)) = \frac{w^r(t(i+1))}{\sum_{k=1}^{R} w^k(t(i+1))}$$

At the last stage, the particles are resampled to form a random sample of size $M$ (with equal weights) using their normalized weights, in order to then be sampled again at the next time step, and so the algorithm is repeated until the end of the time series is reached.

In this dissertation, however, my main objective is not just prediction of the state of the system $\hat{f}(\lambda(t(i))|y(t(1:i)))$, but evaluation of the log-likelihood function

$$\ln L(\Theta) = \log f(y(t(1),...,t(n))|\Theta) = \sum_{i=1}^{n} \ln f(y(t(i))|\Theta, y(t(1:i-1))) \tag{5.7}$$

where $\Theta$ represents a vector of parameters, possibly occurring in both the transition and the observation equations. To construct the likelihood, I will be using the one-step-ahead density obtained via prediction decomposition:

$$f(y(t(i))|\Theta, y(t(1:i{-}1))) = \int f(y(t(i))|\Theta, \lambda(t(i)))f(\lambda(t(i))|\Theta, y(t(1:i{-}1)))d\lambda(t(i)) \tag{5.8}$$

As shown in Pitt [113], the likelihood for the basic SIR can be obtained using the conventional Monte Carlo estimator:

$$\hat{f}(y(t(i))|\Theta, y(t(1:i-1))) = \frac{1}{R}\sum_{j=1}^{R} f(y(t(i+1))|\lambda^j(t(i+1))) \tag{5.9}$$

This likelihood can be evaluated and maximized in a rather straightforward manner, and will produce an unbiased [96] and efficient (both statistically and computationally)[2] estimate of $\Theta$ as $R \rightarrow \infty$, given functional specifications for the transition equation and the observation equation, which are described in the next section.

## 5.2 The SIR Particle Filtering Model

In this section, I present a model of mortgage termination via default and prepayment. The model is set up using the state space framework, i.e. assuming that mortgage terminations, as described by their own evolution processes, are observed with error in a non-linear, non-Gaussian setting. This state space specification allows me to maintain the doubly stochastic nature of the pricing model.[3]

The observation equation of the model of mortgage termination is specified as

$$g(y_{i,j}^l = n | \lambda_{0,j}^l(t_i)) = \frac{\mu_{i,j}^l}{n!} e^{-\mu_{i,j}^l} \qquad - \text{ Poisson} \tag{5.10}$$

with parameter

$$\mu_i^l, j = \sum_{k=1}^{n(i,j)} \lambda_{i,k}^l \tag{5.11}$$

[2] For statistical efficiency, weights $w^r(t(i))$ should have a small variance; see Lui [101].

[3] In an earlier version of the dissertation, I used a simulated maximum likelihood estimation (SMLE) method to estimate hazards, which effectively assumes that the sampled mean rate of termination coincides with the true population mean, so that one has a trivial observation equation. This approach can still be consistent, but has poor finite sample properties. Despite the fact that there are usually many thousand mortgages under observation within any stratum, the monthly probability of termination is so low that, at least for default, the law of large numbers need not apply. As an example of the difficulties, there are numerous months without default, though the likelihood of default is presumably never actually zero. I currently employ this faster SMLE method to yield starting parameter values for our particle filter estimation.

where $l$ is an indicator variable of either default or prepayment, $y^l_{i,j} = y^d_{i,j}$ ($y^l_{i,j} = y^p_{i,j}$) is the actual number of defaults (prepayments) in the $j$'th stratum at the $i$'th observation (payment) date, and $\lambda^l_{i,k} = \mathbf{e}^{\beta'\mathbf{x}^k(t(i))}\lambda^l_{0,j}(t(i))$ is the realized probability of default or prepayment for the $k$'th loan in the $j$'th stratum at time $t(i)$, specified within the proportional hazard framework. At any time $t(i)$, $n(i,j)$ is the number of loans at risk in strata $j$. The vector of coefficients $\beta$ has been estimated at the previous stage using the stratified version of the Breslow partial likelihood estimator, and $\mathbf{x}$ is the vector of realized loan and time-specific covariate values.[4]

The Poisson specification is based on the assumption that, given the estimated hazard rate, individual mortgages terminate independently of one another, with the number of actual terminations binomially distributed. Since at any specific data the number of terminations is low and the number of loans in the risk set is high, one can use the Poisson approximation of the binomial process.[5]

The second part of the state space system, the transition equation, represents the stochastic evolution of the hazard rate, i..e the second level of stochasticity in the model. In the general form, one can write the processes governing the evolution of discrete baseline hazard $\lambda^l_0(t)$ as

$$d\lambda^l_0(t) = a(\lambda^l_0(t), t)dt + b(\lambda^l_0(t), t)dz^{\mathbb{P}}(t), \tag{5.12}$$

[4] In the estimation of the termination processes, one must replace the unknown $\boldsymbol{\beta}$ with the $\hat{\boldsymbol{\beta}}$ estimated in the previous stage, but this has no effect on consistency, and given the large number of individual houses, little effect on standard errors, which remain asymptotically correct.

[5] Since there is both default and prepayment, one actually has a multinomial distribution. Furthermore, the baseline intensities must be weighted by their covariates, so the actual probabilities, while conditionally independent, are not identically distributed. Nonetheless, the resulting distribution is well approximated by two independent Poisson variables, each with a mean that is the sum of the individual means for that form of termination. See, for instance, [135] for a modern form of the appropriate Poisson approximation result. Note that the interaction of default and prepayment is second-order small, and so disappears in the limit.

where $a$ and $b$ are the drift and diffusion of the process $\lambda_0^l(\cdot)$ , respectively, and $dz^{\mathbb{P}}(t)$ is the standard Brownian motion under the physical measure $\mathbb{P}$.

I choose to specify both diffusions as widely known Cox-Ingersoll-Ross square-root processes, since CIR processes have an attractive feature of precluding negative values and are relatively easy to estimate and interpret. In general form, a CIR process is defined as

$$d\lambda_0^l(t) = \kappa^l(\theta^l(t) - \lambda_0^l(t))dt + \sigma^l\sqrt{\lambda_0^l(t)}dW(t) \tag{5.13}$$

where $\kappa^l$ is the mean-reversion parameter, $\theta^l(t)$ is the mean, and $\sigma^l$ is the instant volatility parameter.

Based on prior information about general patterns of termination in mortgage time, I further specify $\theta^l(t)$. For default, I choose the specification

$$\theta^d(t) = \frac{\alpha_d}{\Gamma(\rho_d)2^{\rho_d}}t^{\rho_d-1}e^{-t/2} \qquad - \text{ Unnormalized chi} - \text{square} \tag{5.14}$$

It allows one, by using only two parameters, to specify the way the hazard apparently rises relatively fast, and then trails off slowly during the life of a loan. This specification closely resembles the standard SDA schedule of the Public Securities Association. Note that the mean parameter does not depend on the calendar time—only mortgage time. This specification is later tested versus models with polynomial (up to seven degrees) trends, as well as a more general gamma model, and is found to be the preferred one according to the Schwartz Bayesian Criterion.

For prepayment, I originally chose a specification that closely resembles the conventional PSA prepayment function[6]

[6] The Public Securities Association Prepayment Model (PSA) assumes that when the PSA curve is at the 100% level, the prepayment curve starts at 2% in the first month, and then rises by 0.2% each month until month 30, when it levels at 6% annually.

$$\theta^p(t) = \alpha_p(1 - e^{-\rho_p t}) \qquad - \text{ Adjusted exponential} \tag{5.15}$$

However, this specification does not fit the data particularly well, and is rejected in favor of a model with a sixth-degree polynomial trend[7]

$$\theta^p(t) = \sum_{k=1}^{6} \alpha_i t^k \qquad - \text{ Polynomial} \tag{5.16}$$

Having specified (5.14) and (5.16), I can now proceed to estimating parameters of the model using a SIR particle filter.

## 5.3 Implementation

Particle filters are still a relatively novel class of empirical techniques, and computational routines for their estimation are not included in standard statistical packages, which leaves one with the necessity to write one's own code. Fortunately, collections of algorithms, including pseudo-code, are readily available for public use (see [7][113]).

My program is built in FORTRAN90 around the conventional mathematical libraries LAPACK and SLATEC,[8] as well as some random number and optimization routines provided by Garland Durham, previously used in [60].

The implementation of a particle-filtering algorithm is very similar to that of any other simulated likelihood program. Before starting the algorithm, one must consider a few minor things. First, one has to set the seed of the random number generator (I am using RAN1 of Park and Miller[9]) across iterations, so that different values of

[7] I also try to use a more flexible incomplete beta function specification (used to construct the beta C.D.F., which approximates the exponential C.D.F. used in (5.15)). However, the results were adversely affected by the fact that computations involved in calculating the incomplete beta are extremely time-consuming.

[8] The libraries are available at `www.netlib.org.`

[9] The algorithm the can be found in the collection of "numerical recipes" of Press et al. [117].

$\Theta$ are run through the optimizer with the same set on random numbers. Then, one has to choose proper starting values for $\Theta$. The filter is a relatively slow algorithm compared to conventional SMLE (the latter not having the observation equation), and I use SMLE to produce starting values for the parameters. In addition, one has to find a way to sample $\lambda_0^{l(m)}(t(0))$, i.e. the values of the particles before any observations are made. One way to do so is to make $\{\lambda^{l(m)}(t(0))\}_{m=1}^{M}$ a parameter, as in Durham and Gallant [60]. Another way is to sample the particles from a prior distribution, e.g. using sample statistics or some *a priori* choice of a sampling density. Empirically, all these methods provide similar results, with the first one usually making optimization much slower. In the end, I decided to use the sample mean of the empirical estimates of the baseline hazard as a starting value for the underlying state. Finally, it appears to be helpful to adjust the mean level of the covariates to make the average level of the underlying baseline hazard closer to unity. This adjustment does not present any problems for the likelihood estimation, since the method is invariant to reparametarization. On the contrary, it helps avoid many numerical problems, especially at the propagation stage.

These small computational details out of the way, I can now proceed to the optimization routine. At any point in time $t_i$, one has a sample of $M$ particles, $\lambda_0^{l(1)}(t(i)), ..., \lambda_0^{l(M)}(t(i))$. One has to choose a method of sampling $R$ particles from $M$, obtaining $\lambda_0^{l(1)}(t(i)), ..., \lambda_0^{l(R)}(t(i))$. Following Pitt and Shephard [112], I use the stratified sampling method, which enables more efficient estimation and reduces sample impoverishment. The algorithm, which is based on drawing a single uniform variate and inverting empirical C.D.F.'s, is discussed in detail in Pitt [113].

After the sample of $R$ particles is obtained, one can proceed to the next step, which is the propagation of the particles through the transition equation to the time of the next observation. Since the transition equation in the model is set in continuous time, i.e. as a stochastic differential equation, I need to find some way to approximate the continuous diffusion.

There are several conventional methods for approximating continuous-time diffusions that don't have a closed-form solution for the transition density. The two most widely used are the Euler scheme and the Milstein scheme. Letting $\Delta = (t(i) - t(i-1))/S$, where $S$ is a specified number of simulation subintervals, and $z_0 = \lambda_0^{l(r)}(t(i))$, $r = 1, ..., R$, one can specify the Euler scheme as

$$z_{s+1} = z_s + a(s, z_s)\Delta + b(s, z_s)\sqrt{\Delta} u_s, \quad s = 0, ..., S-1 \tag{5.17}$$

with $z_S = \lambda_0^{l(r)}(t(i+1))$. Here, $a(\cdot)$ and $b(\cdot)$ are defined in (5.13), and $u_s \sim N(0,1)$. The Euler scheme is easily implemented and provides strong convergence of order $\frac{1}{2}$ as $S \to \infty$. The Milstein scheme provides better convergence, but requires the the first-order derivative $b'(s, z_s) = \frac{\partial b(s,z_s)}{\partial z_s}$ to be computed for each value of $z_s$. The scheme is given by:

$$z_{s+1} = z_s + a(s, z_s)\Delta + b(s, z_s)\sqrt{\Delta} u_s + \frac{1}{2} b(s, z_s) b'(s, z_s)(\Delta u_s^2 - \Delta) \tag{5.18}$$

The Milstein scheme, while more time-consuming, does not seem to improve the results much. To keep the model more computationally efficient, I thus use the Euler scheme. To decrease the variance of the Monte Carlo estimator, I use antithetic variables for simulations, i.e. unique random numbers are drawn only for R/2 simulations, while for the other half, $\{u_s\}_{r=R/2+1}^{R} = \{-u_s(r)\}_{r=1}^{R/2}$.

After propagation to time period $t(i+1)$, each variable is assigned its corresponding weight, given by (5.10). At this stage, any scaling/adjustment has to be taken into account (by adjusting the particles back up or down, according to the previous adjustment of the covariates).

After weights are normalized, the particles can be resampled using the same stratified sampling algorithm as the one used at the original sampling stage, with a minor, yet important, modification. At the resampling stage, I have to insure that the sampled particles $\lambda_0^{l(r)}(t(i+1))$ are smooth as a function of $\Theta$.

The logic behind the need for a smoothing step is as follows. Suppose the samples $\lambda_0^{l(m)}(t(i))$ drawn from the filtering density change by a small quantity. Then, the proposal samples $\lambda_0^{l(r)}(t(i+1))$, as well as their discrete probabilities, will change as well. Since the resampling algorithm works by inverting the discrete C.D.F, the resampled particles need not then be close. Empirically, this can represent a considerable problem for the maximizing algorithm. To avoid the problem, I use the smooth bootstrapping algorithm discussed in Pitt [113].

Now (or before the resampling routine), one can use the unnormalized weights of $R$ particles to compute the contribution to the log-likelihood, as given by (5.9). I am using Pitt's [113] suggestion to correct the likelihood bias by first-order Taylor expansion, producing

$$\ln \hat{f}(y(t(i+1))|y(t(1:t))) = \ln \bar{X}(t(i+1)) + \frac{1}{2}\frac{\hat{\sigma}^2(t(i+1))}{R\bar{X}(t(i+1))^2} \tag{5.19}$$

as the final estimate of the contribution to the true likelihood at time $t(i)$, where $\bar{X}(t(i+1))$ is the sample mean of all particles at time $t(i+1)$, and $\hat{\sigma}^2(t(i+1))$ is the corresponding variance.

After running the filter through all observations, I collect the estimates given in (5.19) and construct the log-likelihood in the usual manner as:[10]

[10] In this description of the implementation algorithm, I am abstracting away from the fact that I actually have multiple series, i.e. individual series of defaults and prepayments for each stratum. Empirically, it means that I have to re-initialize the filter each time the algorithm hits the beginning of a new series. Also, the effective number of observations is not the number of observation periods, but the number of observation periods times the number of strata.

$$\hat{l}(\Theta) = \sum_{i=1}^{I} \ln f(y(t(i+\widehat{1}))|y(t(1:t)))$$

The pseudo-code, describing the logic of the algorithm, is given in the Appendix.

## 5.4 Data and Results

The data used for estimation of the state space models described by (5.10), (5.13), (5.14) and (5.16) are represented by series of monthly defaults and prepayments in the sample, together with the scaling terms, which I will call "covariate factors," i.e. $\sum_{k=1}^{n(i,j)} e^{\beta' \mathbf{x}^k(t(i))}$, for each stratum, for the number of periods each stratum remained under observation (i.e. for a maximum of eleven years, for 1970-1990 strata). Hence, the design of the data resembles an unbalanced panel. For the default process, I use the entire sample of 12,111 observations. For prepayment, the sample was truncated to include only the first 29 years in mortgage time, because, in the last year, virtually all mortgages in the sample were prepaid, driving the monthly conditional hazard of prepayment from less than 1% to almost 100%. It has proven to be impossible to fit this spike in the data properly, and any marginal improvement in the fit for this "tail" in the data would always come at an expense of a worse fit for earlier years. Since fitting the prepayment process in the earlier years of a mortgage is much more important both in terms of present value and in terms of the number of mortgages in the strata that are observed in those yearlies years,[11] the observations for the last year were deleted from the sample. This leaves me with 11,985 observations for estimation of the prepayment process.

Since the covariate factors are much greater in magnitude than the number of observed terminations, some prior adjustments are made to the data. The covariate

[11] Mortgages observed in their thirtieth year are originated in 1970-72, and represent 0.36% of the loans in the data set.

factors are scaled by $10^{-7}$ for default and by $10^{15}$ for prepayment, with the starting values for diffusions changed accordingly.

The number of intermediate steps is set to be 30 for both processes, representing a daily frequency for the simulated data. The number of starting particles $M$ is set to 1,000, and the number of propagated particles $R$ is set to 1,300. To avoid computational difficulties, I use the Euler scheme for approximating diffusions.

Before proceeding any further, I need to choose a functional form for the trend of each processes. Since my particle filter method is based on maximizing a likelihood function, general likelihood-based tests of model specification are applicable. The original trends of both the default process and the prepayment process are specified as polynomials,[12] up to the seventh order.[13] Since these polynomial models are nested, the likelihood ratio test can be used. In the default model, the third-degree polynomial is the preferred trend. In the prepayment model, the sixth-degree polynomial trend is preferred. The log-likelihood functions for each polynomial trend model (orders zero to seven), along with corresponding parameter estimates, are reported in Tables 5.1 & 5.2.

After estimating different polynomial models, I specify some additional ones, based on prior evidence given by structural simulations, as well as by the models used by Wall Street practitioners: the SDA model of default, and the PSA model of prepayment.

For default, the unnormalized chi-square density function and the more general unnormalized gamma density function are used as smooth approximations of the conventional SDA function. Since these two models are not nested within the class of polynomial models, the Akaike Information Criterion and the Schwartz Bayesian

[12] Specifying trends of default and prepayment *functions* is rather common in mortgage duration literature. See, for example, Deng et al. [45].

[13] Due to numerical problems, estimating a model with more than a seventh-order polynomial seems to be impossible.

Criterion are employed for model selection. As Panel A of Table 5.3 indicates, the chi-square model is preferred according to both AIC and SBC.

For prepayment, I specify the unnormalized exponential C.D.F. as a smooth approximation of the PSA function. The results reported in Panel B of Table 5.3 indicate that this specification is rejected in favor of polynomial models, i.e. the model with the sixth-degree polynomial trend remains the preferred one.

The results of particle filter estimation are presented in Tables 5.1 & 5.2.[14] The reported results include coefficient estimates and corresponding standard errors. In this discussion, I will concentrate only on the estimates of the preferred models: the chi-square model of default and the sixth-degree polynomial model of prepayment.

Important characteristics of a CIR model include the long-run mean $\theta(t)$ and volatility $\sigma$. The half-life of a process is defined as the amount of time it takes to reach a value halfway between the current level and the long-run mean, given by $\ln 2/\kappa$. The interpretation of the mean reversion parameter is as follows: if $\kappa = 0$, the process is a random walk, if $\kappa = 1$, it has has a unit root, and finally, if $\kappa < 0$, the process is diverging from its long-run mean.

In the chi-square default model, the combination of the scaling parameter $\alpha_0$ and the chi-square parameter $\rho_1$ gives a general idea about the long-term mean of the process $\theta(t)_d$. As seen in *Panel (j)* of Figure 5.2, the hazard of default rises approximately until the end of the fifth year, and then trails off smoothly, asymptotically reaching zero by the time the loan reaches maturity. This is very similar to the behavior of the conventional SDA function. The half-life of the process is 0.5 years, representing a strong degree of mean reversion. In *Panel (j)*, the trend of the

[14] Note that it is relatively hard to interpret the results for the mean and the volatility parameters, since they represent scaled baseline hazards, i.e. the monthly chance of default or prepayment given that all covariates are set to be equal to zero, after an additional adjustment by the scaling factor.

process is superimposed on the cloud of stratified proportional hazard estimates.[15] The fit seems to be adequate. Figure 5.1 and the rest of the panels of Figure 5.2 graphically present trends of the other default models.

In the sixth-degree polynomial prepayment model, as *Panel (j)* of Figure 5.4 indicates, the combination of the trend parameters produces a time-varying mean closely resembling the conventional PSA curve in earlier years, rising for the first few years of the loan life, and staying relatively stable after that. However, in the final years of the mortgage, the prepayment rate increases dramatically, in a manner earlier reported by Jegadeesh and Ju [83], since after about 18 years of payments, mortgage becomes a nuisance for the borrower, and the borrower prepays the remaining balance. The fit of the other models of prepayment is presented in Figure 5.3 and the rest of Table 5.4. The estimate of $\kappa$ implies a slightly lower degree of mean reversion for prepayment, indicating the half-life of 0.8 years.

Overall, the results seem satisfactory. In the default model, all coefficient estimates are statistically significant at the conventional levels, while in the prepayment model, the estimates of the fifth and the sixth polynomial terms are not significant, due to obvious multicollinearity problems. However, the assumption of independence on which these standard errors are based is not necessarily true. There are potential autocorrelation problems, both in mortgage time and in origination time.[16] I perform standard heteroscedasticity-robust tests for correlation in both dimensions, and find

[15] Stratified proportional hazard estimates are obtained using the formula displayed under Tables 4.7 and 4.9, without aggregation across $J$ strata.

[16] There is also the issue of correlation in calendar time; however, the calendar time of each observation is hard to pin down, since the data are stratified quarterly, not monthly. Approximate time of each observation can be obtained if one assumes that all mortgages in the same stratum are originated in any given month of that stratum. For example, results displayed in the first line of Table 5.5 (for both default and prepayment) are based on the assumption that all loans are originated in the second month of their stratum. However, since this involved approximations, correlation in calendar time will be ignored.

significant evidence that such correlation indeed exists. The results of these tests are shown in Table 5.5.

To correct for this two-dimensional correlation problem, I construct a two-dimensional Newey-West covariance matrix, as in [34]. The corrected standard errors, based on the Newey-West covariance matrix with 12 lags, are also reported in Tables 5.1 & 5.2 for the two preferred models.

Residual analysis was also performed on the model. Figure 5.6 displays the conventional Pearson residuals versus the actual counts, versus the predicted counts, and versus the normal score. For comparison, the Anscombe residuals are also graphed versus the normal score. Overall, residual analysis does not seem to indicate any major misspecification problems.

Additional, slightly more rigorous analysis of the overall fit of the model can be performed by studying the actual versus predicted probabilities for different count cells. I specify 12 count cells for each termination process. For default, each cell except the last contains one value, from 0 to 10. For prepayment, since it is a much more frequent event, each cell contains 10 values, except the first one and the last one. The results of the tests are reported in Panel A of Table 5.4 and Panel A of table 5.5. Overall, the results seem to be very plausible, particularly for the default model. The same results are presented graphically in Figure 5.5.

The table also reports the Pearson statistic for each of the models. If model specification is exactly right, the statistic should be equal to the number of observations less the number of estimated coefficients. In both models, the statistic indicates a slight degree of underdispersion in the data compared to the model prediction, especially in the prepayment model.

The second panel of Tables 5.4 and 5.5 presents the results of Conditional Moment (CM) tests of the overall model fit, as described in Cameron and Trivedi [22]. According on the CM test, which makes use of the Newey-West covariance matrix, the default model cannot be rejected at any conventional significance level, while the prepayment model is definitely rejected.

Finally, Figure 5.7 presents the one-step-ahead likelihood function over 5-year calendar time periods. For default, the likelihood function is relatively stable over the entire period. The prepayment model is apparently not as good in fitting the entire period; in particular, there are several observations in the late nineties, as well as 2000 and 2001, for which the model exhibits a particularly bad fit, most likely, because of exceptionally low interest rates during that period.[17]

[17] This can also be an indication of the misspecification of the effect of interest rates on prepayment at the proportional hazard stage of the analysis.

**TABLE 5.1**

**Particle Filter Estimates for the Default Diffusion Process**

Sample Size: N=12,111

Number of Particles: M=1,000 per stratum

Number of Particles: R=1,300 per stratum

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | Constant | First-Order | Second-Order |
|---|---:|---:|---:|
| | Model | Polynomial | Polynomial |
| | | Model | Model |
| **Variable** | **Estimate** | **Estimate** | **Estimate** |
| | **(Std. Error)** | **(Std. Error)** | **(Std. Error)** |
| $\kappa$ | 0.208 | 0.224 | 0.262 |
| | (0.056) | (0.050) | (0.060) |
| $\alpha_0$ | 1.848 | 3.294 | 2.484 |
| | (0.419) | (0.561) | (0.563) |
| $\alpha_1$ | | -0.121 | 0.0023 |
| | | (0.025) | (0.032) |
| $\alpha_2$ | | | $-4.857 \times 10^{-3}$ |
| | | | $(8.347 \times 10^{-4})$ |
| $\sigma$ | 1.529 | 1.482 | 1.556 |
| | (0.074) | (0.067) | (0.079) |
| *Log-Likelihood* | -5,412.14 | -5,394.35 | -5,390.58 |

I always let $\kappa$ represent the CIR reversion parameter and $\sigma$ the volatility parameter. Other coefficients concern the assumed shapes of the trend term $\theta_d(t)$. The $\alpha_i$ are coefficients of polynomials in the obvious sense.

**TABLE 5.1**

**Particle Filter Estimates for the Default Diffusion Process**

Sample Size: N=12,111

Number of Particles: M=1,000 per stratum

Number of Particles: R=1,300 per stratum

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | Third-Order Polynomial Model | Fourth-Order Polynomial Model | Fifth-Order Polynomial Model |
|---|---:|---:|---:|
| **Variable** | **Estimate** | **Estimate** | **Estimate** |
| | **(Std. Error)** | **(Std. Error)** | **(Std. Error)** |
| $\kappa$ | 0.375 | 0.365 | 0.377 |
| | (0.067) | (0.068) | (0.067) |
| $\alpha_0$ | 1.579 | 1.557 | 1.697 |
| | (0.360) | (0.350) | (0.289) |
| $\alpha_1$ | 0.244 | 0.256 | 0.229 |
| | (0.063) | (0.022) | (0.014) |
| $\alpha_2$ | -0.024 | -0.024 | -0.024 |
| | $(3.431 \times 10^{-3})$ | $(4.375 \times 10^{-4})$ | $(4.826 \times 10^{-4})$ |
| $\alpha_3$ | $4.620 \times 10^{-4}$ | $4.462 \times 10^{-4}$ | $4.624 \times 10^{-4}$ |
| | $(4.811 \times 10^{-5})$ | $(2.192 \times 10^{-5})$ | $(6.296 \times 10^{-6})$ |
| $\alpha_4$ | | $3.167 \times 10^{-7}$ | $7.774 \times 10^{-7}$ |
| | | $(1.168 \times 10^{-6})$ | $(1.982 \times 10^{-6})$ |
| $\alpha_5$ | | | $-4.991 \times 10^{-9}$ |
| | | | $(8.088 \times 10^{-8})$ |
| $\sigma$ | 1.558 | 1.566 | 1.591 |
| | (0.081) | (0.082) | (0.082) |
| *Log-Likelihood* | -5,383.92 | -5,383.72 | -5,383.05 |

I always let $\kappa$ represent the CIR reversion parameter and $\sigma$ the volatility parameter. Other coefficients concern the assumed shapes of the trend term $\theta_d(t)$. The $\alpha_i$ are coefficients of polynomials in the obvious sense.

**TABLE 5.1**

**Particle Filter Estimates for the Default Diffusion Process**

Sample Size: N=12,111

Number of Particles: M=1,000 per stratum

Number of Particles: R=1,300 per stratum

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | Sixth-Order Polynomial Model | Seventh-Order Polynomial Model |
|---|---|---|
| **Variable** | **Estimate** | **Estimate** |
| | **(Std. Error)** | **(Std. Error)** |
| $\kappa$ | 0.377 | 0.377 |
| | (0.070) | (0.070) |
| $\alpha_0$ | 1.697 | 1.697 |
| | (0.358) | (0.358) |
| $\alpha_1$ | 0.230 | 0.230 |
| | (0.032) | (0.032) |
| $\alpha_2$ | -0.024 | -0.024 |
| | $(1.458 \times 10^{-3})$ | $(1.458 \times 10^{-3})$ |
| $\alpha_3$ | $4.624 \times 10^{-4}$ | $4.624 \times 10^{-4}$ |
| | $(2.201 \times 10^{-5})$ | $(2.201 \times 10^{-5})$ |
| $\alpha_4$ | $7.729 \times 10^{-7}$ | $7.729 \times 10^{-7}$ |
| | $(2.297 \times 10^{-6})$ | $(1.982 \times 10^{-6})$ |
| $\alpha_5$ | $-4.766 \times 10^{-9}$ | $-4.766 \times 10^{-9}$ |
| | $(1.816 \times 10^{-7})$ | $(1.816 \times 10^{-7})$ |
| $\alpha_6$ | $-3.056 \times 10^{-11}$ | $-3.056 \times 10^{-11}$ |
| | $(8.861 \times 10^{-9})$ | $(8.861 \times 10^{-9})$ |
| $\alpha_7$ | | 0.000 |
| | | (0.000) |
| $\sigma$ | 1.590 | 1.590 |
| | (0.084) | (0.084) |
| *Log-Likelihood* | -5,381.79 | -5,381.79 |

I always let $\kappa$ represent the CIR reversion parameter and $\sigma$ the volatility parameter. Other coefficients concern the assumed shapes of the trend term $\theta_d(t)$. The $\alpha_i$ are coefficients of polynomials in the obvious sense.

**TABLE 5.1**

**Particle Filter Estimates for the Default Diffusion Process**

Sample Size: N=12,111

Number of Particles: M=1,000 per stratum

Number of Particles: R=1,300 per stratum

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | Chi-Square<br>Model | Gamma<br>Model |
|---|---|---|
| **Variable** | **Estimate**<br>**(Std. Error)**<br>**(Corrected**<br>**Std. Error)** | **Estimate**<br>**(Std. Error)** |
| $\kappa$ | **1.382** | 1.496 |
| | **(0.168)** | (0.176) |
| | **(0.273)** | |
| $\alpha_0$ | **38.780** | 38.788 |
| | **(1.839)** | (1.756) |
| | **(2.283)** | |
| $\rho_1$ | **3.587** | 3.795 |
| | **(0.073)** | (0.186) |
| | **(0.117)** | |
| $\rho_2$ | | 1.870 |
| | | (0.105) |
| $\sigma$ | **1.939** | 1.989 |
| | **(0.113)** | (0.114) |
| | **(0.193)** | |
| *Log-Likelihood* | -5,299.34 | -5,298.96 |

I always let $\kappa$ represent the CIR reversion parameter and $\sigma$ the volatility parameter. Other coefficients concern the assumed shapes of the trend term $\theta_d(t)$. I let $\alpha_d$ represent a factor unnormalizing either the chi-square or gamma P.D.F.s, while $2\rho_1$ is the single parameter of a chi-square P.D.F., and hence, $\rho_1$ is one of the parameters of the gamma P.D.F., while $\rho_2$ is the second one. The first standard error of the preferred, chi-square model is the standard one; the second is a double Newey-West correction.

**TABLE 5.2**

**Particle Filter Estimates for the Prepayment Diffusion Process**

Sample Size: N=11,985

Number of Particles: M=1,000 per stratum

Number of Particles: R=1,300 per stratum

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | Constant | First-Order | Second-Order |
|---|---|---|---|
| | Model | Polynomial | Polynomial |
| | | Model | Model |
| **Variable** | **Estimate** | **Estimate** | **Estimate** |
| | **(Std. Error)** | **(Std. Error)** | **(Std. Error)** |
| $\kappa$ | 0.377 | 0.599 | 0.644 |
| | (0.039) | (0.041) | (0.054) |
| $\alpha_0$ | 4.598 | 0.656 | 2.369 |
| | (0.371) | (0.037) | (0.155) |
| $\alpha_1$ | | 0.273 | 0.120 |
| | | (0.021) | (0.061) |
| $\alpha_2$ | | | $2.328 \times 10^{-3}$ |
| | | | $(2.841 \times 10^{-3})$ |
| $\sigma$ | 2.352 | 2.363 | 2.374 |
| | (0.007) | (0.007) | (0.007) |
| *Log-Likelihood* | -24,956.89 | -24,920.53 | -24,912.47 |

I always let $\kappa$ represent the CIR reversion parameter and $\sigma$ the volatility parameter. Other coefficients concern the assumed shapes of the trend term $\theta_p(t)$. The $\alpha_i$ are coefficients of polynomials in the obvious sense.

**TABLE 5.2**

**Particle Filter Estimates for the Prepayment Diffusion Process**

Sample Size: N=11,985

Number of Particles: M=1,000 per stratum

Number of Particles: R=1,300 per stratum

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | Third-Order | Fourth-Order | Fifth-Order |
|---|---|---|---|
| | Polynomial | Polynomial | Polynomial |
| | Model | Model | Model |
| **Variable** | **Estimate** | **Estimate** | **Estimate** |
| | **(Std. Error)** | **(Std. Error)** | **(Std. Error)** |
| $\kappa$ | 0.830 | 0.891 | 0.891 |
| | (0.058) | (0.069) | (0.062) |
| $\alpha_0$ | 1.235 | 1.311 | 1.033 |
| | (0.070) | (0.101) | (0.058) |
| $\alpha_1$ | 0.968 | 0.898 | 1.460 |
| | (0.119) | (0.208) | (0.177) |
| $\alpha_2$ | -0.092 | -0.083 | -0.251 |
| | (1.270) | (0.037) | (0.036) |
| $\alpha_3$ | $2.479 \times 10^{-3}$ | $2.111 \times 10^{-3}$ | 0.019 |
| | $(3.421 \times 10^{-4})$ | $(2.223 \times 10^{-3})$ | $(2.769 \times 10^{-3})$ |
| $\alpha_4$ | | $4.731 \times 10^{-6}$ | $-7.011 \times 10^{-4}$ |
| | | $(4.234 \times 10^{-5})$ | $(1.021 \times 10^{-4})$ |
| $\alpha_5$ | | | $1.021 \times 10^{-5}$ |
| | | | $(1.563 \times 10^{-6})$ |
| $\sigma$ | 2.364 | 2.393 | 2.397 |
| | (0.007) | (0.008) | (0.008) |
| *Log-Likelihood* | -24,882.59 | -24,881.91 | -24,878.75 |

WI always let $\kappa$ represent the CIR reversion parameter and $\sigma$ the volatility parameter. Other coefficients concern the assumed shapes of the trend term $\theta_p(t)$. The $\alpha_i$ are coefficients of polynomials in the obvious sense.

## TABLE 5.2

**Particle Filter Estimates for the Prepayment Diffusion Process**

Sample Size: N=11,985

Number of Particles: M=1,000 per stratum

Number of Particles: R=1,300 per stratum

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | Sixth-Order | Seventh-Order |
|---|---:|---:|
| | Polynomial | Polynomial |
| | Model | Model |
| **Variable** | **Estimate** | **Estimate** |
| | **(Std. Error)** | **(Std. Error)** |
| | **(Corrected** | |
| | **Std. Error)** | |
| $\kappa$ | **0.871** | 0.907 |
| | **(0.057)** | (0.059) |
| | **(0.089)** | |
| $\alpha_0$ | **0.388** | 0.198 |
| | **(0.101)** | (0.091) |
| | **(0.186)** | |
| $\alpha_1$ | **1.720** | 1.573 |
| | **(0.213)** | (0.191) |
| | **(0.388)** | |
| $\alpha_2$ | **-0.281** | -0.214 |
| | **(0.042)** | (0.040) |
| | **(0.064)** | |
| $\alpha_3$ | **0.019** | 0.012 |
| | $\mathbf{(3.213 \times 10^{-3})}$ | $(3.322 \times 10^{-3})$ |
| | $\mathbf{(3.878 \times 10^{-3})}$ | |

*Continued on the next page*

**TABLE 5.2**

**Particle Filter Estimates for the Prepayment Diffusion Process**

Sample Size: N=11,985

Number of Particles: M=1,000 per stratum

Number of Particles: R=1,300 per stratum

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | Sixth-Order | Seventh-Order |
|---|---:|---:|
| | Polynomial | Polynomial |
| | Model | Model |
| **Variable** | **Estimate** | **Estimate** |
| | **(Std. Error)** | **(Std. Error)** |
| | **(Corrected** | |
| | **Std. Error)** | |
| $\alpha_4$ | $\boldsymbol{-6.151 \times 10^{-4}}$ | $-2.344 \times 10^{-4}$ |
| | $\boldsymbol{(8.754 \times 10^{-5})}$ | $(1.156 \times 10^{-4})$ |
| | $\boldsymbol{(8.970 \times 10^{-5})}$ | |
| $\alpha_5$ | $\boldsymbol{4.720 \times 10^{-6}}$ | $-2.210 \times 10^{-6}$ |
| | $\boldsymbol{(2.876 \times 10^{-6})}$ | $(2.601 \times 10^{-6})$ |
| | $\boldsymbol{(4.388 \times 10^{-6})}$ | |
| $\alpha_6$ | $\boldsymbol{9.354 \times 10^{-8}}$ | $1.146 \times 10^{-7}$ |
| | $\boldsymbol{(1.697 \times 10^{-8})}$ | $(5.680 \times 10^{-8})$ |
| | $\boldsymbol{(1.109 \times 10^{-7})}$ | |
| $\alpha_7$ | | $1.215 \times 10^{-10}$ |
| | | $(5.036 \times 10^{-12})$ |
| $\sigma$ | **2.297** | 2.335 |
| | **(0.006)** | (0.008) |
| | **(0.011)** | |
| *Log-Likelihood* | -24,832.17 | -24,831.42 |

I always let $\kappa$ represent the CIR reversion parameter and $\sigma$ the volatility parameter. Other coefficients concern the assumed shapes of the trend term $\theta_p(t)$. The $\alpha_i$ are coefficients of polynomials in the obvious sense. The first standard error of the preferred, sixth-degree polynomial model is the standard one; the second is a double Newey-West correction.

**TABLE 5.2**

**Particle Filter Estimates for the Prepayment Diffusion Process**

Sample Size: N=11,985

Number of Particles: M=1,000 per stratum

Number of Particles: R=1,300 per stratum

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | Exponential<br>Model |
|---|---:|
| **Variable** | **Estimate**<br>**(Std. Error)** |
| $\kappa$ | 0.755 |
| | (0.050) |
| $\alpha_0$ | 4.359 |
| | (0.254) |
| $\rho_1$ | 1.518 |
| | (0.131) |
| $\sigma$ | 2.349 |
| | (0.007) |
| *Log-Likelihood* | -24,895.43 |

I always let $\kappa$ represent the CIR reversion parameter and $\sigma$ the volatility parameter. Other coefficients concern the assumed shapes of the trend term $\theta_p(t)$. I let $\alpha_p$ represent a factor unnormalizing the exponential C.D.F., while $2\rho_1$ is the single parameter of the exponential C.D.F.

## TABLE 5.3

### Panel A
### Model Selection Tests:
### Default Model

Sample Size: N=12,111

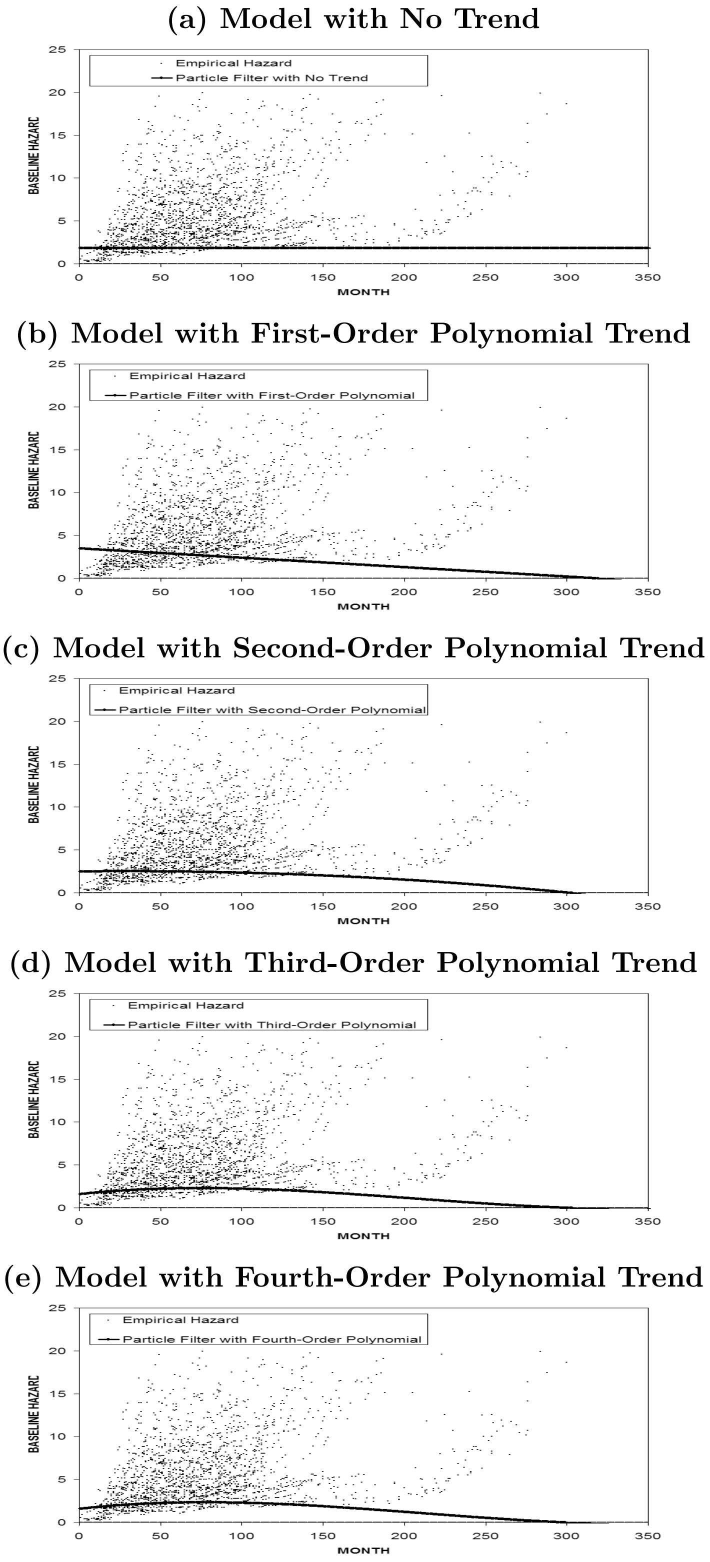
| Trend | AIC | SBC |
|---|---:|---:|
| None | 10,830.3 | 10,852.5 |
| 1-Order Polynomial | 10,796.7 | 10,826.3 |
| 2-Order Polynomial | 10,791.2 | 10,828.2 |
| 3-Order Polynomial | 10779.8 | 10,824.3 |
| 4-Order Polynomial | 10,781.4 | 10,833.3 |
| 5-Order Polynomial | 10,784.1 | 10,843.3 |
| 6-Order Polynomial | 10,781.6 | 10,848.2 |
| 7-Order Polynomial | 10,783.6 | 10,857.6 |
| Chi-Square | 10,606.7 | 10,636.3 |
| Gamma | 10,607.9 | 10,644.9 |

### Panel B
### Model Selection Tests:
### Prepayment Model

Sample Size: N=11,985

| Trend | AIC | SBC |
|---|---:|---:|
| None | 49,919.7 | 49,941.9 |
| 1-Order Polynomial | 49,849.1 | 49,878.6 |
| 2-Order Polynomial | 49,834.9 | 49,871.9 |
| 3-Order Polynomial | 49,777.2 | 49,821.5 |
| 4-Order Polynomial | 49,777.8 | 49,829.6 |
| 5-Order Polynomial | 49,773.5 | 49,832.6 |
| 6-Order Polynomial | 49,682.3 | 49,748.9 |
| 7-Order Polynomial | 49,682.8 | 49,756.7 |
| Exponential | 49,798.8 | 49,828.4 |

Figure 5.1: **The Trend of the Particle Filter Estimate of the Square-Root Baseline Default Process vs. Stratified Empirical Proportional Hazard Estimates**



**(a) Model with No Trend**

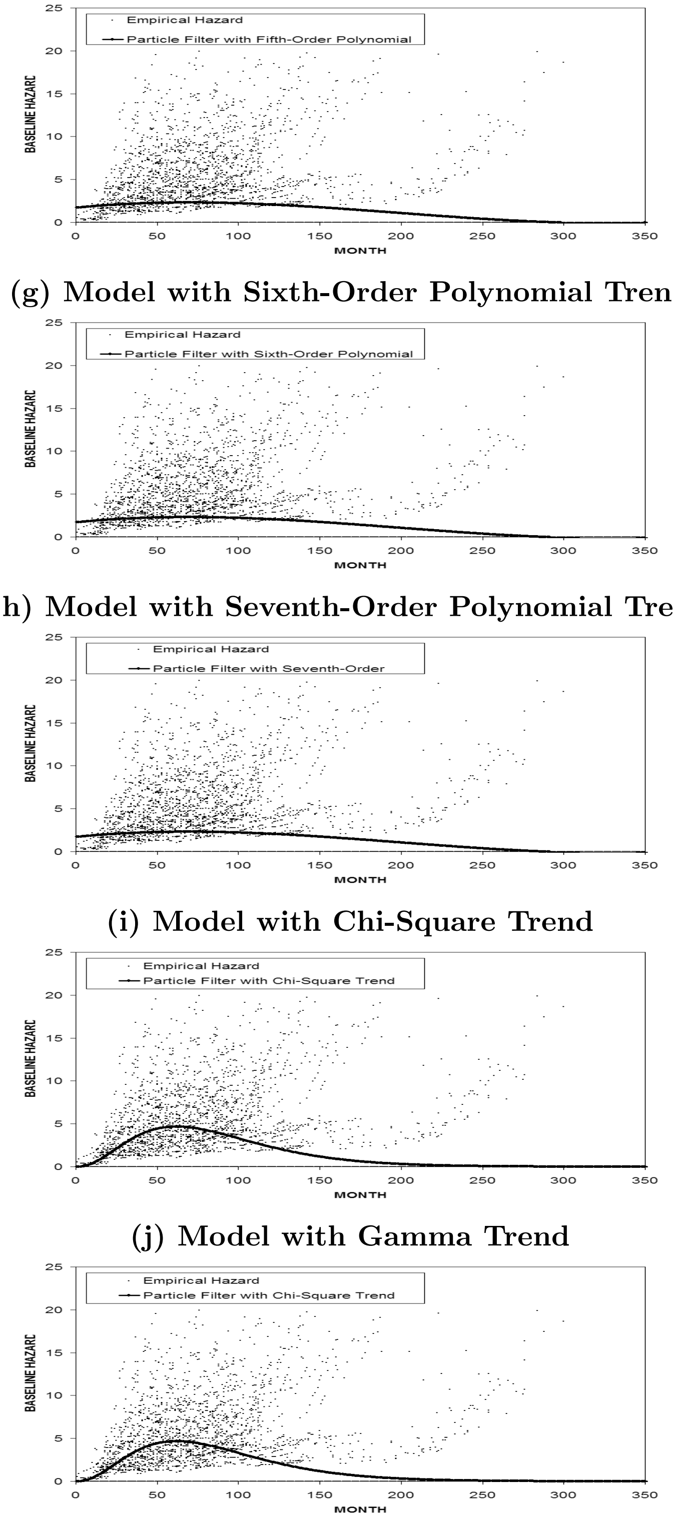
**(b) Model with First-Order Polynomial Trend**

**(c) Model with Second-Order Polynomial Trend**

**(d) Model with Third-Order Polynomial Trend**

**(e) Model with Fourth-Order Polynomial Trend**

Figure 5.2: **The Trend of the Particle Filter Estimate of the Square-Root Baseline Default Process vs. Stratified Empirical Proportional Hazard Estimates**



**(f) Model with Fifth-Order Polynomial Trend**

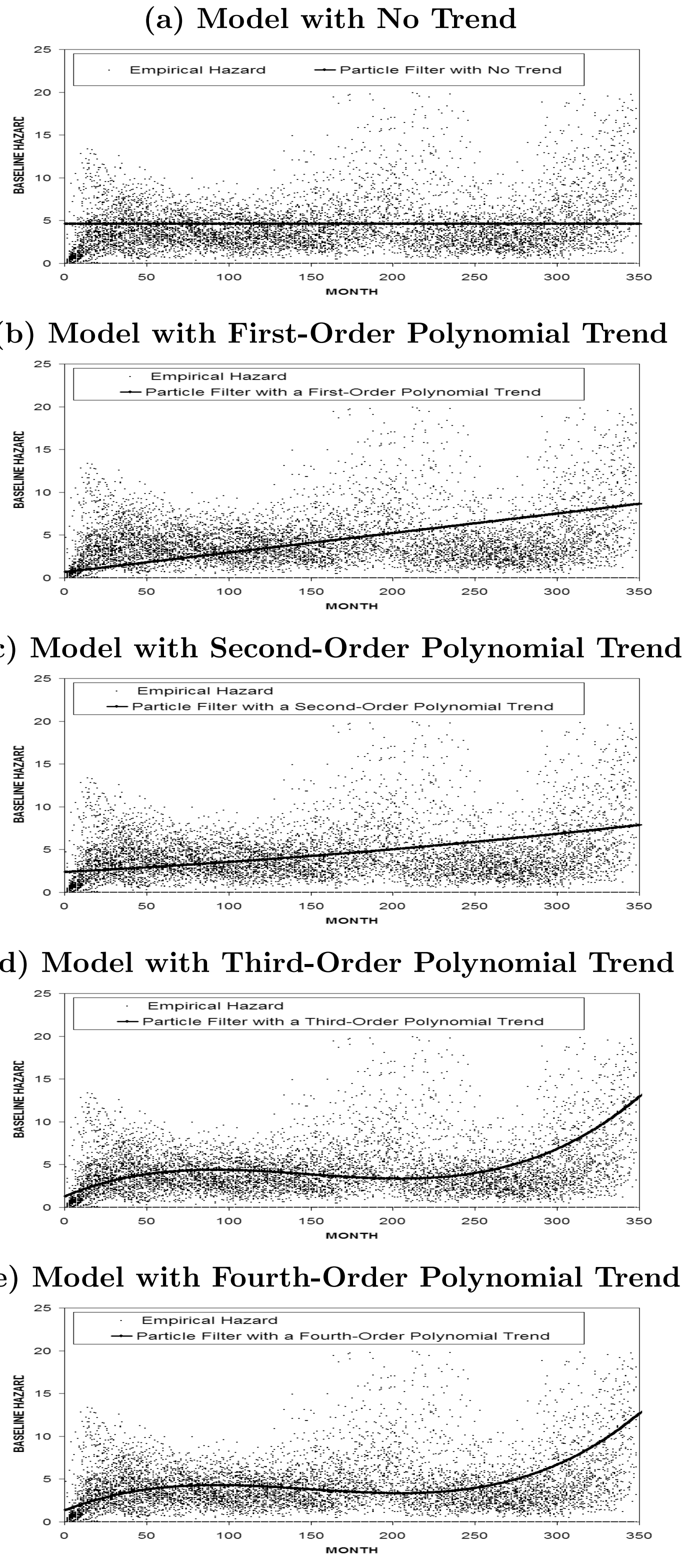
**(g) Model with Sixth-Order Polynomial Trend**

**(h) Model with Seventh-Order Polynomial Trend**

**(i) Model with Chi-Square Trend**

**(j) Model with Gamma Trend**

Figure 5.3: **The Trend of the Particle Filter Estimate of the Square-Root Baseline Prepayment Process vs. Stratified Empirical Proportional Hazard Estimates**



(a) Model with No Trend

(b) Model with First-Order Polynomial Trend

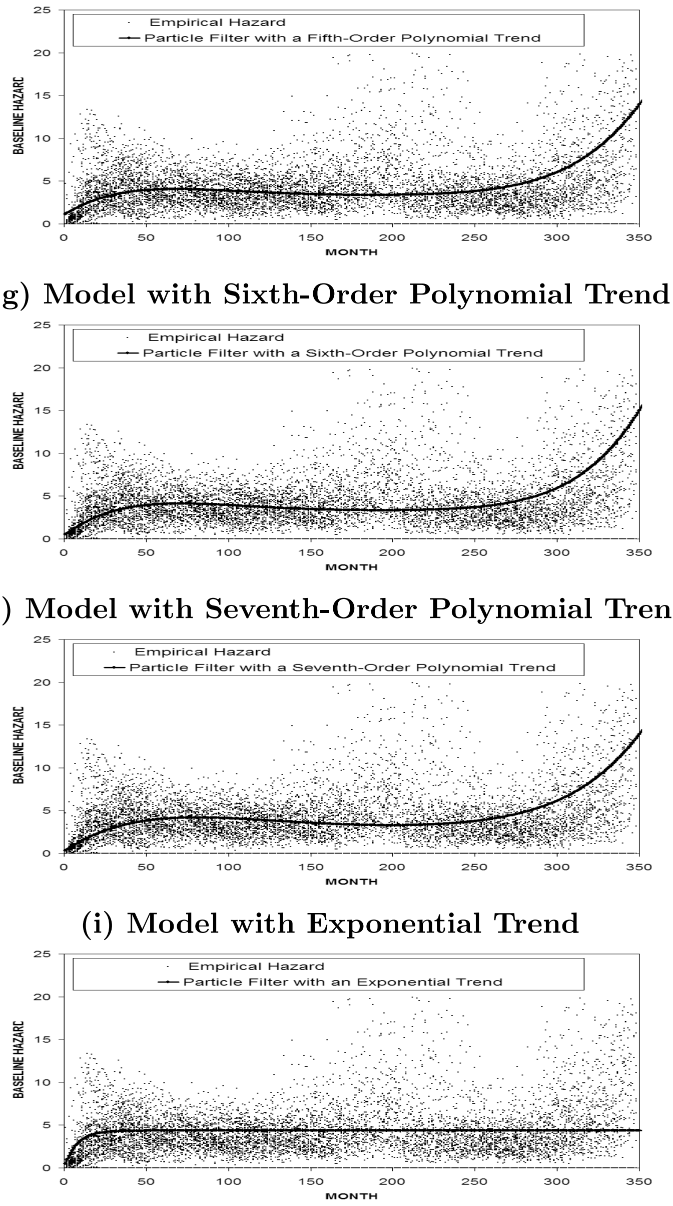
(c) Model with Second-Order Polynomial Trend

(d) Model with Third-Order Polynomial Trend

(e) Model with Fourth-Order Polynomial Trend

Figure 5.4: **The Trend of the Particle Filter Estimate of the Square-Root Baseline Prepayment Process vs. Stratified Empirical Proportional Hazard Estimates**

**(f) Model with Fifth-Order Polynomial Trend**



**(g) Model with Sixth-Order Polynomial Trend**

**(h) Model with Seventh-Order Polynomial Trend**

**(i) Model with Exponential Trend**

**TABLE 5.4**

**Tests of Model Predictions**

**Default Model**

**Panel A**

**Predicted vs. Actual Probabilities**

| **Counts** | **Actual** | **Predicted** | **\|Diff.\|** | **Pearson** |
|---|---:|---:|---:|---:|
| 0 | 0.8264 | 0.8271 | 0.0008 | 0.0092 |
| 1 | 0.0860 | 0.0891 | 0.0031 | 1.3039 |
| 2 | 0.0421 | 0.0383 | 0.0038 | 4.6479 |
| 3 | 0.0192 | 0.0198 | 0.0006 | 0.2073 |
| 4 | 0.0118 | 0.0108 | 0.0010 | 1.2410 |
| 5 | 0.0057 | 0.0059 | 0.0002 | 0.1254 |
| 6 | 0.0031 | 0.0033 | 0.0002 | 0.1570 |
| 7 | 0.0013 | 0.0019 | 0.0006 | 2.3576 |
| 8 | 0.0014 | 0.0012 | 0.0002 | 0.6086 |
| 9 | 0.0004 | 0.0007 | 0.0003 | 1.6736 |
| 10 | 0.0006 | 0.0005 | 0.0001 | 0.2325 |
| $> 10$ | 0.0019 | 0.0013 | 0.0006 | 3.9549 |
| Pearson Statistic: *P* | | | | 10,115.53 |
| Degrees of Freedom: *N-K* | | | | 12,107.00 |

**Panel B**

**CM Tests**

Degrees of Freedom: Q = 12

| **Covariance Used** | **Test Statistic** | **P-Value** |
|---|---:|---:|
| OPG | 30.28 | 0.002 |
| Newey-West | 17.32 | 0.138 |

The predicted probability of a termination count is its combined probability over cohorts and months, using one-step particle-filter forecasts and the now estimated parameters. This prediction is then compared to the actual percent of occasions that terminations for some cohort and month was, indeed, of that count, with the difference yielding the unstandardized residual (|Diff|). The final column calculates each category's contribution to the total Pearson statistic $P$, which is then compared to the degrees of freedom $Q = N - K$, with a lower value of the statistic taken to be an indication of underdispersion. The second table presents the Conditional Moment (CM) tests of the overall model fit. Under the null hypothesis of an adequate fit, the test statistic has a chi-square distribution with the number of degrees of freedom equal to the number of count cells specified.

### TABLE 5.5

**Tests of Model Predictions**

**Prepayment Model**

**Panel A**

**Predicted vs. Actual Probabilities**

| **Counts** | **Actual** | **Predicted** | **\|Diff.\|** | **Pearson** |
|---|---|---|---|---|
| 0 | 0.3441 | 0.3319 | 0.0122 | 5.3556 |
| 1-10 | 0.4145 | 0.4219 | 0.0074 | 1.5568 |
| 11-20 | 0.0679 | 0.0727 | 0.0048 | 3.7343 |
| 21-30 | 0.0234 | 0.0286 | 0.0053 | 11.6077 |
| 31-40 | 0.0159 | 0.0191 | 0.0031 | 6.2293 |
| 41-50 | 0.0186 | 0.0154 | 0.0032 | 7.8740 |
| 51-60 | 0.0139 | 0.0129 | 0.0009 | 0.7869 |
| 61-70 | 0.0117 | 0.0109 | 0.0008 | 0.6217 |
| 71-80 | 0.0074 | 0.0093 | 0.0019 | 4.7016 |
| 81-90 | 0.0085 | 0.0081 | 0.0004 | 0.3008 |
| 91-100 | 0.0080 | 0.0070 | 0.0010 | 1.6987 |
| $> 100$ | 0.0661 | 0.0621 | 0.0040 | 3.0733 |
| Pearson Statistic: $P$ | | | | 8,214.61 |
| Degrees of Freedom: $N$-$K$ | | | | 11,976.00 |

**Panel B**

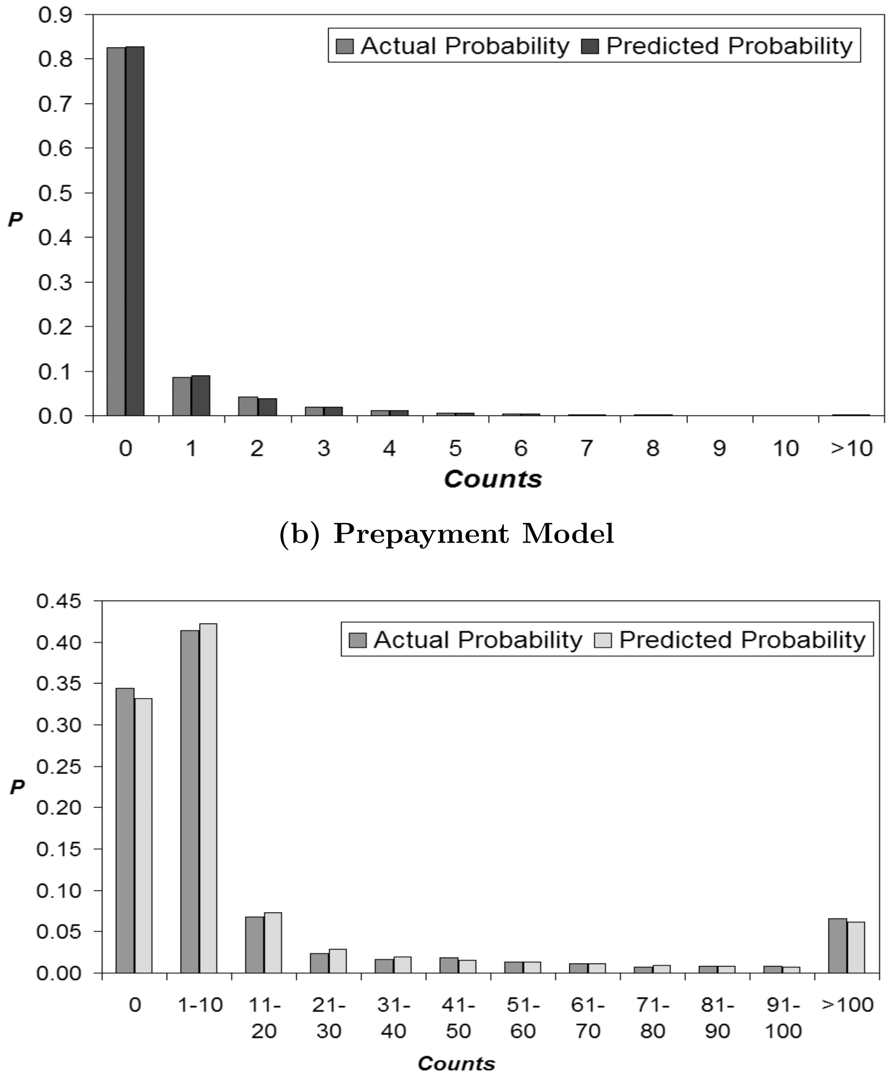
**CM Tests**

Degrees of Freedom: Q = 12

| **Covariance Used** | **Test Statistic** | **P-Value** |
|---|---|---|
| OPG | 157.00 | 0.000 |
| Newey-West | 43.39 | 0.000 |

The predicted probability of a termination count is its combined probability over cohorts and months, using one-step particle-filter forecasts and the now estimated parameters. This prediction is then compared to the actual percent of occasions that terminations for some cohort and month was, indeed, of that count, with the difference yielding the unstandardized residual (|Diff|). The final column calculates each category's contribution to the total Pearson statistic $P$, which is then compared to the degrees of freedom $Q = N - K$, with a lower value of the statistic taken to be an indication of underdispersion. The second table presents the Conditional Moment (CM) tests of the overall model fit. Under the null hypothesis of an adequate fit, the test statistic has a chi-square distribution with the number of degrees of freedom equal to the number of count cells specified.

Figure 5.5: **Predicted vs. Actual Probabilities**



**(a) Default Model**

**(b) Prepayment Model**

## TABLE 5.5

**Percentage of Series for which $H_0$ : *No Autocorrelation* is Rejected**

Default Model

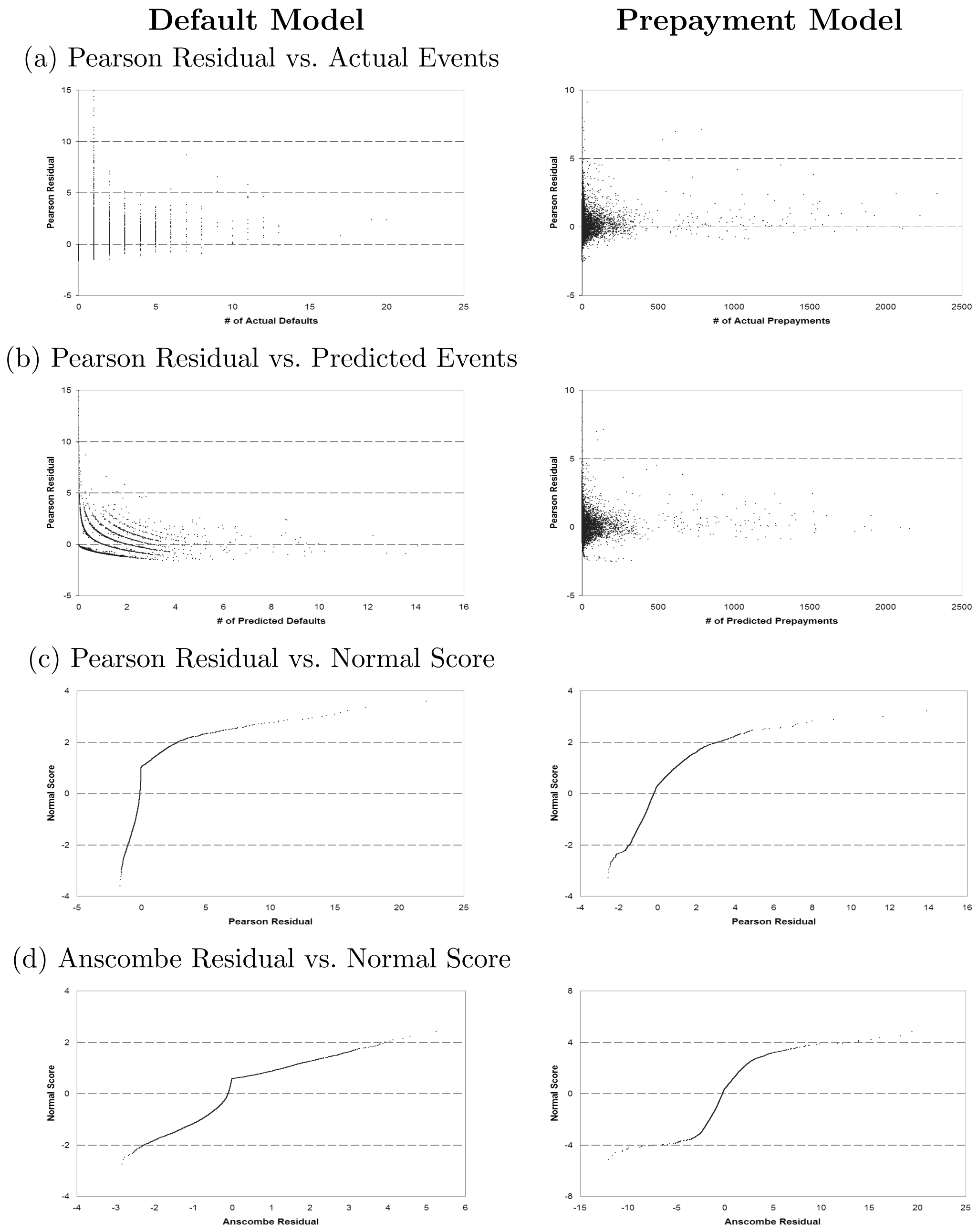
| | **6 Lags** | | | **12 Lags** | | |
|---|---|---|---|---|---|---|
| | $\alpha = 1\%$ | $\alpha = 5\%$ | $\alpha = 10\%$ | $\alpha = 1\%$ | $\alpha = 5\%$ | $\alpha = 10\%$ |
| Calendar Time | 28.6% | 46.8% | 51.9% | 55.8% | 70.1% | 80.5% |
| Mortgage Time | 46.8% | 59.7% | 66.2% | 68.8% | 74.0% | 75.3% |
| Origination Time | 44.3% | 56.8% | 65.9% | 69.5% | 75.0% | 77.2% |

Prepayment Model

| | **6 Lags** | | | **12 Lags** | | |
|---|---|---|---|---|---|---|
| | $\alpha = 1\%$ | $\alpha = 5\%$ | $\alpha = 10\%$ | $\alpha = 1\%$ | $\alpha = 5\%$ | $\alpha = 10\%$ |
| Calendar Time | 57.1% | 71.4% | 75.3% | 75.3% | 85.7% | 89.6% |
| Mortgage Time | 53.3% | 62.34% | 68.8% | 88.3% | 93.5% | 94.8% |
| Origination Time | 23.5% | 36.5% | 45.9% | 65.8% | 83.5% | 85.8% |

The calendar time sample and the origination time sample were restricted to include only series with at least 100 observations. The mortgage time sample was restricted to include only series with at least 38 observations.

Figure 5.6: **Residual Analysis**



**Default Model**

**Prepayment Model**

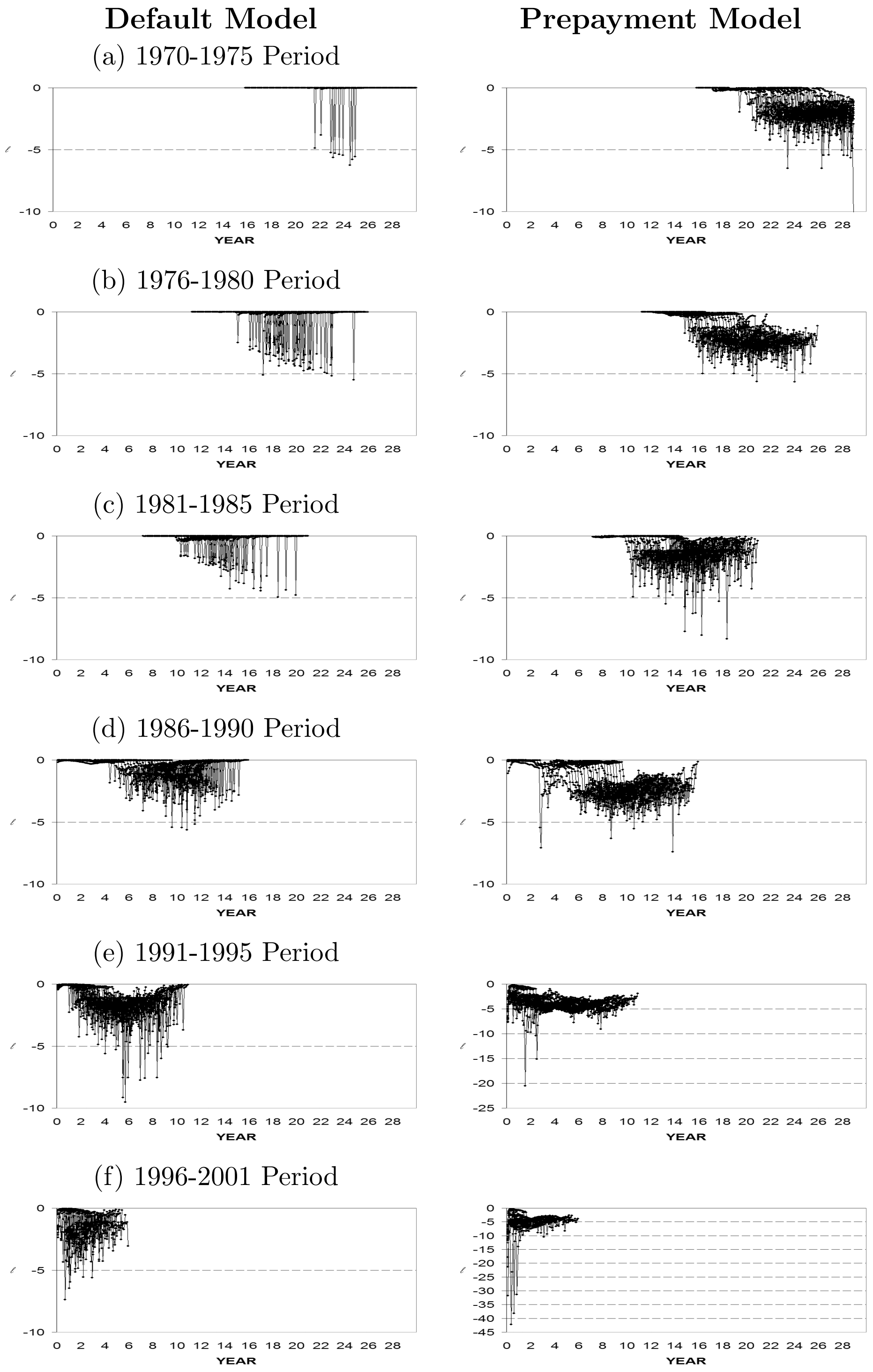
(a) Pearson Residual vs. Actual Events

(b) Pearson Residual vs. Predicted Events

(c) Pearson Residual vs. Normal Score

(d) Anscombe Residual vs. Normal Score

Figure 5.7: **One-Step Prediction Density for Different Time Periods**



**Default Model** **Prepayment Model**

(a) 1970-1975 Period

(b) 1976-1980 Period

(c) 1981-1985 Period

(d) 1986-1990 Period

(e) 1991-1995 Period

(f) 1996-2001 Period

# Chapter 6

## Estimation of Risk-Free Term Structure

I model term structure of risk-free interest rates as a multifactor square-root process of the Cox, Ingersoll and Ross (CIR) type [38]. The key advantage of this model is the presence of explicit analytical expressions for the equilibrium interest rate dynamics and bond prices, allowing me to estimate parameters of the model without resorting to any simulations. These close form formulae become a significant convenience when I perform the Monte Carlo pricing of mortgages.

The observed bond prices are treated as exact, i.e. observed without error. While the baseline hazards estimated at the previous stage, on the other hand, do have noisy observation equations, the meaning there is considerably different from what noise introduced into the term structure observation equations would represent. The noise in the baseline hazard observation equations arises from the assumed *double stochasticity* of the termination processes, and is then required by the model, despite the fact that it is assumed that the number of defaults and prepayments is observed without error. In contrast, any noise term introduced into the term structure observation equation would have to represent some sort of "measurement error." Although it may be useful for robustness of estimation,[1] such an error term would be totally ad hoc, not having been built into the model, but rather, having been tacked on in

[1] Introducing noise in the observation equation is usually motivated by the need to evaluate the entire yield curve via including more bond series in the estimated model. In my case, I am not particularly interested in pricing the entire range of different bond series, but rather in obtaining estimates of two interest rates: the short-term (instantaneous) rate used for discounting, and the 10-year rate used as a covariate in termination processes.

the estimation stage. Thus, making the term structure estimation like the hazard estimation requires that there be no noise.[2]

## 6.1 Existing Literature on Multi-Factor CIR Models

In this section, I provide a short discussion of relevant literature related to developing and estimating multi-factor Cox-Ingersoll-Ross models.[3]

One of the first multi-factor CIR models was developed by Pearson and Sun [111], who estimate and test a two-factor CIR model and its extended version using data on US Treasury bills, bonds and notes observed monthly during 1971-1986. Pearson and Sun develop the model by appending a price level process to the one-factor model of the real economy, and estimate it using exact maximum likelihood.

Chen and Scott [27] develop a similar model, with several major distinctions. First, they assume all factors in their model are real factors, meaning that inflation is not a factor. Second, they introduce Gaussian pricing/observation errors in the model, which allows them to include more bond series in estimation, improving the accuracy of bond pricing. For identification purposes, one of the bond series is assumed to be priced without an error. Using two data sets, one for 1960-97 and one for 1980-1988, Chen and Scott estimate models with up to three factors, however, without analyzing the “extended” form of the CIR model. After comparing the models, the authors conclude that more than one factor is necessary to explain the changes in the shape and the slope of the yield curve. For most cases, the two-factor

[2] Note, also, that, were this not an exact approach, then when one is pricing a mortgage, it would become difficult to determine out-of-sample initial values of the state variables describing the term structure, as is required for valuation.

[3] In the discussion of existing literature, I will concentrate on likelihood-based methods. Another popular approach is the simulated method of moments (SMM). See Dai and Singleton [41] for an application of SMM to a family of exponential affine term structure models, including CIR.

model is found to adequately capture the general shape of the yield curve,[4] with the third factor providing only a marginal improvement.

In their next paper, Chen and Scott [29][5] use a nonlinear Kalman filter to estimate CIR models with one, two and three factors. The use of a Kalman filter allows one to assume that *all* bond series used for estimation are observed with error, at the expense of a potential large sample bias in parameter estimates (due to approximating the non-central chi-square distribution of innovations as Gaussian). The general conclusions are similar to those of Chen and Scott [27]: the two-factor model is generally adequate, though there are periods during which the third factor becomes important.[6] Finally, Monte Carlo simulations are used to test the extent of large-sample biases. The results confirm that, even though there are significant biases in individual parameter estimates, the parameter combinations relevant for asset pricing do not seem to exhibit such biases.

Duan and Simonato [48] re-estimate the two-factor model of Chen and Scott [29] using the same Kalman filter method, but with different data, and compare its predictions to those the original CIR and the Vasicek model. Geyer and Pichler [66] perform a similar exercise, employing a state-space approach for estimation of a multifactor CIR model. Using data from the period 1964-93, the authors estimate models with up to five factors, and obtain estimates similar to those of Chen and Scott [29], but with a higher degree of statistical significance. The difference in statistical significance is due to a greater number of bond series used for estimation (sixteen vs. four).

[4] Two-factor CIR models without noise are also used by Chen and Scott in the context of option pricing. See [26][28] for examples.

[5] This paper has been at the stage of a working paper since 1992. It was published only in 2003 in the *Journal of Real Estate Finance and Economics*.

[6] The data are the same as the 1960-87 data set used in [27].

In reduced-form literature, the multi-factor model has been used by Duffie and Singleton [55] and Duffee [49]. Both papers estimate the “extended” term structure model with two factors, though the data and the estimation method are significantly different. Duffie and Singleton estimate the defaultable term structure directly, under the RMV assumption, using exact likelihood with noisy observations, as in [27]. Duffee employs the extended Kalman filter to estimate the risk-free term structure, later used for estimation of the credit risk spread parameters.

## 6.2 The Two-Factor Independent CIR Model

In the basic multi-factor CIR model, the instantaneous risk-free interest rate is expressed as the sum of $K$ independent state variables:

$$r(t) = \sum_{\imath=1}^{K} y_\imath(t)$$

The evolution of the state variables under the physical measure $\mathbb{P}$ is described by the stochastic differential equation of the following form:

$$dy_\imath(t) = \kappa_\imath(\theta_\imath - y_\imath(t))dt + \sigma_\imath\sqrt{y_\imath(t)}dz_\imath^{\mathbb{P}} \tag{6.1}$$

In the absence of arbitrage, this is equivalent to the following SDE under the risk-neutral measure $\mathbb{Q}$:

$$dy_\imath(t) = (\kappa_\imath(\theta_\imath - y_\imath(t)) - \nu_\imath y_\imath(t))dt + \sigma_\imath\sqrt{y_\imath(t)}dz_\imath^{\mathbb{Q}} \tag{6.2}$$

For each $\imath$, $\kappa_\imath$ is the speed of mean reversion, $\theta_\imath$ is the long-run mean, $\sigma_\imath$ is the volatility parameter, $\nu_\imath$ is the risk premium, and $dz_\imath^{\mathbb{P}}$ and $dz_\imath^{\mathbb{Q}}$ are independent standard Brownian motions under $\mathbb{P}$ and $\mathbb{Q}$, respectively. The presence of $\sqrt{y_\imath}$ in the volatility term precludes the factors from having negative values.

The solution for the nominal price at time $t$ of a risk-free bond that pays one dollar at time $s$ is determined by

$$N(t,s) = \prod_{k=1}^{K} A_k(t,s)\mathbf{e}^{-\sum_{j=1}^{K} B_k y(s,t)}$$

where $A_\imath(t,s)$ and $B_\imath(t,s)$ have the form

$$A_\imath(t,s) = \left[\frac{2\gamma_\imath \mathbf{e}^{\frac{1}{2}(\kappa_\imath+\nu_\imath-\gamma_\imath)(s-t)}}{2\gamma_\imath \mathbf{e}^{-\gamma_\imath(s-t)} + (\kappa_\imath+\nu_\imath+\gamma_\imath)(1-\mathbf{e}^{-\gamma_\imath(s-t)})}\right]^{\frac{2\kappa_\imath\theta_\imath}{\sigma_\imath^2}} \tag{6.3}$$

$$B_\imath(t,s) = \frac{2(1-\mathbf{e}^{-\gamma_\imath(s-t)})}{2\gamma_\imath \mathbf{e}^{-\gamma_\imath(s-t)} + (\kappa_\imath+\nu_\imath+\gamma_\imath)(1-\mathbf{e}^{-\gamma_\imath(s-t)})} \tag{6.4}$$

where:

$$\gamma_\imath = \sqrt{(\kappa_\imath+\nu_\imath)^2 + 2\sigma_\imath^2}$$

Finally, the continuously compounded yield for a discount bond is defined as:

$$R(t,s) = -\frac{\ln N(t,s)}{s-t} \tag{6.5}$$

In the extended model, the instantaneous interest rate is just the sum of the factors and a *constant*, which yields:

$$r(t) = \sum_{\imath=1}^{K} y^\imath(t) + \bar{y} \tag{6.6}$$

Empirically, including the constant mitigates the problem of negative state values, which are not possible under the square-root specification of diffusions.

Given the extended specification, a price of a zero coupon bond with \$1 paid at maturity is simply:

$$N(t,s) = \prod_{k=1}^{K} A_k(t,s)\mathbf{e}^{-\sum_{j=1}^{K} B_k y(s,t) - \bar{y}s} \tag{6.7}$$

The joint density of factors $y_\imath(t)$ conditional on $y_\imath(t')$ , for $t > t'$, is

$$f(y_\imath(t)|y_\imath(t')) = \prod_{\imath=1}^{K} d_\imath(t,t')\mathbf{e}^{-u_\imath(t,t')-v_\imath(t,t')}\left[\frac{v_\imath(t,t')}{u_\imath(t,t')}\right]^{\frac{q_\imath}{2}} I_{q_\imath}\left(2\sqrt{u_\imath(t,t')v_\imath(t,t')}\right) \quad (6.8)$$

where

$$d_\imath(t,t') = \frac{2\kappa_\imath}{\sigma_\imath^2[1-\mathbf{e}^{-\kappa_\imath(t'-t)}]}$$

$$u_\imath(t,t') = d_\imath(t,t')y_\imath(t^\imath)\mathbf{e}^{-\kappa_\imath(t'-t)}$$

$$v_\imath(t,t') = d_\imath(t,t')y_\imath(t)$$

$$q_\imath = 2\kappa_\imath\theta_\imath/\sigma_\imath^2 - 1$$

and $I_q(\cdot)$ is the modified Bessel function of order $q$. Maximization of the log-likelihood function based on (6.8) will produce consistent estimates of all parameters of the model. However, since I do not observe the factors $y_\imath(t)$ directly, I will use a change of variables and the inverted pricing formula given by (6.7) to estimate the model based on just the observed bond prices (or yields), without imposing any assumptions other than those present in the original CIR model.

Before I proceed to a description of the data and my empirical results, a couple of words should be said about the method of getting the bond price data used for estimation. The data are obtained using the Bliss-Nelson-Siegel method (also called the extended Nelson-Siegel method [13]), which fits the discount rate function $R(s)$ directly to observed bond prices by using the asymptotically flat approximating function:

$$R(s) = \beta_0 + \beta_1\left[\frac{1-\mathbf{e}^{-s/\tau_1}}{s/\tau_1}\right] + \ \beta_2\left[\frac{1-\mathbf{e}^{-s/\tau_2}}{s/\tau_2} - \mathbf{e}^{-s/\tau_2}\right] \quad (6.9)$$

The parameter vector $\Theta = \{\beta_0, \beta_1, \theta_1, \theta_2\}$ is obtained by minimizing the sum of squared weighted pricing errors $\epsilon_i$

$$\min_{\Theta} \sum_{i=1}^{n_t} (w_i \epsilon_i)^2 \tag{6.10}$$

where $\epsilon_i$ is defined in terms of the bid price $N_i^B$, the ask price $N_i^A$, and the predicted price $\hat{N}_i$ as

$$\epsilon_i = \begin{cases} N_i^A - \hat{N}_i & \text{if } \hat{N}_i < N_i^A \\ N_i^B - \hat{N}_i & \text{if } \hat{N}_i < N_i^B \\ 0 & \text{otherwise} \end{cases} \tag{6.11}$$

and the weights $w_i$ are defined in terms of Macaulay duration $d_i$ measured in days:[7]

$$w_i = \frac{1/d_i}{\sum_{j=1}^{n_t} 1/d_j}$$

The minimization has the following constraints

$$R(s_{\min}) \geq 0$$

$$R(\infty) \geq 0$$

and:

$$\mathbf{e}^{-R(s_k)s_k} \geq \mathbf{e}^{-R(s_{k+1})s_{k+1}} \quad \forall s_k < s_{\max}$$

Using an extensive data set on Treasury securities, Bliss demonstrates that this method provides better overall results, while being relatively parsimonious, when compared to other traditional methods of obtaining rates of zero-coupon discount bonds: the unsmoothed Fama-Bliss method, the smoothed Fama-Bliss method, the McCulloch cubic spline method, and the Fisher et al. method. Hence, this is the

[7] Macaulay duration is the weighted average maturity of bond's cash flows.

method I use to obtain prices of riskless zero-coupon discount bonds used for the estimation of risk-free term structure.

## 6.3 Data and Results

I use two bond series, three-month Treasury bills and ten-year Treasury bonds, covering the same period as the loan observations (1990:01-2001:12), and obtained using the Bliss-Nelson-Siegel method (see Bliss [13]). This gives me the total of 144 observations for each of the two series. I choose these particular series to capture the shorter end of the term structure required for discounting purposes, and the longer end required to build one of the covariates that determine patterns of termination. I decided to use 10-year bonds, rather than, say, 30-year ones, to better capture the planning horizon of prepayment decisions.[8] Given these two series, the bond formulae are inverted to yield the state values, contingent on parameter values. The exact likelihood that these states could have occurred is given in (6.8). It can be maximized with respect to the unknown parameters to yield the estimated parameter values of the term structure.

One obvious difference distinguishing term structure estimation from hazard estimation is that I am now working with market data, and so obtain estimates in the risk-neutral form. This means that, assuming standard forms for the risk adjustments, the term structure estimation yields these adjustments together with all the other parameters.[9]

The descriptive statistics for the data used in estimation are presented in Table 6.1. The 3-month rate has the average of 4.8%, with the minimum of 1.76% (December 2001), and the maximum of 8% (February 1990). The 10-year rate has

[8] The distinction between 10-year and 30-year bonds is perhaps of little consequence, given the close correlation between the two.

[9] On the other hand, because the hazard rate dynamics are real, I will have to use a separate procedure to get their risk adjustments.

the average of 6.6%, with the minimum of 4.67% (October 2001), and the maximum of 8.93% (March 1990).

The empirical results of estimation are presented in Table 6.2. In general, they are consistent with the existing literature. The first factor reflects the evolution of the 10-year Treasury rate, while the second can be used as an indicator of the evolution of the spread between the 3-month rate and the 10-year rate. This interpretation is standard in term structure studies, and can be found in Chen and Scott [28][29], Geyer and Pichler [66], Duan and Simonato [48], Duffee [49], among others.

As expected, the first factor has very strong correlation with the 10-year Treasury bond rate (see Table 6.3). The coefficient of correlation between the factor and the long rate is -0.99, while the correlation between first differences is also -0.99. The factor has a long-run average $\theta$ of 0.011, a relatively low volatility $\sigma$ of 0.055, and the half-life of 5.25 years, indicating a relatively low degree of mean-reversion. The market price of risk parameter $\lambda$ is -0.106, indicating that, under the risk-neutral measure $\mathbb{Q}$, the factor is very close to being a random walk. The evolution of the factor is presented graphically in Figure 6.1, along with the evolution of the 10-year Treasury Bond yield during the observation period. The graph also indicates a very close correlation between the factor and the 10-year rate.

The second factor has a strong correlation with the interest rate spread. The correlation coefficient between the factor and the spread is 0.98, while the correlation between the first differences is 0.99. This factor has a mean of 0.045, a volatility of 0.111, with a half-life of 0.86 years, which indicates strong mean-reversion of the interest rate spread. Since the risk-adjustment parameter $\lambda$ is only -0.152, this factor exhibits a relatively high degree of mean-reversion under the risk-neutral measure, as well. Figure 6.2 presents the evolution of the second factor and the (negative) interest rate spread over the observation period.

The constant term $\bar{y}$ is estimated to be $-0.016$, which is generally consistent with the existing empirical literature on extended CIR models, e.g. Pearson and Sun [111], Duffee [49]. Since the value of the constant is negative, it mitigates the problem of potential negative values for the state variables, while making the term structure more flexible. See Duffee [49] for a discussion.

Not all coefficient estimates presented in Table 6.2 are statistically significant, due to the relatively small sample size. The problem is particularly obvious for the risk adjustment parameters. However, the parameter combinations used for asset pricing ( $\theta\kappa$ and $(\kappa + \lambda)$ ) are all statistically significant at 10%.

Finally, Figure 6.3 presents factor loadings of the two factors, which are the $B$ coefficients of (6.7), as a function of maturity. The first factor has an almost equal effect for all maturities, which supports its interpretation as the general interest rate level. The second factor is much more related to the short end, and thus, represents the slope of the term structure.

Overall, the results are consistent with the existing literature, both in terms of magnitude and in terms of interpretation. The average instantaneous rate for the 1990-2001 period is estimated to be about 4.13%, as given by the sum of the two $\theta$'s and the constant $\bar{y}$. The statistical significance of estimated coefficients is somewhat lower than desired, since the data are restricted to the period during which the loans in the sample are observed. To test the model, I also perform estimation of the model using data for the entire 1970-2001 period, and find that estimates are in fact more significant; however, they are also quite different in magnitude. The major difference is that the instantaneous rate is estimated to be over 5.5%. While this is a plausible number for that period, it will cause problems at the next stage, where I use Monte Carlo simulation to price loans originated in the nineties and therefore having much lower contract rates than those originated in the seventies or eighties.[10] Hence, to obtain accurate simulated values of the instantaneous rate and the 10-year bond rate used as a covariate, I use the term structure data from the nineties.

[10] Loans from the seventies and eighties are not used at the pricing stage, to avoid obvious left truncation issues.

**TABLE 6.1**

**Summary Statistics of Interest Rate Data**

Observation Period: January 1990 - December 2001

Sample size: N = 144

| **Variable** | **Mean** | **Median** | **Std.Dev.** | **Min.** | **Max.** |
|---|---|---|---|---|---|
| *3-month Treasury Bill* | 0.048 | 0.050 | 0.013 | 0.017 | 0.080 |
| *10-year Treasury Bond* | 0.066 | 0.064 | 0.010 | 0.047 | 0.089 |

**TABLE 6.2**

**Exact Maximum Likelihood Estimates**

**of a Two-Factor Square-Root Model of Term Structure**

Observation Period: January 1990 - December 2001

Sample size: N = 144

| | **First Factor** | | **Second Factor** |
|---|---|---|---|
| **Variable** | **Estimate** | **Estimate** | **Estimate** |
| | **(Std. Error)** | **(Std. Error)** | **(Std. Error)** |
| $\kappa$ | 0.132 | | 0.803 |
| | (0.128) | | (0.427) |
| $\theta$ | 0.011 | | 0.045 |
| | (0.010) | | (0.023) |
| $\sigma$ | 0.055 | | 0.111 |
| | (0.003) | | (0.014) |
| $\lambda$ | -0.106 | | -0.152 |
| | (0.130) | | (0.414) |
| $R$ | | -0.016 | |
| | | (0.004) | |
| *Log-Likelihood* | | 1,143.03 | |

**Estimates of Parameter Combinations for Asset Pricing**

| | | | |
|---|---|---|---|
| $\kappa\theta$ | 0.002 | | 0.036 |
| | (0.001) | | (0.006) |
| $\kappa + \lambda$ | 0.026 | | 0.652 |
| | (0.014) | | (0.066) |

As always, $\kappa$ us the CIR reversion coefficient for the state variable and $\sigma$ is the volatility parameter, while here, $\theta$ is the constant trend, $\nu$ is the additive risk adjustment, and $\bar{y}$ is a constant, which when added to the sum of the two state variables, yields the spot rate, and makes this an "extended" two-factor CIR model. Note that the half-life, $\ln 2/\kappa$, of the first factor is found to be 5.25 years, whereas the half-life of the second is 0.86 years.

**TABLE 6.3**

**Correlation Coefficients**

| **Variables** | **Correlation** |
|---|---|
| *First Factor and Spread* | 0.503 |
| *First Factor and Spread (First Differences)* | 0.736 |
| *Second Factor and Spread* | -0.991 |
| *Second Factor and Spread (First Differences)* | -0.994 |
| *First Factor and 10-year Treasury Bond* | 0.980 |
| *First Factor and 10-year Treasury Bond (First Differences)* | 0.988 |
| *Second Factor and 10-year Treasury Bond* | -0.448 |
| *Second Factor and 10-year Treasury Bond (First Differences)* | -0.708 |

Figure 6.1: **First Factor of the Term Structure Model and the 10-year Treasury Bond**

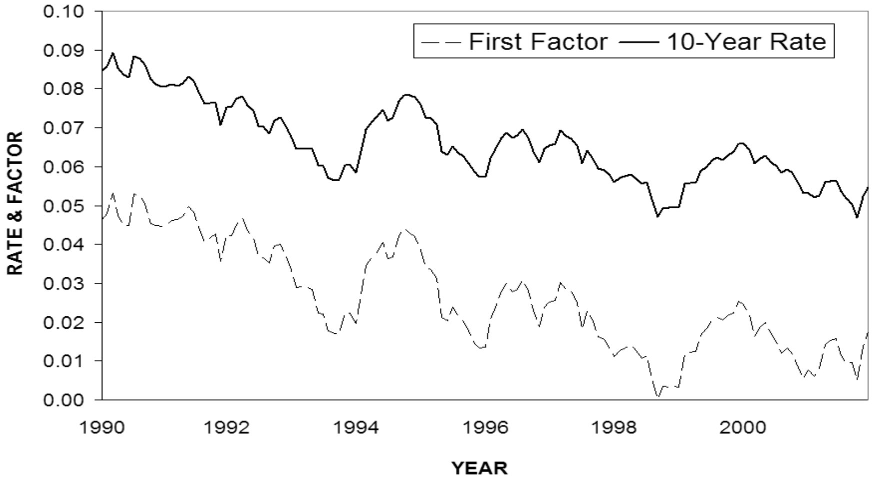


Figure 6.2: **Second Factor of the Term Structure Model and the Spread**

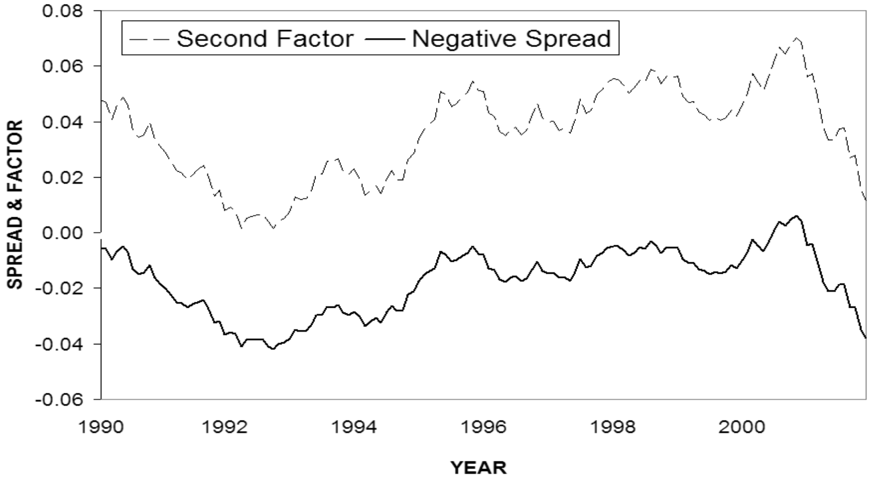
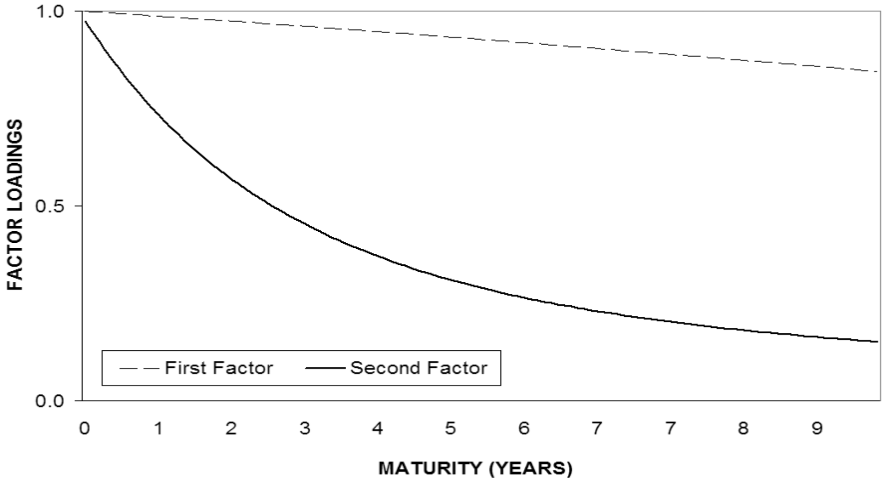


Figure 6.3: **Factor Loadings and the Term Structure Model**



# Chapter 7

# Calibration of Risk-Adjustment Parameters and Pricing

## 7.1 The Monte Carlo Calibration Model

At this stage of the empirical procedure, I am employing Monte Carlo integration to evaluate (3.6) numerically. Forward-looking methods, such as Monte Carlo simulations, are a more natural choice for valuation in a model with exogenous termination (reduced-form models), compared to backward-looking methods, the latter being the standard approach in the structural pricing literature. The most important constraint of backward-looking methods is the complexity of incorporating multiple state variables—there are four of them in the model—while forward methods are less sensitive to this problem of dimensionality.[1] To be able to test the pricing performance of the model, I have to obtain estimates for the risk adjustment parameters, as well as some other variables that enter the pricing equation. The procedure for calibrating these parameters, based on Monte Carlo simulations, is described below.[2]

In the arbitrage-free environment, the value at origination $V(t(0))$ of a mortgage should equal the face amount of the loan $L$, adjusted for points $\delta$, so that

$$V(t(0))/L - (1 - \delta) = 0. \tag{7.1}$$

[1] See Boyle et al. [14] for a thorough discussion of Monte Carlo methods in asset pricing. An even more complete discussion can be found in a recent monograph by Glasserman [69].

[2] Parameter calibration is a tool commonly employed in mortgage finance. See, for example, Titman and Torous [131], who use data on pass-through securities to obtain parameters of the house price process.

where V(t(0) is defined by (3.6), or equivalently, can be expressed in a more intuitive form (also more convenient from the computational point of view) as

$$
\begin{aligned}
V(t(0)) = E^{\mathbb{Q}}_{t(0)}\Bigg[\sum_{i=1}^{I} \mathrm{e}^{-\int_{t(0)}^{t(i)}((1-\tau_F)r(s)+\ell)ds}\Bigg(\prod_{j=1}^{i-1}(1-\lambda^d_j-\lambda^p_j)\Bigg)\Bigg]\lambda^d_i W(i) + \\
E^{\mathbb{Q}}_{t(0)}\Bigg[\sum_{i=1}^{I} \mathrm{e}^{-\int_{t(0)}^{t(i)}((1-\tau_F)r(s)+\ell)ds}\Bigg(\prod_{j=1}^{i-1}(1-\lambda^d_j-\lambda^p_j)\Bigg)\Bigg]\lambda^p_i A(i) + \\
E^{\mathbb{Q}}_{t(0)}\Bigg[\sum_{i=1}^{I} \mathrm{e}^{-\int_{t(0)}^{t(i)}((1-\tau_F)r(s)+\ell)ds}\Bigg(\prod_{j=1}^{i}(1-\lambda^d_j-\lambda^p_j)\Bigg)\Bigg]M \\
I = 360 \qquad (7.2)
\end{aligned}
$$

where the first term represents the expected value of future mortgages payments, the second term represents the expected cash flows on prepayment, and the third term represents the expected recovery on default.

By looking at (7.2), one can see the parameters that have to be estimated. First, I have to obtain a liquidity term $\ell$ (introduced in the model following Duffie and Singleton [56]), reflecting the illiquidity of mortgage markets relative to the market for Treasuries.

Second, there are two default risk adjustments $(\mu^d, \nu^d)$ and two prepayment risk adjustments $(\mu^p, \nu^p)$. One can either attempt to estimate all of them, or assume that the multiplicative parameter is equal to one, following the conditional diversification hypothesis of Jarrow, Lando and Yu [81].

Third, there exists recovery on default. The recovery value $W(i)$ depends on whether the loan is insured, which it usually is as long as the unpaid principal is of more than 80% LTV, so it is assumed that all loans with LTV$>$ 80% have insurance. I include insurance in recovery value, since it is part of the contract from the lender's point of view, and will affect the appropriate terms of the contract used below to predict termination risk adjustments. Actual recovery will then be

$W(i) = (\phi + \psi)A(i)$, where $\phi$ is the percent insured and $\psi$ is the percent recovered in the absence of insurance. This formulation corresponds to the case known as "recovery of face value" (RFV) in the literature. I am going to assume, however, that the recovery rate is $\psi = 1 - \omega$, where the loss rate $\omega$ is stochastic, ex ante, though proportionate to the current loan-to-value ratio, so $\omega = k(U(i)/H(0))$, with $k$ being a random variable and $H(0)$ the original house value. If I assume that that recovery (loss) is distributed independently of the other stochastic state variables, then (see Schönbucher [123]), the relevant expression for the expected recovery $W^e(i)$ becomes $W^e(i) = (1 + \phi - k^e(U(i)/H(0))A(i)$. Note that in considering the expectation, I am working with the risk-neutral measure, and so the risk-adjusted expected loss parameter $k^e$ need not indicate the real expected loss.[3]

The expressions for the mortgage payment $M$ and the unpaid balance $A(i)$ are contractually specified and entirely deterministic. The expression for the mortgage payment with an FRM is $M' = L[\frac{c/12}{1-\frac{1}{(1+c/12)^{360}}}]$, with $c$ being the contract rate, and $L$ being the loan size. Interest from such loans, though, is subject to state and Federal taxes, at the rates $\tau_S$ and $\tau_F$, respectively. These tax rates can also be either estimated, or imposed using external information. Accounting for taxation results in the net receipts being only $M = (1-\tau_F)(1-\tau_S))(1+c)U(i-1)+(M'-(1+c)U(i-1))$, where $U(i) = M'[\frac{1-\frac{1}{(1+c/12)^{360-i}}}{c/12}]$ is the unpaid principal in period $i$, after a payment. The receipt upon prepayment is then simply $A(i) = U(i) + M$.

For the calibration procedure, I choose a random selection of loans from within the category with substantial LTV values (over 50%) and a relatively recent origi-

[3] For convenience, I have assumed matters are such that, with 100% recovery, the payoff to default is indistinguishable from a prepayment. A tax adjustment may need to be introduced if losses are deductible, but I have not bothered to explicitly express this, since even if it is present, it will simply be absorbed into the estimate of the recovery rate. The expression for $W(i)$ ought to really be $\min(1, \phi + \psi)A(i)$, so the above expression ignores ignore the possibility that insurance and recovery rates will ever together reach unity. In the estimations, I take $\phi$ to be 20%, at an 80% loan to value.

nation period,[4] and solve for parameters values that minimize the sum-of-squared errors

$$\sum_{h=1}^{H}(V_h(t(0))/L_h - (1-\delta_h))^2 \tag{7.3}$$

where $H$ is the number of loans in the sub-sample.

For each iteration of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm, 1,000 draws of the Monte Carlo simulations were conducted, with 30 time increments per month over the 30-year life of the mortgage. Starting values of the term structure are dictated by the actual bond values at origination, while the starting values of the baseline hazards are always taken to be near zero, in conformity with the assumed shapes of the processes' trends.

The first step of the Monte Carlo calibration requires simulation of diffusions of state variables under the risk-neutral measure $\mathbb{Q}$. The model includes four state variables, i.e. two baseline hazard processes $\lambda_0^l(t)$ and two variables $y_\imath(t)$, used to construct the instantaneous risk-free rate $r(t)$.

For the term structure, the two factors are assumed to follow the CIR diffusion process given by (6.2) under $\mathbb{Q}$.

One can easily use the conventional Euler (5.17) or Milstein (5.18) schemes to simulate the process, or the exact transition density, based on the fact that, if a factor transformation is given by $x_\imath(t) = 2cy_\imath(t)$, then $x_\imath(t)|x_\imath(t-\Delta)$ is distributed as non-central chi-squared, with $4\kappa'_\imath\theta'_\imath/\sigma_\imath$ degrees of freedom and the non-centrality parameter $x_\imath(t-\Delta)\mathbf{e}^{-\kappa'_\imath\Delta}$, where $c = 2\kappa'_\imath/[\sigma_\imath(1-\mathbf{e}^{-\kappa'_\imath\delta})]$, and $\kappa'_\imath$ and $\theta'_\imath$ are the risk-adjusted parameters, $\kappa'_\imath = \kappa_\imath + \nu_\imath$ and $\theta'_\imath = \kappa_\imath\theta_\imath/(\kappa_\imath+\nu_\imath)$. All diffusion parameters are those earlier obtained by exact maximum likelihood for $\imath = 1, 2$. Unfortunately,

[4] The latter is intended to mitigate the problem of tax values, as well as market liquidity, not being constant over time.

obtaining random non-central chi-square variates is a very time-consuming procedure, and so I use the Euler scheme to propagate term structure factors. Empirical tests show that the results are very similar to those provided by the exact propagation scheme.

Random numbers are drawn using the generator of Park and Miller (RAN1). To reduce the variance of the MC estimator, antithetic variables are used in a manner identical to that used at the particle-filtering stage. Starting values for interest rate factors are obtained by inverting the pricing formulae for bond prices at the month of each loan's origination.

After propagating the processes, the estimate of the interest rate at each point is obtained by summing up the values of the two factors and the constant given in (6.6).[5] Given values of the interest rate, its integrals are obtained by Monte Carlo integration (summing up and dividing by the number of intervals). The 10-year Treasury rate, used as a covariate, is obtained by the pricing formula given in (6.7). Its value at the end of each month is used as a covariate for modelling the default/prepayment hazards.

At the next step, I propagate baseline hazard processes of default and prepayment. The baseline intensities are propagated by the Euler method, based on the corresponding diffusions (5.14) and (5.16), using the parameter estimates previously obtained via particle-filtering estimation. Simulated hazard rates at the end of each month are used to build the realized hazard rates, according to (3.2).

Once estimates of all discounting integrals are obtained, I use (7.2) to find the mortgage value at origination for each simulation. Summing up the values across simulations and dividing the sum by the total number of simulations produces the

[5] As a test, I perform pricing of a default/prepayment free bond using a closed-form bond solution. The results based on 1,000 simulations are found to be accurate

usual Monte Carlo estimate of the expected value of the loan. The optimization algorithm is then used to find the parameter values that minimize (7.3).

## 7.2 Parameter Calibration Results

Since Monte Carlo is a relatively time consuming algorithm,[6] I use a relatively small number of simulations to calibrate necessary parameters: 1,000 simulations, with 30 intermediate step each.

Some earlier tests of the performance of the Monte Carlo calibration device indicate that it is virtually impossible to obtain estimates of the complete set of eight parameters, i.e. the two tax rates, the liquidity premium, the loss on default, and the four risk-adjustment parameters. Therefore, I restrict the set of parameters to include only those directly related to the mortgage market: the risk-adjustment parameters, the loss rate, and the liquidity premium in the mortgage market. The tax rates are set to their historical median values, 28% for the Federal rate,[7] and 4.32% for the effective state rate.[8]

Because of difficulties in identifying the prepayment parameters, however, rather than calibrate all the parameters simultaneously, I used a two-stage process. The reason for difficulties in prepayment risk-parameter calibration is not hard to find: by arbitrarily increasing the additive or multiplicative, one can always induce immediate prepayment, which then automatically balances the loan, up to points. Thus, endlessly increasing the prepayment additive or multiplicative can be an attractive

[6] In the current implementation, it takes approximately 6.5 seconds per evaluation with 1,000 simulations and 30 intermediate steps on a 2GHz PC with 1Gb RAM.

[7] The Federal tax rate of 28% is also consistent with the results reported in a recent paper published by the Congressional Budget Office [1], which finds the total effective Federal rate for the highest income quintile to vary between 25.1% and 28% during 1990-2001.

[8] The effective state rate of 4.32% is obtained by taking the median state tax rate (6%) and multiplying it times $(1 - \tau_F)$.

way for the optimizer to reduce its sum-of-squared errors. To avoid this dilemma, I first isolated the effect of prepayment by restricting myself to sets of loans with loan-to-value ratios of less than 50%, so that their default component became numerically trivial, and could therefore be ignored. I then further restricted these loans to only be ones having points, thus making it feasible to calibrate their prepayment parameters. Having determined these, I subsequently calibrated the remaining parameters, those of greater interest to me, using more typical sets of loans with greater than 50% LTV ratios.

While my entire empirical estimation has been built on the assumption that the real hazard processes are stable over time, nothing requires this to be the case for the risk adjustments or liquidity.[9] In particular, the measured difference between the yield on GNMA pass-through securities and the conventional mortgage rate, which is easily obtained using outside data, may be thought to serve as an imperfect measure of the required liquidity premium. The data on the conventional mortgage rates are obtained from the FRED II database of the Federal Reserve Bank of St. Louis. The GNMA yield data are obtained from the *Federal Reserve Bulletin*. These data are averages of the yields published in Friday editions on the *Wall Street Journal*, and assume a 12-year average life.[10] Descriptive statistics of the data are given in Table 7.1, and Figure 7.1 graphically presents the evolution of the liquidity premium obtained in this manner over the observation period. The premium has the mean of 47.81 basis points, which is indeed not far from the liquidity premium of 38.9 basis points reported in the second stage of my calibration, when using a sample of loans from over the entire observation period.[11] However, the average monthly absolute

[9] Note that the reported liquidities in the first stage of calibration, the ones for exceptionally low LTV loans, exceed the second-stage ones, as might be expected.

[10] The original data are provided to the *Wall Street Journal* by Bear, Stearns & Co.

[11] My results are also consistent with estimates earlier reported in empirical literature, e.g. Marathe and Shawky [104], who estimate the liquidity premium to be 45 basis points during 1989-96.

deviation of the GNMA liquidity measure turns out to be 17 basis points, making it quite volatile.

This suggests the need to recalibrate liquidity on a monthly basis, and so I include two such calibrations in Table 7.2, for the months of January and July 2001, using only loans from those respective months. While the liquidity measure does change noticeably, the other parameters remain reasonably stable, in comparison to one another and to the calibration over the entire observation period. As then also might be expected with volatile liquidity, the RMSE (root mean squared error) of the monthly calibrations are considerably smaller than the RMSE of the calibration done over the entire observation period.[12]

Jarrow, Lando and Yu [81] have formally demonstrated the natural result that, with sufficient diversification, the multiplicative risk parameter for default should approach one in value. The general idea, though, that premia for idiosyncratic jump risk can be diversified away over sufficient contracts goes back at least to Merton [107], and should apply to the risk of prepayment, as well as to default. My setting is a particularly appropriate one in which to test the applicability of such large-number reasoning; my sample of nearly a million mortgages is but a fraction of the entire population, and my driving assumption has been that all loans are based on common hazard processes. Furthermore, my formulation and data uniquely put me in the position of being able to determine both the additive and multiplicative risk parameters of default and prepayment, using the very same data set employed to estimate the real processes. Duffee [49], who uses only market data, is unable to determine default multiplicatives at all, whereas Driessen [47] does calibrate the default multiplicatives, but only after introducing a second set of data, which being

[12] The RMSE I obtain over the entire observation period is similar to the 1.6% RMSE obtained in Titman & Torous [131], who were, however, working with commercial mortgages, which feature the possibility of default, but not prepayment.

real, is rather disparate from the market data used to obtain the other parameters of the hazard processes.

As the results in 7.2 reveal, the multiplicative risk parameter on default is substantially greater than one in value, in violation of the diversification hypothesis. Since I regard the mortgage market as the best candidate imaginable for the applicability of jump-risk diversification reasoning, this suggests that one is unlikely to find any other situation where it satisfactorily applies. The multiplicative risk parameter for prepayment, on the other hand, is much closer to one, but still substantially greater in value.

The additive risk adjustments $\nu^\ell$ that I obtain indicate aversion to both the systematic prepayment and default risk components. However, they are not so negative as to prevent the market-adjusted reversion factors from remaining mean-reverting for both the default and prepayment processes. This is in contrast to the previous results of Duffee [49] and Driessen [47], where in their various estimations, at least one of the default multiplicatives is always mean-averting. I would argue that mean aversion should be taken as an indication of likely misspecification. A negative market reversion factor effectively means that the market regards the hazard trend as negative. Since the initial hazard value is presumably positive, it will be far removed from this trend, and will with great likelihood explode away from it over time. It is hard to then give such processes, particularly their trends, a satisfactory interpretation, even if one regards market probabilities as merely a device for calculating present values. If, as in this previous literature, one then infers the real process from the market one, and it is stable, then the two trends, real and market, have opposite signs, and any relation between them seems somewhat obscure. Indeed, a market process being unstable would seem to call into doubt the significance of any real process that might be inferred from it. Since I have already separately estimated our real processes and do attach considerable significance to them, particularly default, it seems especially important that my market processes maintain a close connection with the real ones, and this then calls for the market-adjusted reversion factor to be positive.

It is also worthwhile emphasizing that my approach allows me to determine recovery along with the risk parameters. This is not possible in the framework of Duffee [49] and Driessen [47], due to the interaction of a series of restrictions in which they allow only default, use primarily market data, and take a recovery of market value (RMV) approach. Indeed, relaxing any one of these three assumptions, and I relax all three, allows one to then determine recovery. My calibrated loss parameter indicates that, at a 100% LTV ratio, the expected loss is 28.24% of the remaining loan balance, as calibrated over the entire observation period.[13]

## 7.3 Pricing Performance

After calibrating our model, I performed out-of-sample tests, using a selection of mortgages, coming from within the original pool of mortgages, but not including any of those mortgages used in the calibration. The results are reported in 7.3. Using data for the entire sample period, I find the RMSE values to range between 1.407% and 2.131% for different LTV groups. The loans with the loan-to-value ratio greater than 80% and less than 85% have the highest pricing error. Overwise, there are no clear patterns in pricing errors among these groups. When loans in the sample (as well as calibrated parameters) are restricted to a single month of origination, the pricing performance shows a clear improvement: RMSE varies between 1.267% and 1.648% for the January 2001 sample, and between 1.148% and 1.788% for the July 2001 sample.

[13] While I have commented before that the loss rate involves market expectations and need not be real, it nonetheless appears quite stable and corresponds well with my priors concerning average real loss rates.

It is also of some interest to consider the impact of the various components of a mortgage: payments, prepayments, and default. Randomly picking a typical mortgage,[14] I calculate that its value would have been \$104,875 in the absence of either prepayment or default, while its value should have been \$104,251 with prepayment only, and \$100,836 with only the possibility of default.[15] On the other hand, the actual contract, with both the possibility of prepayment and default, is valued at \$102,802. Thus, prepayment, alone, lowers the value of the loan to the lender by \$624, default, alone, lowers the value by \$4,039, and both of them lower the value by \$2,073. The \$2,590 difference between \$2,073 and $\$624 + \$4,039 = \$4,663$ reflects the fact that prepayment precludes default and vice versa, a phenomenon familiar from structural models of mortgage termination.

Prepayment is much more common than default, but the harm of a prepayment is limited in scope, and so, default plays the greater role in explaining the spread in contract rates over Treasury rates. A default is always to a lender's disadvantage, and though rare, can lead to substantial loss, while prepayment may or may not be harmful to the lender, depending on whether interest rates are high or low. Of course, as my estimation of the spread covariates for prepayment confirms, there is substantially more prepayment in those low interest rate cases that are harmful to the lender, but nonetheless, the beneficial cases partly offset the harmful ones, and with the consequences of either being limited, prepayment requires less of a premium than does default.

It is also interesting to consider the decomposition of just one mortgage. If I take my sample mortgage and decompose its predicted value, I find that the value

[14] The mortgage randomly picked for the example originated in July, 2001, and had an original loan balance of \$102,940, with 0.75 points, an 80% LTV, and a contract rate of 7.25%. Parameters obtained from the July 2001 sample were used in all of following pricing exercises.

[15] I continue to hold the contract rate constant at 7.25% throughout this discussion.

of its likely payments is \$34,295 the value of its likely prepayment settlements is \$61,771, and the value of any likely default recovery is \$6,737. At first, it appears a bit implausible that the value of default recovery could be anything like that of prepayment, given that the former is much more infrequent and my estimations show there to be far from full recovery. However, I am working with market-adjusted probabilities, and the aversion to default risk effectively inflates the likelihood of default, allowing for a considerable "expected" value to the default component. To see how this can be consistent with the fact that more default brings down the value of the mortgage, one has to realize that all this artificial probability being attached to default is taking away market-adjusted likelihood from prepayment and continuation, alternatives which would end with higher payments for the lender.

My ability to predict mortgage values, while respectable, is not all that one might wish.[16] One feature undoubtedly present is variation in the personal characteristics of the borrower. An important function of lending institutions is to evaluate the creditworthiness of the borrower, and to set contract rates accordingly; however, I do not have access to this information, and so, cannot explain this part of the variation in contract rates. In building models of corporate debt, it is not unusual to introduce a separate hazard process component for every firm (see Duffee [49] or Driessen [47]). In my case, this would correspond to to estimating nearly a million separate stochastic processes, a clearly absurd possibility, even were there data available to support such an endeavor. Of necessity, I assume, instead, that different mortgages share common hazards, up to covariates, with these covariates including only externally observable characteristics of the mortgage.[17]

[16] Note, however, that my results are about the same out-of-sample as in-sample. I have avoided such devices as calibrating a term structure to one particular time, in order to increase the ability to predict values near that time.

[17] Presumably, my predictive powers will appear more impressive when used to evaluate a mortgage-backed security, where individual borrower differences will be washed out by the large number of mortgages.

Ultimately, I am more interested in explanation than prediction; one might have run a simple regression of our covariates against the spread of contract rates over corresponding 10-year Treasury rates, and perhaps done as good a job as our model of predicting contract rates. What my model offers, instead, is an explanation of why the resulting regression coefficients would take on the values they do, an explanation in terms of the observed behaviors of borrowers as to how they do prepay or default, and the need to explain the spread of contract rates in a way consistent with the absence of arbitrage.

**TABLE 7.1**

**Summary Statistics of Liquidity Data**

Observation Period: January 1990 - December 2001

Sample size: N = 144

| **Variable** | **Mean** | **Median** | **Std.Dev.** | **Min.** | **Max.** |
|---|---|---|---|---|---|
| Conventional Mortgage Rate | 7.999 | 7.855 | 0.965 | 6.620 | 10.480 |
| GNMA Pass-Through Rate | 7.517 | 7.425 | 0.955 | 5.860 | 9.770 |
| Liquidity | 0.478 | 0.480 | 0.229 | 0.000 | 1.110 |

Figure 7.1: **Liquidity Premium, 1990-2001**

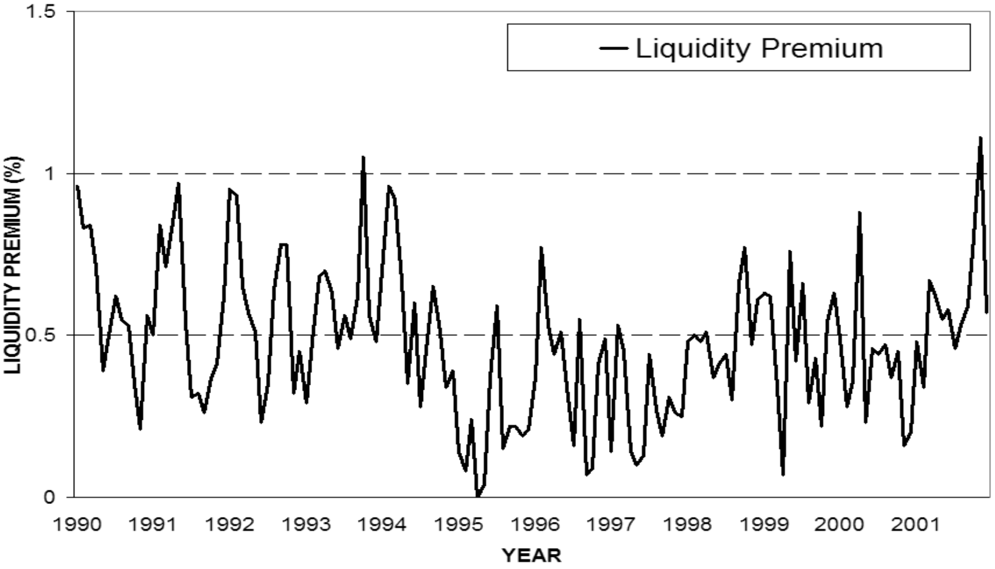


**TABLE 7.2**

**Monte Carlo Calibration Results**

Number of Loans in Each Sample: H=100

Number of Simulations: N=1,000

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | **1990-2001 Sample** | **January 2001 Sample** | **July 2001 Sample** |
|---|---|---|---|
| **Variable** | **Estimate** | **Estimate** | **Estimate** |
| $\mu_p$ | 1.909 | 1.630 | 1.613 |
| $\nu_p$ | -0.323 | -0.586 | -0.555 |
| *Liquidity (%)*[18] | 0.507 | 0.669 | 0.755 |
| *RMSE (%)* | 1.980 | 1.579 | 1.328 |
| | | | |
| $\mu_d$ | 9.018 | 9.024 | 9.013 |
| $\nu_d$ | -1.224 | -1.248 | -1.076 |
| *Loss Rate (%)* | 28.242 | 26.176 | 28.044 |
| *Liquidity (%)* | 0.389 | 0.553 | 0.643 |
| *RMSE (%)* | 1.699 | 1.365 | 1.136 |

[18] The liquidity premium estimated at the first stage represents that for the loans with LTVs under 50%. It is not used for the following tests of pricing performance.

**TABLE 7.3**

**Out-of-Sample Pricing Errors**

Number of Loans in Each Sample: H=100

Number of Simulations: N=1,000

Propagation Method: Euler

Monthly Time Subintervals: $1/\Delta = 30$

| | **1990-2001 Sample** | **January 2001 Sample** | **July 2001 Sample** |
|---|---|---|---|
| **LTV Category (%)** | **RMSE (%)** | **RMSE (%)** | **RMSE (%)** |
| $70^{+} - 75$ | 2.079 | 1.267 | 1.260 |
| $75^{+} - 80$ | 1.766 | 1.274 | 1.148 |
| $80^{+} - 85$ | 2.131 | 1.473 | 1.296 |
| $85^{+} - 90$ | 1.813 | 1.648 | 1.412 |
| $90^{+} - 95$ | 1.407 | 1.488 | 1.409 |
| $95^{+} - 100$ | 2.004 | 1.465 | 1.788 |

### 7.3.1 Conclusions

Default and prepayment are uncertain events that determine a mortgage's life and hence affect its present value. Duration analysis is a common statistical technique used to analyze patterns of default and prepayment. However, as usually employed, the baseline of a general duration model is a passive object; all the explanation lies with the observable covariates. Thus, in employing a duration model to describe mortgage termination one is effectively assuming that the probability a mortgage will terminate is commonly known, once the observable covariates become known. With a large number of similar mortgages, one then knows the exact proportion that will default, i.e. with sufficient diversification, the only risk premium required is that which reflects the market risk of the observable covariates. In practice, the term structure is likely to be the only uncertain covariate, in which case, there would be no more premium for a defaultable contract's risk than for a default-free one. Reduced-form modelling of mortgage termination recognizes that there are sources of uncertainty other than those represented by the available covariates. These sources are described by a "doubly stochastic" model, where, not only the event of default, but also the probability of such default, remains uncertain.

I unite the duration analysis and reduced-form pricing literatures, by dividing mortgages into strata by time of origination, and treating the baseline of each stratum as a separate draw from a common stochastic process. This second source of uncertainly then offers the possibility of explaining how the contract rate on, say, a default-free, prepayable mortgage-backed security can so significantly exceed the long-term Treasury rate. A strict duration model, with only a term structure as the time-varying covariate, could never do this without resorting to implausibly high values for the premium to the idiosyncratic risk of prepayment.

In order to capture the doubly stochastic nature of the model, I use a "state-space" structure to model the baseline hazard processes of prepayment and of default. Each baseline model consists of a stochastic transition equation, representing the first layer of uncertainty, together with a non-dynamic observation equation, which has the second stochastic component. The stochastic variation in the transition equation reflects the external economic conditions common to a cohort of loans. The equation also has a time-varying trend that reflects the internal evolution in the likelihood of loan termination. The stochastic component of the observation equation, on the other hand, reflects the fact that the actual number of terminations is also influenced by idiosyncratic risk.

Kalman filtering is the traditional way to estimate such a state-space model; however, it provides consistent estimates only if the model is linear and Gaussian, whereas hazard models, like mine, typically are neither. I, thus, use a particle-filter approach, since it continues to provide consistent estimates in a non-linear, non-Gaussian setting.

The term structure of interest rates, on the other hand, is not a doubly stochastic process. Any noise in the term structure process is usually attributed to measurement errors. To avoid arbitrary assumptions about the nature of this measurement error, I estimate a two-factor model of term structure using the method of exact maximum likelihood.

Unlike firms, homeowners do not have a wide variety of associated contracts, individually reflecting their risk of default or prepayment. However, this lack of market data for loans is compensated for by the prevalence of historical data, since mortgage terminations are far more frequently observed than are firms going into default. To obtain parameters of the risk-adjusted processes given the physical ones, and so be able to price mortgages, I go back to mortgage contracts, and invoke a no-arbitrage principle: the present value of the loan's payments, priced according to the market's assessment of the riskiness of the loan, must be equal to the face value of the loan. This then yields the risk adjustments that convert the actual hazard processes into risk-neutralized ones, as well as yielding expected recovery on default.

With two hazard processes and a two-state term structure, I have four state variables, used in the Monte Carlo estimation of mortgage values. Since the term structure also appears in the hazard covariates, I am explicitly allowing for dependency between the stochastic interest rates and the hazards of prepayment and default.

What is absent from the model is any account of the individual characteristics of the borrower, such as, for example, the FICO scores commonly used by the lender to assess borrower's likelihood of default. Given the proportional hazard basis, the model would be ideally suited for including such factors, if I only had such data. Another particularly interesting possibility, also limited only by the availability of data, would be to introduce information about the evolution in house prices associated with the various originating mortgages. Such a reduced-form model could then be easily compared with the earlier structural models of prepayment and default, which typically use both interest rates and house prices as their driving variables. The current model could also easily be adapted to value the various instruments available in the secondary mortgage market, or even applied to price mortgage insurance.

# Bibliography

[1] Effective federal tax rates: 1979-2001. Congressional Budget Office, April 2004.

[2] Odd Aalen. Non-parametric inference for a family of counting processes. *Annals of Statistics*, 6(4):701–726, 1978.

[3] Brent Ambrose and Anthony Sanders. Commercial mortgage-backed securities: Prepayment and default. *Journal of Real Estate Finance and Economics*, 26(2-3), 2003.

[4] P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, 10(4):1100–1120, Dec. 1982.

[5] Per Kragh Andersen, Ornulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical Models Based on Counting Processes.* Springer Series in Statistics. Springer-Verlag, 1993.

[6] Wayne R. Archer, David C. Ling, and Gary A. McGill. The effect of income and collateral constraints of residential mortgage terminations. *Regional Science and Urban Economics*, 26:235–261, 1996.

[7] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[8] Alain Bélanger, Steven E. Shreve, and Dennis Wong. A general framework for pricing credit risk. Working Paper, April 2003.

[9] C. Berzuini, N. G. Best, W. R. Gilks, and C. Larizza. Dynamic conditional independence models and Markov chain Monte Carlo methods. *Journal of American Statistical Association*, 92:1403–1412, 1997.

[10] Tomasz R. Bielecki and Marek Rutkowski. *Credit Risk: Modelling, Valuation and Hedging*. Springer, 2002.

[11] Tomas Björk, Yuri Kabanov, and Wolfgang Runggaldier. Bond market structure in the presence of marked point processes. *Mathematical Finance*, 7(2):211–223, April 1997.

[12] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The Journal of Political Economy*, 81(3):637–654, May-June 1973.

[13] Robert R. Bliss. Testing term structure estimation methods. *Advances in Futures and Options Research*, 9:191–231, 1997.

[14] P. Boyle, M. Broadie, and P. Glasserman. Monte Carlo methods and security pricing. *Journal of Economic Dynamics and Control*, 21:1267–1321, 1997.

[15] Pierre Brémaud. *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag, 1981.

[16] Michael J. Brennan and Eduardo S. Schwartz. Determinants of GNMA mortgage prices. *AREUEA Journal*, 13(3):209–228, 1985.

[17] N. E. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99, Mar. 1974.

[18] Andrew L. Brunson, James B. Kau, and Donald C. Keenan. A fixed-rate mortgage valuation model in three state variables. *The Journal of Fixed Income*, 11(1):17–27, June 2001.

[19] Stephen A. Buser and Patrick H. Hendershott. Pricing default-free fixed-rate mortgages. *Housing Finance Review*, 3(4):405–429, 1984.

[20] Stephen A. Buser, Patrick H. Hendershott, and Anthony B. Sanders. Pricing life of loan caps on default-free adjustable-rate mortgages. *AREUEA Journal*, 13(3):248–260, 1985.

[21] Charles A. Calhoun and Yongheng Deng. A dynamic analysis of fixed- and adjustable-rate mortgage terminations. *Journal of Real Estate Finance and Economics*, 24:9–33, 2002.

[22] A. Colin Cameron and Pravin K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, 1998.

[23] Tim S. Campbell and Kimball Dietrich. The determinants of default on insured residential mortgage loans. *The Journal of Finance*, 38(5):1569–1581, Dec. 1983.

[24] J. Carpenter, P. Clifford, and P. Fearnhead. Improved particle filter for nonlinear problems. *IEE Proceedings on Radar and Sonar Navigation*, 1999.

[25] Amitava Chatterjee, Robert O. Edmister, and Gay B. Hatfield. An empirical investigation of alternative contingent claims models for pricing residential mortgages. *Journal of Real Estate Finance and Economics*, 17(2):139–162, 1998.

[26] Ren-Raw Chen and Loius Scott. Pricing interest rate options in a two-factor Cox-Ingersoll-Ross model of the term structure. *The Review of Financial Studies*, 5(4):613–636, 1992.

[27] Ren-Raw Chen and Loius Scott. Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates. *Journal of Fixed Income*, 3:14–31, December 1993.

[28] Ren-Raw Chen and Loius Scott. Interest rate options in multifactor Cox-Ingersoll-Ross models of the term structure. *The Journal of Derivatives*, 20:53–72, 1995.

[29] Ren-Raw Chen and Louis Scott. Multi-factor Cox-Ingersoll-Ross models of the term structure: Estimates and tests form a Kalman filter model. *Journal of Real Estate Finance and Economics*, 27(2):143–172, 2003.

[30] Brian Ciochetti, Yongheng Deng, Gail Lee, James Shilling, and Rui Yao. A proportional hazard model of commercial mortgage default with originator bias. *Journal of Real Estate Finance and Economics*, 27(1), 2003.

[31] Brian A. Ciochetti, Yongheng Deng, Bin Gao, and Rui Yao. The termination of mortgage contracts through prepayment and default in the commercial mortgage markers: A proportional hazard approach with competing risks. *Real Estate Economics*, 30(4):595–633, 2002.

[32] John M. Clapp, Gerson M. Goldberg, John P. Harding, and Michael LaCour-Little. Movers and shuckers: Interdependent prepayment decisions. *Real Estate Economics*, 29(3):411–450, 2001.

[33] P. Collin-Dufresne and J. Harding. A closed-form formula for valuing mortgages. *Journal of Real Estate Finance and Economics*, 19(2):133–146, 1999.

[34] T. G. Conley. GMM estimation with cross-sectional dependence. *Journal of Econometrics*, 91:1–45, 1999.

[35] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, Aug. 1975.

[36] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, 34(2):187–220, 1972.

[37] John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. An analysis of variable rate loan contracts. *Journal of Finance*, 40:293–308, 1980.

[38] John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. A theory of term structure of the interest rates. *Econometrica*, 53(2):385–408, Mar. 1985.

[39] Donald F. Cunningham and Charles A. Capone. The relative termination experience of adjustable- to fixed-rate mortgages. *The Journal of Finance*, 45(5):1687–1703, Dec. 1990.

[40] Donald F. Cunningham and Patrick H. Hendershott. Pricing FHA mortgage default insurance. *Housing Finance Review*, 3(4):383–392, 1984.

[41] Qiang Dai and Kenneth J. Singleton. Specification analysis of affine term structure models. *The Journal of Finance*, 55(5):1943–1978, Oct. 2000.

[42] Enrico DeGiorgi. An intensity based non-parametric default model for residential mortgage portfolios. Working Paper, Swiss Federal Institute of Technology, 2001.

[43] Yongheng Deng. Mortgage termination: An empirical hazard model with stochastic term structure. *Journal of Real Estate Finance and Economics*, 14(3):309–331, 1997.

[44] Yongheng Deng, John M. Quigley, and Robert Van Order. Low downpayment loans for homeowners: The cost of public subsidy. *Regional Science and Urban Economics*, 26(3-7):263–287, 1996.

[45] Yongheng Deng, John M. Quigley, and Robert Van Order. Mortgage termination, heterogeneity and the exercise of mortgage options. *Econometrica*, 68(2):275–307, March 2000.

[46] A. Doucet, J. F. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Cambridge: Cambridge University Press, 2001.

[47] Joost Driessen. Is default event risk priced in corporate bonds? Working Paper, University of Amsterdam, March 2002.

[48] Jin-Chuan Duan and Jean-Guy Simonato. Estimating and testing exponential-affine term structure models by Kalman filter. Working Paper, CIRANO, 1995.

[49] Gregory Duffee. Estimating the price of default risk. *Review of Financial Studies*, 12:197–226, 1999.

[50] D. Duffie, M. Schroder, and C. Skiadas. Recursive valuation of defaultable securities and the timing of resolution of uncertainty. *Annals of Applied Probability*, 6(1):1075–1090, 1996.

[51] Darrell Duffie. Defaultable term structure models with fractional recovery of par. Working Paper, August 1998.

[52] Darrell Duffie. First-to-default valuation. Working Paper, Stanford University, May 1998.

[53] Darrell Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, Princeton, New Jersey, 3rd edition, 2001.

[54] Darrell Duffie and David Lando. Term structure of credit risk spreads with incomplete accounting information. *Econometrica*, 69(3):633–664, 2001.

[55] Darrell Duffie and Kenneth J. Singleton. An econometric model of the term structure of interest-rate swap yields. *The Journal of Finance*, 52(4):1287–3121, 1997.

[56] Darrell Duffie and Kenneth J. Singleton. Modelling term structures of defaultable bonds. *Review of Financial Studies*, 12(4):687–720, 1999.

[57] Darrell Duffie and Kenneth J. Singleton. *Credit Risk*. Princeton University Press, Princeton, New Jersey, 2003.

[58] Kenneth B. Dunn and John J. McDonnell. A comparison of alternative models for pricing GNMA mortgage-backed securities. *The Journal of Finance*, 36(3):375–392, June 1981.

[59] Kenneth B. Dunn and John J. McDonnell. Valuation of GNMA mortgage-backed securities. *The Journal of Finance*, 36(3):599–616, June 1981.

[60] Garland B. Durham. Likelihood-based specification analysis of continuous-time models of the short-term interest rate. Working Paper, University of Iowa, 2002.

[61] Garland B. Durham and A. Ronald Gallant. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business and Economic Statistics*, 20(3):297–316, Jul. 2002.

[62] Bradley Efron. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 76(359):557–565, Sep. 1977.

[63] James F. Epperson, James B. Kau, Donald C. Keenan, and Walter J. Muller III. Pricing default risk in mortgages. *AREUEA Journal*, 13(3):152–167, 1985.

[64] James Follain, Jan Ondrich, and Gyan Sinha. Ruthless prepayment? Evidence from multifamily mortgages. *Journal of Urban Economics*, 41:78–101, 1997.

[65] Qiang Fu, Michael LaCour-Little, and Kerry D. Vandell. Commercial mortgage prepayments under heterogenous prepayment penalty structures. *Journal of Real Estate Research*, forthcoming, 2003.

[66] Alois L. J. Geyer and Stefan Pichler. A state-space approach to estimate and test multifactor cox-ingersoll-ross models of the term structure. *The Journal of Financial Research*, 22(1):107–130, 1999.

[67] Kay Giesecke. Default compensator, incomplete information, and the term structure of credit spreads. Working Paper, August 2002.

[68] Kay Giesecke. Credit risk modelling and valuation: An introduction. Working Paper, Cornell University, January 2003.

[69] Paul Glasserman. *Monte Carlo Methods on Financial Engineering*, volume 53 of *Applications of Mathematics: Stochastic Modelling and Applied Probability.* New York: Springer-Verlag, 2003.

[70] N. Gordon, D. J. Salmond, and C. M. Ewing. Bayesian state estimation for tracking and guidance using the bootstrap filter. *AIAA Journal of Guidance, Control and Dynamics*, 18:1434–1443, 1995.

[71] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. A novel approach to non-linear and non-Gaussian Bayesian state estimation. *IEE-Proceedings Part F*, 140:107–133, 1993.

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

[73] Arden R. Hall. Valuing the mortgage borrower's prepayment option. *AREUEA Journal*, 13(3):229–247, 1985.

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

[75] Oskar Ragnar Harmon. A hazard rate analysis of single-family conventional mortgage loan delinquency payment patterns. *Housing Finance Review*, 8:291–303, 1989.

[76] J. Heckman and B. Singer. A method for minimizing the effect of distributional assumptions in econometric models for duration data. *Econometrica*, 52(2):271–320, March 1984.

[77] Wenyi Huang and Jan Ondrich. Stay, pay, or walk away: A hazard rate analysis of FHA-insured multifamily mortgage terminations. *Journal of Housing Research*, 13(1):85–117, 2002.

[78] M. Izard and A. Blake. *Computer Vision – ECCV'96*, chapter Contour Tracking by Stochastic Propagation of Conditional Density. New York: Springer, 1996.

[79] Jean Jacod and Albert N. Shiryaev. *Limit Theorems for Stochastic Processes.* A Series of Comprehensive Studies in Mathematics. Springer-Verlag, 1987.

[80] Robert A. Jarrow, David Lando, and Stuart M. Turnbull. A Markov model for the term structure of credit spreads. *Review of Financial Studies*, 10(2):481–523, 1997.

[81] Robert A. Jarrow, David Lando, and Fan Yu. Default risk and diversification: Theory and applications. Working Paper, Cornell University, 2003.

[82] Robert A. Jarrow and Stuart M. Turnbull. Pricing options on financial securities subject to default risk. *Journal of Finance*, 50(1):53–86, 1995.

[83] Narashimahn Jegadeesh and Xiongwei Ju. A non-parametric prepayment model and valuation of mortgage-backed securities. *Journal of Fixed Income*, pages 50–67, June 2000.

[84] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley and Sons, second edition edition, 2002.

[85] E. L. Kaplan and Paul Meier. Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, June 1958.

[86] James B. Kau and Donald C. Keenan. An overview of the option theoretic pricing of mortgages. *Journal of Housing Research*, 6(2):217–244, 1995.

[87] James B. Kau, Donald C. Keenan, Walter J. Muller III, and James F. Epperson. Rational pricing of adjustable rate mortgages. *AREUEA Journal*, 13:117–128, 1985.

[88] James B. Kau, Donald C. Keenan, Walter J. Muller III, and James F. Epperson. The valuation and securitization of commercial and multifamily mortgages. *Journal of Banking and Finance*, 11:1525–546, 1987.

[89] James B. Kau, Donald C. Keenan, Walter J. Muller III, and James F. Epperson. Pricing commercial mortgages and their mortgage-backed securities. *Journal of Real Estate Finance and Economics*, 3(4):333–356, 1990.

[90] James B. Kau, Donald C. Keenan, Walter J. Muller III, and James F. Epperson. The valuation and analysis of adjustable rate mortgages. *Management Science*, 36(12):1417–1431, 1990.

[91] James B. Kau, Donald C. Keenan, Walter J. Muller III, and James F. Epperson. A generalized valuation model for fixed-rate residential mortgages. *Journal of Money, Credit and Banking*, 24(3):279–299, Aug. 1992.

[92] James B. Kau, Donald C. Keenan, Walter J. Muller III, and James F. Epperson. Option theory and floating-rate securities with a comparison of adjustable and fixed-rate mortgages. *Journal of Business*, 66(4):595–618, 1993.

[93] James B. Kau, Donald C. Keenan, and Taewon Kim. Default probabilities of mortgages. *Journal of Urban Economics*, 35(3):278–296, Aug. 1994.

[94] Nicholas M. Keifer. Economic duration data and hazard functions. *Journal of Economic Literature*, 26(2):646–679, June 1988.

[95] Masaaki Kijima. *Stochastic Processes with Applications to Finance*. Chapman and Hall, 2003.

[96] A. Kong, J. S. Lui, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89:278–288, 1994.

[97] Tony Lancaster. *The Economic Analysis of Transition Data*. Number 17 in Econometric Society Monographs. Cambridge University Press, 1990.

[98] David Lando. *Mathematics of Derivative Securities*, chapter Modelling Bonds and Derivatives with Default Risk, pages 369–393. Cambridge University Press, 1997.

[99] David Lando. On Cox processes and credit risky securities. *Review of Derivatives Research*, 2(1):99–120, 1998.

[100] Wai K. Leung and C. F. Sirmans. A lattice approach to fixed-rate mortgage pricing with default and prepayment options. *AREUEA Journal*, 18(1):91–104, 1990.

[101] J. S. Lui. Metropolizaed independent sampling with comparison to rejection sampling and importance sampling. *Statistics and Computing*, 6:113–119, 1996.

[102] J. S. Lui and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.

[103] D. Madan and H. Unal. Pricing the risk of default. *Review of Derivatives Research*, 2:121–160, 1998.

[104] Achla Marathe and Hany A. Shawky. The structural relation between mortgage and market interest rates. *Journal of Business Finance and Accounting*, 30(9-10):1235–1251, Nov.-Dec. 2003.

[105] Brian P. McCall. Unemployment insurance rules, joblessness, and part-time work. *Econometrica*, 64(3):647–682, May 1996.

[106] Brian P. McCall. The determinants of full-time versus part-time reemployment following job displacement. *Journal of Labor Economics*, 15(4):714–734, 1997.

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

[108] Bruce D. Meyer. Unemployment insurance and unemployment spells. *Econometrica*, 58(4):757–782, July 1990.

[109] Wiji Narendranathan and Mark B. Stewart. Modelling the probability of leaving unemployment: Competing risks models with flexible base-line hazards. *Applied Statistics*, 42(1):63–83, 1993.

[110] Andrey D. Pavlov. Competing risks of mortgage termination: Who refinances, who moves, and who defaults? *Journal of Real Estate Finance and Economics*, 23(2):185–210, 2001.

[111] N. D. Pearson and T. Sun. Exploiting the conditional density in estimating the term structure: An application to the Cox, Ingersoll and Ross model. *Journal of Finance*, 49:1279–1304, 1994.

[112] M. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94:590–599, 1999.

[113] Michael K. Pitt. Smooth particle filters for likelihood evaluation and maximization. Working Paper, July 2002.

[114] Randell J. Pozdena and Ben Iben. Pricing mortgages: An options approach. *Economic Review*, 2:39–55, 1984.

[115] R. L. Prentice and L. A. Gloecker. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34(1):57–67, March 1978.

[116] Ross L. Prentice and John D. Kalbfleisch. Mixed discrete and continuous Cox regression model. *Lifetime Data Analysis*, 9:195–210, 2003.

[117] William H. Press, Saul T. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, volume 1. Cambridge University Press, 1999.

[118] John M. Quigley and Robert Van Order. Efficiency on the mortgage market: the borrower's perspective. *AREUEA Journal*, 18(3):237–354, 1990.

[119] John M. Quigley and Robert Van Order. Explicit tests of contingent claims models of mortgage default. *Journal of Real Estate Finance and Economics*, 11(2):99–117, 1995.

[120] Scott F. Richard and Richard Roll. Prepayments on fixed-rate mortgage-backed securities. *Journal of Portfolio Management*, 15(3):73–82, Jan. 1989.

[121] Geert Ridder and Insan Tunali. Stratified partial likelihood estimation. *Journal of Econometrics*, 92:193–232, 1999.

[122] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley, 1987.

[123] Philipp J. Schönbucher. *Credit Derivatives Pricing Models: Model, Pricing and Implementation.* John Wiley and Sons, 2003.

[124] Eduardo S. Schwartz and Walter N. Torous. Prepayment and the valuation of mortgage-backed securities. *The Journal of Finance*, 44(2):375–392, 1989.

[125] Eduardo S. Schwartz and Walter N. Torous. *Financial Markets and Financial Crises*, chapter Caps of Adjustable-Rate Mortgages: Valuation, Insurance, and Hedging. University of Chicago Press, 1991.

[126] Eduardo S. Schwartz and Walter N. Torous. Prepayment, default, and the valuation of mortgage pass-through securities. *Journal of Business*, 65(2):221–239, 1992.

[127] Richard Stanton. Rational prepayment and value of mortgage-backed securities. *The Review of Financial Studies*, 8:677–708, 1995.

[128] Richard Stanton and Nancy Wallace. Mortgage choice: What is the point? *Real Estate Economics*, 26(2):173–205, 1998.

[129] Glenn T. Sueyoshi. Semiparametric proportional hazards estimation of competing risks models with time-varying covariates. *Journal of Econometrics*, 51:25–28, 1992.

[130] Glenn T. Sueyoshi. A class of binary response models for grouped duration data. *Journal of Applied Econometrics*, 10:411–431, 1995.

[131] Sheridan Titman and Walter Torous. Valuing commercial mortgages: An empirical investigation of the contingent-claims approach to pricing risky debt. *The Journal of Finance*, 44(2):345–373, June 1989.

[132] G.J. van den Berg. *Handbook of Econometrics*, volume 5 of *Handbooks in Econometrics 2*, chapter Duration Models Specification, Identification and Multiple Durations, pages 3380–3460. Elsevier Science B.V., 2001.

[133] Kerry Vandell, Walter Barnes, David Hartzell, Dennis Kraft, and William Wendt. Commercial mortgage defaults: Proportional hazard estimation using individual loan histories. *Journal of American Real Estate and Urban Economics Association*, 20(4):451–480, 1993.

[134] Kerry D. Vandell and Thomas Thibodeau. Estimation of mortgage defaults using disaggregate loan history data. *AREUEA Journal*, 13(3):292–316, 1985.

[135] Y. H. Wang. Coupling methods in approximations. *Canadian Journal of Statistics*, 14:69–74, 1986.

# Appendix A

## Pseudo-Code for the SIR Particle Filter

**Simulated Maximum Likelihood Estimation Using Sampling/Importance Resampling (SIR) Filter of Gordon et al. [71], with suggestions from Pitt [113].**

Choose between default and prepayment (use either (5.14) or (5.16) as a transition equation):

```
l=DEFAULT or l=PREPAYMENT
```

Allocate the filter depending on how many particles you want to use:

```
M=PARAMETER, R=PARAMETER, ALLOCATE(PARTICLES,WEIGHTS)
```

Initialize the filter: Set $\{\lambda_0^{l(m)}(t(0))\}_{m=1}^M$ (sample mean intensity, some prior guess, or a parameter to be estimated):

```
$\{\lambda_0^{l(m)}(t(0))\}_{m=1}^M$ = MEAN($\{\lambda_0^{l(m)}\}_{m=1}^M$) or
($\{\lambda_0^{l(m)}(t(0))\}_{m=1}^M$) = PARAMETER or
$\{\lambda_0^{l(m)}(t(0))\}_{m=1}^M = \Theta($# OF OPTIMIZATION PARAMETERS$+$1)
```

Begin likelihood maximization (BFGS or Newton-Raphson) with an initial guess about the parameter vector $\Theta = \Theta_0$:

```
LIKELIHOOD($\lambda_0^l(t(0)), \Theta$)
1 WHILE k $\leq$ (MAXIMUM # OF ITERATIONS) or until CONVERGENCE
    Set the seed for the random number generator (using RAN1):
    SEED=PARAMETER
    Loop through observations. For each observation, keep track of the observation time indicator $i$ and the stratum indicator $j$:
```

```
FOR n = 0 : (# OF OBSERVATIONS - 1)
    Remember to re-initialize the filter for each new series:
    IF j(n)≠j(n-1) THEN RESET($\{\lambda_0^{l(m)}(t(0))\}_{m=1}^{M}$)
    Sample $R$ particles with equal weights (use stratified sampling):
    FOR r = 1 : R
        Draw $\lambda_0^{l(r)}(t(i))$ using $\{\lambda_0^{l(m)}(t(i))\}_{m=1}^{M}$
    END FOR r
    Propagate particles according to (5.13) and either (5.14) or (5.16) with $Q$ subperiods:
    FOR r = 1 : R
        $\lambda_0^{l(r)}(t(i+1))$ = PROPAGATE($\lambda_0^{l(r)}(t(i)), Q$)
    END FOR r
    Get weights according to (5.10):
    FOR r = 1 : R
        $w^{l(r)}(t(i+1))$ = POISSON($y_{i+1,j}^{l(r)}, \mu_{i+1,j}^{l(r)}$)
    END FOR r
    Compute likelihood contribution of observation n+1 given either by logarithm of (5.9) or, if desired, by (5.19):
    MEAN = MEAN($\{\lambda_0^{l(r)}(t(i+1))\}_{r=1}^{R}$)
    VAR = VARIANCE($\{\lambda_0^{l(r)}(t(i+1))\}_{r=1}^{R}$)
    $f_{n+1}$ = CONTRIBUTION(MEAN,VAR)
    Compute normalized weights of particles:
    SUM = SUM($\{\lambda_0^{l(r)}(t(i+1))\}_{r=1}^{R}$)
    FOR r = 1 : R
        $\pi^{l(r)}(t(i+1))$ = $\frac{w^{l(r)}(t(i+1))}{\texttt{SUM}}$
    END FOR r
    Sort particles in ascending order. Use smooth stratified resampling to avoid
    degeneracy and non-continuity problems:
    $\{\lambda_0^{l(r)}(t(i+1))\}_{r=1}^{R}$ = SORT$\{\lambda_0^{l(r)}(t(i+1))\}_{r=1}^{R}$
    FOR m = 1 : M
       Draw $\lambda_0^{l(m)}(t(i+1))$ using $\{\lambda_0^{l(r)}(t(i+1))\}_{r=1}^{R}$, $\pi^{l(r)}(t(i+1))$
    END FOR m
  END FOR n
  Compute total likelihood as given by (5.19)
  Compute gradient(numerical)
  IF CONVERGED
    GO TO 2
  ELSE
    MODIFY $\Theta$ and GO TO 1
  END IF
END WHILE k
Display results, including parameter estimates $\hat{\Theta}$ and their standard errors (square root of the diagonal of the inverted Hessian)
2 TERMINATE
```