risks

MDPI

Article

Limiting Loss Distribution of Default and Prepayment for Loan Portfolios and Its Application in RMBS

Chenxi Xia [1], Xin Zang [2,*], Lan Bu [3], Qinhan Duan [1] and Jingping Yang [1]

[1] School of Mathematical Sciences, Peking University, Beijing 100871, China; xiacx@pku.edu.cn (C.X.); 2301110079@pku.edu.cn (Q.D.); yangjp@math.pku.edu.cn (J.Y.)
[2] School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China
[3] Financial Technology Thrust, Hong Kong University of Science and Technology, Guangzhou 511400, China; lanbu@hkust-gz.edu.cn
* Correspondence: xzang@bjtu.edu.cn

Abstract

This paper studies the joint distribution of the default and prepayment losses for a large portfolio of loans, based on a bottom-up approach. The repayment behaviors of loans in the portfolio are determined by both systematic and idiosyncratic risk factors and are conditionally independent given the systematic factors. The joint two-dimensional limit distributions of the portfolio default and prepayment losses are obtained, including the strong law of large numbers and the central limit theorem. A numerical study for the portfolio losses is performed for some simplified models. Finally, we conduct the empirical analysis on the residential mortgage-backed security (RMBS) based on Freddie Mac's dataset. The empirical results reveal the impacts of different factors on the default and prepayment behaviors, and the distributions of the portfolio losses are simulated based on empirical estimation results to show its difference with the log-normal distributions.

Keywords: loan portfolio; default risk; prepayment risk; limiting loss distribution; RMBS

check for updates

Academic Editors: Tak Kuen Ken Siu and Hailiang Yang

Received: 31 May 2025
Revised: 4 August 2025
Accepted: 8 August 2025
Published: 15 August 2025

Citation: Xia, Chenxi, Xin Zang, Lan Bu, Qinhan Duan, and Jingping Yang. 2025. Limiting Loss Distribution of Default and Prepayment for Loan Portfolios and Its Application in RMBS. *Risks* 13: 153. https://doi.org/10.3390/risks13080153

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

For individual customers, a mortgage and personal loan are two common options for borrowing money from the financial institutions. A mortgage is a secured loan typically used for real estate financing, while a personal loan is an unsecured loan and can be flexibly used for various purposes. If the borrower fails to repay the mortgage, the lender has the right to foreclose the loan and repossess the property. The repayment process of a given loan may be subject to both default risk and prepayment risk. Default occurs when the borrower lacks sufficient funds to meet their debt obligations, or when the property value falls below the outstanding loan balance, potentially leading to a strategic default. Prepayment may occur, for instance, when there is a drop in the market interest rates, allowing the borrower to pay off the existing loan and refinance at a lower rate. Our study focuses on a large number of similar loans, which can be pooled and securitized as a portfolio for investing and trading, e.g., an asset-backed security (ABS) consisting of a specific pool of underlying loans. The total loss of portfolio includes default loss and prepayment loss, which come from the potential losses of each loan in the pool.

In early periods, the literature on modeling the loan for individual customers mainly focused on either default or prepayment. For example, in applying the option pricing models to mortgage defaults, Cunningham and Hendershott (1984) and Epperson et al. (1985) priced the default as a put option. The motivation is that the default occurs if the

house price falls sufficiently below the mortgage value so that the decision of default follows a similar manner as the exercise of a compound European put option. Under the option pricing framework, Foster and Van Order (1984) and Quigley and Van Order (1995) empirically estimated the default models. In contrast, Dunn and McConnell (1981), Buser and Hendershott (1984), and Brennan and Schwartz (1985) priced the mortgages with prepayment risks by using the option-based approach. The introduction of the option pricing method lies in the fact that the prepayment gives the option to buy the remaining part of the loan (including the outstanding debt and prepayment cost) and therefore can be interpreted as an American call option on a bond. The estimation of prepayment models can be found in Schwartz and Torous (1989) and Quigley and Van Order (1990).

Since the 1990s, a number of works jointly modeled the default and prepayment as competing risks for the personal loan or mortgage. Deng et al. (1996) and Deng et al. (2000) extended the traditional option models to jointly consider default and prepayment as dependent risks in a proportional hazard framework for mortgages. Banasik et al. (1999) introduced the competing risks approach of survival analysis to study the prepayment and default for personal loans, where the applications of Cox's proportional hazard (CPH) model were further extended by Stepanova and Thomas (2002). In later extensions, Quercia and Spader (2008) applied the multinomial logit model to investigate the homeownership education and counseling completion on the prepayment and default. Steinbuks (2015) studied the effect of prepayment penalty restrictions on the probability of prepayment and default. Zhang et al. (2019) proposed a mixture cure PH model under competing risks for an online loan. Some techniques for estimating the competing risks models include, e.g., a maximum likelihood technique based on constrained optimization in Thackham and Ma (2022) and a nonparametric life-table method in Li et al. (2023).

The existing research on the limiting portfolio loss mainly focused on default losses. Vasicek (1991) initiated this field by deriving the asymptotic distribution of portfolio credit losses under a single-factor Gaussian copula model. Credit Suisse Financial Products (1997) introduced a Poisson mixture model framework for analyzing the limiting portfolio loss. Gordy (2003) developed a single-factor model for the limiting loss of homogeneous portfolios, providing the theoretical basis for regulatory capital framework under Basel II/III[1]. Giesecke et al. (2015) derived a law of large numbers for default losses in heterogeneous portfolios using the stochastic PDE methods. Sirignano and Giesecke (2019) proposed data-driven asymptotic approaches for the large loan pools, demonstrating their computational efficiency in computing the distributions of default rates, prepayment rates, and loss from default. In the industry, the top-down approach has been widely used to assess the limiting portfolio loss, focusing on directly modeling the distribution of total portfolio loss (e.g., the log-normal distribution for modeling the default loss in Moody's Investors Service 2024).

In this paper, we model the limiting loss distribution for a large portfolio of loans by considering both default and prepayment risk as competing events (i.e., only the first occurred event is observed) and apply the model to a specific loan dataset. Our framework adopts a bottom-up approach to start by modeling the repayment behaviors of individual loans. At each discrete payment date, the repayment of a loan may end with either one of the three events: scheduled repayment completion, default, or prepayment. Building on this structure, we develop a joint model for the probabilities and loss values of both default and prepayment. The repayment behaviors of all the loans are determined by both systematic and idiosyncratic factors and are conditionally independent given the systematic factors. Under certain regular conditions, the limit properties of portfolio loss are discussed, including the strong law of large numbers (SLLN) and the central limit theorem (CLT). Further, we conduct an empirical analysis to estimate the probabilities of default and prepayment by applying the data of residential mortgage-backed security

(RMBS). Based on the estimation results, we simulate the portfolio loss distribution, which is close to the log-normal distribution.

Our main contributions can be summarized as follows.

(a) We propose a discrete-time bottom-up portfolio loss model incorporating both default and prepayment risks. It is different from the existing frameworks for default losses in the single-period (Gordy 2003) or continuous-time (Giesecke et al. 2013) contexts. Several illustrative examples demonstrate that our discrete-time structure better captures the individual loan repayment behaviors, such as the monthly scheduled repayments and periodic event observations.

(b) Given the conditional independent structure, we investigate the limit properties of portfolio loss as the pool size tends to infinity, including the SLLN and CLT along with an upper error bound for approximating. Especially, under the assumption that the frequency of individual factors converges to a given distribution, the SLLN and CLT show that the limiting loss distribution of the portfolio is determined only by the systematic factors. We further verify this result by the numerical simulations of simplified models and find that the limiting distribution is approximately log-normal. Moreover, the copula of default and prepayment loss distributions are exhibited.

(c) Our empirical analysis applies the proposed framework to the empirical data, employing a CPH model to estimate the factor coefficients and baseline hazards. The estimated coefficients in the CPH model indicate that the systematic factors significantly influence both hazards. In terms of the influence of individual factors, the default risk is affected by loan-level credit quality, and the prepayment risk mainly depends on refinancing capacity. The out-of-sample validation reveals the good predictive performance of the model, and the estimated loss distribution is consistent with the CLT result. These findings align with the established theoretical and empirical evidence (see Calhoun and Deng 2002; Campbell and Cocco 2015). Finally, we perform the Monte Carlo simulation of portfolio default and prepayment losses based on the empirically estimated parameters. The asymptotic portfolio loss distributions display reasonable agreement with the log-normal distribution, which complies with Moody's methodology (Moody's Investors Service 2024).

The rest of this paper is organized as follows. Section 2 introduces the portfolio loss model considering both default and prepayment risk. Section 3 derives the limit properties of loss distribution under the conditional independent structure. Section 4 presents the empirical work and a further simulation study. Section 5 concludes.

2. The Portfolio Loss Model Considering Default and Prepayment

In this section, a general model is proposed for a portfolio of loans, where the repayment processes are exposed to both default risk and prepayment risk. The repayment behaviors of all loans are independent conditional on some common systematic factors, and each loan is also affected by some extra idiosyncratic factors.

2.1. The Portfolio Loss for a Pool of Loans

Building on the highly diversified characteristics of loan pools, we specify a general portfolio loss model that reflects the collective behavior of underlying assets. Assume that there are N unit-principal loans in the asset pool, which are issued at the same time and repaid periodically. For simplicity, the common scheduled periodical repayment dates of all loans are denoted as $1, 2, \ldots, K$, where $K \in \mathbb{Z}_+$ is the terminal time.

During the loan repayment process, the default or prepayment may occur on each loan. For the ith loan, let $\tau_i^{(1)} \in \mathbb{Z}_+$ and $\tau_i^{(2)} \in \mathbb{Z}_+$ be its time of potential default and

time of potential prepayment, respectively. Since the default and prepayment are mutually exclusive, we assume that they do not simultaneously happen.

Assumption 1. $\tau_i^{(1)} \neq \tau_i^{(2)}$ *a.s.*, $1 \leq i \leq N$.

Based on the potential event times $\tau_i^{(1)}, \tau_i^{(2)}$, define

$$\tau_i := \min\{\tau_i^{(1)}, \tau_i^{(2)}, K\} \text{ and } D_i := \begin{cases} 0, & \text{if } \tau_i^{(1)} > K \text{ and } \tau_i^{(2)} > K, \\ 1, & \text{if } \tau_i^{(1)} < \tau_i^{(2)} \text{ and } \tau_i^{(1)} \leq K, \\ 2, & \text{if } \tau_i^{(2)} < \tau_i^{(1)} \text{ and } \tau_i^{(2)} \leq K. \end{cases} \tag{1}$$

Here, τ_i represents the time-until-termination, which is default at $D_i = 1$, prepayment at $D_i = 2$, and normal repayment at $D_i = 0$. A model for (τ_i, D_i) belongs to the class of multiple decrement models (see, e.g., Deshmukh (2012) and Chapter 7.2 of Rotar (2014)). For the ith loan, denote by $b_i^{(1)}$ and $b_i^{(2)}$ the loss given default and loss given prepayment, which are random variables taking values in $[0, 1]$. Then, the discounted default loss and prepayment loss at time 0 are defined as

$$l_i^{(1)} := 1_{\{D_i=1\}} b_i^{(1)} e^{-r\tau_i} \text{ and } l_i^{(2)} := 1_{\{D_i=2\}} b_i^{(2)} e^{-r\tau_i} \tag{2}$$

respectively, where r is the constant risk-free interest rate. Denote by $\boldsymbol{l}_i = (l_i^{(1)}, l_i^{(2)})$ the ith loss vector. Then, the portfolio loss vector $\boldsymbol{L}_N = (L_N^{(1)}, L_N^{(2)})$ for the mortgage pool of N loans is given by

$$\boldsymbol{L}_N := \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{l}_i, \tag{3}$$

where $L_N^{(1)} = \sum_{i=1}^N l_i^{(1)}/N$ represents the portfolio default loss and $L_N^{(2)} = \sum_{i=1}^N l_i^{(2)}/N$ represents the portfolio prepayment loss, respectively. Throughout this paper, bold symbols are used for (column) vectors.

In practice, loans are classified as defaults only after a period of delinquency or inactivity. While some studies model the delinquency stages when analyzing the repayment behavior, our analysis mainly focuses on the terminal loss outcomes (default/prepayment). Since the observed default losses include those incurred during the delinquency periods, we adopt a single-step default framework for simplicity.

Remark 1. *We develop a discrete-time multi-period model, which naturally incorporates discounting by the risk-free interest rate. Notably, when $K = 1$, the model reduces to a single-period framework where we can set $r = 0$ (i.e., no discounting), making it consistent with many classical single-period models in the literature (i.e., Gordy 2003; Vasicek 1991).*

2.2. The Correlation Structure for the Portfolio Loss

In the following, we assume that the cash flow of payments for the N loans are influenced by some common systematic and macroeconomic factors $\boldsymbol{Y}$, e.g., refinancing interest rate, GDP growth rate, unemployment rate, and industry-performance factors, which can vary through time. For $1 \leq i \leq N$, the repayment behavior of the ith loan is also determined by some extra idiosyncratic factors $\boldsymbol{x}_i$, which remain unchanged through time. Here, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ have the same category and dimension. For example, Sirignano et al. (2016) and Sirignano and Giesecke (2019) consider the credit score, LTV ratio, initial loan rate, type of loan, collateral type, and geographic location as the individual factors. Regarding the factors $\boldsymbol{Y}$ and $\boldsymbol{x}_i$, we make the following two assumptions.

Assumption 2. *For $1 \leq i \leq N$, $j = 1, 2$, and $1 \leq k \leq K$, there exist continuous functions $q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$, $b^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$, and $B^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$ such that*

$$\mathbb{P}(\tau_i^{(j)} = k | \tau_i > k-1, \boldsymbol{Y} = \boldsymbol{y}) = q^{(j)}(k, \boldsymbol{x}_i, \boldsymbol{y}), \tag{4}$$

$$\mathbb{E}[b_i^{(j)} | \tau_i = k, D_i = j, \boldsymbol{Y} = \boldsymbol{y}] = b^{(j)}(k, \boldsymbol{x}_i, \boldsymbol{y}), \tag{5}$$

$$\mathbb{E}[(b_i^{(j)})^2 | \tau_i = k, D_i = j, \boldsymbol{Y} = \boldsymbol{y}] = B^{(j)}(k, \boldsymbol{x}_i, \boldsymbol{y}). \tag{6}$$

Assumption 3. *The N random vectors $(\tau_i^{(1)}, \tau_i^{(2)}, b_i^{(1)}, b_i^{(2)}), 1 \leq i \leq N$ are mutually independent conditional on $\boldsymbol{Y}$.*

Assumption 2 means that the loss probabilities of the ith loan are fully determined by the individual factors $\boldsymbol{x}_i$ and systematic factors $\boldsymbol{Y}$. Given $\boldsymbol{Y} = \boldsymbol{y}$, the repayment behavior of the ith loan is illustrated in Figure 1. At each time $k = 1, \ldots, K$, the ith loan is exposed to the default risk and prepayment risk, with the probability $q^{(1)}(k, \boldsymbol{x}_i, \boldsymbol{y})$ of default and the probability $q^{(2)}(k, \boldsymbol{x}_i, \boldsymbol{y})$ of prepayment. Although the actual loss values are not uniquely determined by the individual factors $\boldsymbol{x}_i$ and systematic factors $\boldsymbol{Y}$, the first and second moments of the loss distribution (equivalently, its expectation and variance) are fully specified by these factors.

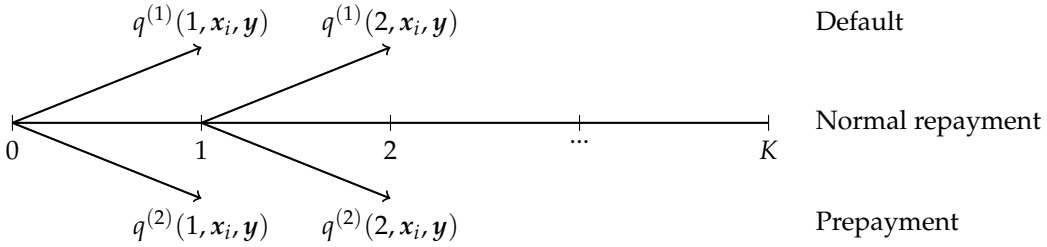


Figure 1. The repayment behavior of the ith loan.

Assumption 3 implies that the repayment behaviors of the N loans are independent conditional on the systematic factors $\boldsymbol{Y}$, which is a standard modeling approach in the literature (see Credit Suisse Financial Products 1997; Gordy 2003). Note that for some $1 \leq i \leq N$, the components in $(\tau_i^{(1)}, \tau_i^{(2)}, b_i^{(1)}, b_i^{(2)})$ need not be conditionally independent.

Some commonly used models for $q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y}), j = 1, 2$ in the literature are listed in Example 1.

Example 1. *Three different models for $q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y}), j = 1, 2$ are presented in Table 1. While the predictions of default and prepayment probabilities for the linear regression model may fall outside the range $[0, 1]$, this limitation is resolved by the subsequent use of the multinomial logistic model (Calhoun and Deng 2002). Additionally, the proportional hazard model incorporates the baseline hazard functions $q_0^{(j)}(k)$, which capture the time-dependent structure of the baseline risk for each event j at the time period k.*

Table 1. Some models for $q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$ in the literature.

Model	Expression of $q^{(j)}(k, \boldsymbol{x}_i, \boldsymbol{y})$
Linear regression model (Quercia and Stegman 1992)	$\beta_0^{(j)}(k) + \boldsymbol{\beta}_X^{(j)}(k)' \boldsymbol{x}_i + \boldsymbol{\beta}_Y^{(j)}(k)' \boldsymbol{y}$
Multinomial logistic model (Calhoun and Deng 2002)	$\frac{\exp(\beta_0^{(j)}(k) + \beta_X^{(j)}(k)' \boldsymbol{x}_i + \beta_Y^{(j)}(k)' \boldsymbol{y})}{1 + \sum_{j=1}^{2} \exp(\beta_0^{(j)}(k) + \beta_X^{(j)}(k)' \boldsymbol{x}_i + \beta_Y^{(j)}(k)' \boldsymbol{y})}$
Proportional hazard model (Deng et al. 2000; Li et al. 2023)	$q_0^{(j)}(k) \exp(\boldsymbol{\beta}_X^{(j)\prime} \boldsymbol{x}_i + \boldsymbol{\beta}_Y^{(j)\prime} \boldsymbol{y})$

Moreover, some widely used models for $b_i^{(j)}$, $b^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$, and $B^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$ in Assumption 2 are given below.

Example 2. *Assume that each loan is repaid by equal installments. For the ith loan, denote by r_i its fixed loan rate, R_i the fixed periodic payment for K periods, and $M_i(k)$ the outstanding balance at time k, $1 \le k \le K$. By the discounted cash flow formula[2], we have*

$$R_i = \frac{e^{r_i}(1 - e^{-r_i})}{1 - e^{-Kr_i}} \text{ and } M_i(k) = \frac{e^{r_i}(1 - e^{-(K-k+1)r_i})}{1 - e^{-Kr_i}}. \tag{7}$$

Note that r_i is one of the individual factors in $\boldsymbol{x}_i$, so both R_i and $M_i(k)$ in (7) are functions of $\boldsymbol{x}_i$.

For the ith loan, the default loss at time k is a percentage of the outstanding balance $M_i(k)$ (Flores et al. 2010). As modeled in Moody's Mortgage Portfolio Analyzer (MPA) methodology (Stein et al. 2011), the loss $b_i^{(1)}$ at time k is given by

$$b_i^{(1)} = M_i(k) \cdot LGD_i \text{ and } LGD_i \sim Beta(\alpha(\boldsymbol{x}_i, \boldsymbol{y}), \beta(\boldsymbol{x}_i, \boldsymbol{y}))$$

conditional on $\boldsymbol{Y} = \boldsymbol{y}$, where LGD_i is a Beta-distributed ratio of loss given default (LGD) with parameters $\alpha(\boldsymbol{x}_i, \boldsymbol{y})$ and $\beta(\boldsymbol{x}_i, \boldsymbol{y})$. Consequently, we have

$$b^{(1)}(k, \boldsymbol{x}_i, \boldsymbol{y}) = M_i(k) \frac{\alpha(\boldsymbol{x}_i, \boldsymbol{y})}{\alpha(\boldsymbol{x}_i, \boldsymbol{y}) + \beta(\boldsymbol{x}_i, \boldsymbol{y})}$$

and

$$B^{(1)}(k, \boldsymbol{x}_i, \boldsymbol{y}) = M_i^2(k) \frac{\alpha(\boldsymbol{x}_i, \boldsymbol{y})(\alpha(\boldsymbol{x}_i, \boldsymbol{y}) + 1)}{(\alpha(\boldsymbol{x}_i, \boldsymbol{y}) + \beta(\boldsymbol{x}_i, \boldsymbol{y}))(\alpha(\boldsymbol{x}_i, \boldsymbol{y}) + \beta(\boldsymbol{x}_i, \boldsymbol{y}) + 1)}.$$

Alternatively, the prepayment losses are determined by interest rate differentials (Jones and Chen 2016; Richard and Roll 1989). For the ith loan, the prepayment loss at time k is given by

$$b_i^{(2)} = b^{(2)}(k, \boldsymbol{x}_i, \boldsymbol{y}) = R_i \sum_{h=0}^{K-k} e^{-hr} - M_i(k) = \frac{1 - e^{-(K-k+1)r}}{1 - e^{-r}} R_i - M_i(k) \tag{8}$$

conditional on $\boldsymbol{Y} = \boldsymbol{y}$, where $R_i \sum_{h=0}^{K-k} e^{-hr}$ represents the risk-free discounted value of the unpaid balance at time k, and $M_i(k)$ is the outstanding loan balance in (7). Consequently, we have $B^{(2)}(k, \boldsymbol{x}_i, \boldsymbol{y}) = b^{(2)}(k, \boldsymbol{x}_i, \boldsymbol{y})^2$ with $b^{(2)}(k, \boldsymbol{x}_i, \boldsymbol{y})$ in (8).

Denote

$$S(k, \boldsymbol{x}, \boldsymbol{y}) := \prod_{h=1}^{k} \left(1 - q^{(1)}(h, \boldsymbol{x}, \boldsymbol{y}) - q^{(2)}(h, \boldsymbol{x}, \boldsymbol{y})\right), \quad 0 \le k \le K. \tag{9}$$

In the following, we derive some results for the conditional distribution of $(\tau_i, D_i), 1 \le i \le N$ defined in (1).

Lemma 1. *Suppose that Assumptions 1–3 hold.*

(i) The conditional survival function of the ith loan at time k is

$$\mathbb{P}(\tau_i > k | \boldsymbol{Y} = \boldsymbol{y}) = S(k, \boldsymbol{x}_i, \boldsymbol{y}), \quad 1 \le k \le K - 1, \tag{10}$$

where $S(k, \boldsymbol{x}, \boldsymbol{y})$ is defined in (9).

(ii) The conditional joint distribution of (τ_i, D_i) *is given by*

$$\mathbb{P}(\tau_i = k, D_i = j|\boldsymbol{Y} = \boldsymbol{y}) = \begin{cases} 0, & \text{if } j = 0,\ 1 \le k \le K-1, \\ S(K, \boldsymbol{x}_i, \boldsymbol{y}), & \text{if } j = 0,\ k = K, \\ S(k-1, \boldsymbol{x}_i, \boldsymbol{y}) q^{(j)}(k, \boldsymbol{x}_i, \boldsymbol{y}), & \text{if } j = 1, 2,\ 1 \le k \le K, \end{cases}$$

where $q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$ *is defined in* (4).

For later use, below we derive the expectations and variances of $\boldsymbol{l}_i$ and $\boldsymbol{L}_N$ conditional on $\boldsymbol{Y} = \boldsymbol{y}$, as functions of $q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$, $b^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$, and $B^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$. For convenience, define

$$\boldsymbol{m}(\boldsymbol{x}, \boldsymbol{y}) = (m_1(\boldsymbol{x}, \boldsymbol{y}), m_2(\boldsymbol{x}, \boldsymbol{y})) \text{ and } \boldsymbol{V}(\boldsymbol{x}, \boldsymbol{y}) = (v_{i,j}(\boldsymbol{x}, \boldsymbol{y}))_{2\times 2} \tag{11}$$

with

$$m_j(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} S(k-1, \boldsymbol{x}, \boldsymbol{y}) q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y}) b^{(j)}(k, \boldsymbol{x}, \boldsymbol{y}) e^{-kr},\ j = 1, 2, \tag{12}$$

$$v_{j,j}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} S(k-1, \boldsymbol{x}, \boldsymbol{y}) q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y}) B^{(j)}(k, \boldsymbol{x}, \boldsymbol{y}) e^{-2kr} - m_j(\boldsymbol{x}, \boldsymbol{y})^2,\ j = 1, 2, \tag{13}$$

$$v_{1,2}(\boldsymbol{x}, \boldsymbol{y}) = v_{2,1}(\boldsymbol{x}, \boldsymbol{y}) = -m_1(\boldsymbol{x}, \boldsymbol{y}) m_2(\boldsymbol{x}, \boldsymbol{y}).$$

Proposition 1. *Suppose that Assumptions 1–3 hold, and then we have*

$$\mathbb{E}[\boldsymbol{l}_i|\boldsymbol{Y} = \boldsymbol{y}] = \boldsymbol{m}(\boldsymbol{x}_i, \boldsymbol{y}) \text{ and } \mathrm{Var}[\boldsymbol{l}_i|\boldsymbol{Y} = \boldsymbol{y}] = \boldsymbol{V}(\boldsymbol{x}_i, \boldsymbol{y}), \tag{14}$$

with $\boldsymbol{m}(\boldsymbol{x}, \boldsymbol{y})$ *and* $\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{y})$ *defined in (11). Moreover,*

$$\mathbb{E}[\boldsymbol{L}_N|\boldsymbol{Y} = \boldsymbol{y}] = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}(\boldsymbol{x}_i, \boldsymbol{y}) \text{ and } \mathrm{Var}[\boldsymbol{L}_N|\boldsymbol{Y} = \boldsymbol{y}] = \frac{1}{N^2}\sum_{i=1}^{N} \boldsymbol{V}(\boldsymbol{x}_i, \boldsymbol{y}). \tag{15}$$

Proposition 1 lays a basic setup for deriving the limit distributions in Section 3. Some specific models are introduced in Section 3.3.

Remark 2. *To better reflect the influence of default and prepayment behaviors, when calculating the discounted loss amounts from default and prepayment in Proposition 1, we employ a constant market interest rate. The extensions to the context of stochastic interest rate can be also similarly considered.*

3. The Limiting Distribution of the Portfolio Loss

In this section, we study the limit properties of the portfolio loss $\boldsymbol{L}_N$ as $N \to \infty$, including the law of large numbers, the central limit theorem, and an upper bound on the error term. For technical uses, two assumptions are introduced below.

Assumption 4. *There exists a distribution function* $F_{\boldsymbol{X}}$*, such that*

$$F_{\boldsymbol{X}}^{(N)}(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N} 1_{\{\boldsymbol{x}_i \le \boldsymbol{x}\}} \xrightarrow{w} F_{\boldsymbol{X}}(\boldsymbol{x}) \text{ as } N \to \infty, \tag{16}$$

where "$\xrightarrow{w}$*" means the weak convergence.*

Assumption 5. *There exists some* $\varepsilon > 0$*, such that* $\det(\mathrm{Var}[\boldsymbol{l}_i|\boldsymbol{Y}]) > \varepsilon$ *a.s.*, $1 \le i \le N$.

Assumption 4 characterizes the limiting behavior of the frequency counts of individual factors $\boldsymbol{x}_i, 1 \leq i \leq N$ in the mortgage pool. Aligning conceptually with Assumption 2.2 of Giesecke et al. (2013) and Condition 2.1 of Giesecke et al. (2015), it implies that as the pool size grows, the distribution of these factors converges to a simpler macroscopic characterization. This assumption allows us to model the overall features of the portfolio instead of focusing on the individual data characteristics under the large-sample context.

Assumption 5 imposes a lower bound on the determinant of the covariance matrix $\mathrm{Var}[\boldsymbol{l}_i|\boldsymbol{Y}]$. This condition rules out some degenerate scenarios by ensuring that (i) both default loss and prepayment loss exhibit non-trivial variability (i.e., their variances are bounded away from zero) and reflect the realistic fluctuations in loan portfolios; (ii) the losses are not perfectly linearly dependent, excluding the artificial cases where one loss becomes completely predictable given the other; and (iii) the joint distribution maintains a genuine two-dimensional randomness structure, preventing pathological concentration of risk. This mathematical formulation effectively captures the natural dispersion and imperfect dependence between default and prepayment behaviors observed practically.

3.1. The Strong Law of Large Numbers

First, the strong law of large numbers on $\boldsymbol{L}_N$ as $N \to \infty$ is provided. For ease of exposition, define

$$\boldsymbol{\mu}(\boldsymbol{y}) = \int \boldsymbol{m}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x}) \text{ and } \boldsymbol{\Sigma}(\boldsymbol{y}) = \int \boldsymbol{V}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x}), \tag{17}$$

with $\boldsymbol{m}(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{y})$ given in (11), and $F_{\boldsymbol{X}}$ introduced in Assumption 4. For later use, denote $\boldsymbol{\mu}(\boldsymbol{y}) = (\mu_1(\boldsymbol{y}), \mu_2(\boldsymbol{y}))$ and $\boldsymbol{\Sigma}(y) = (\Sigma_{i,j}(y))_{2\times 2}$.

Proposition 2. *Suppose that Assumptions 1–3 hold. Let $\boldsymbol{L}_N$ be as in (3).*

(i) As $N \to \infty$,

$$\boldsymbol{L}_N - \mathbb{E}[\boldsymbol{L}_N|\boldsymbol{Y}] \to 0, \ \mathbb{P}\text{–a.s.}, \tag{18}$$

where $\mathbb{E}[\boldsymbol{L}_N|\boldsymbol{Y} = \boldsymbol{y}]$ is given in (15).

(ii) Suppose further that Assumption 4 holds. Then, as $N \to \infty$,

$$\boldsymbol{L}_N \to \boldsymbol{\mu}(\boldsymbol{Y}), \ \mathbb{P}\text{–a.s.}, \tag{19}$$

with $\boldsymbol{\mu}(\boldsymbol{y})$ defined in (17).

Intuitively, Proposition 2 (i) states that for a portfolio with a large number of individuals, $\boldsymbol{L}_N$ converges to its conditional expectation. Note that $\mathbb{E}[\boldsymbol{L}_N|\boldsymbol{Y} = \boldsymbol{y}]$ in (15) is determined by both systematic factors $\boldsymbol{Y}$ and individual factors $\boldsymbol{x}_i, 1 \leq i \leq N$. Compared to (i), Proposition 2 (ii) establishes a simpler, almost sure limit of $\boldsymbol{L}_N$ in terms of the limit distribution $F_{\boldsymbol{X}}$. As shown in (19), the limit $\boldsymbol{\mu}(\boldsymbol{Y})$ depends only on the vector of systematic factors $\boldsymbol{Y}$, which requires less underlying information and provides greater computational efficiency via the simplified expression.

Recall that $\boldsymbol{\mu}(\boldsymbol{y}) = (\mu_1(\boldsymbol{y}), \mu_2(\boldsymbol{y}))$. The following examples demonstrate some straightforward applications of Proposition 2 to the cases where the dimension of $\boldsymbol{Y}$ is one or two.

Example 3. *(i) Suppose that $\boldsymbol{Y}$ is a one-dimensional random variable (denoted by Y), $\mu_1(y)$ is a continuous increasing function, and $\mu_2(y)$ is a continuous decreasing function. Then, we have*

$$\lim_{N\to\infty} \mathbb{P}(L_N^{(1)} \leq x) = \mathbb{P}(\mu_1(Y) \leq x) = \mathbb{P}(Y \leq \mu_1^{-1}(x)) = F_Y(\mu_1^{-1}(x))$$

and

$$\lim_{N\to\infty} \mathbb{P}(L_N^{(2)} \le x) = \mathbb{P}(\mu_2(Y) \le x) = \mathbb{P}(Y \ge \mu_2^{-1}(x)) = 1 - F_Y(\mu_2^{-1}(x)),$$

where F_Y is the distribution function of Y. In this setting, $L_N^{(1)}$ and $L_N^{(2)}$ nearly exhibit the counter-monotonicity as $N \to \infty$. This is because the limit of $L_N^{(1)}$ increases in Y, while the limit of $L_N^{(2)}$ decreases in Y.

(ii) Suppose that $\boldsymbol{Y}$ is a two-dimensional random vector and $\boldsymbol{\mu}(\boldsymbol{y}) : \mathbb{R}^2 \to \mathbb{R}^2$ is a continuous differentiable bijection. Then, the limit of probability density function of $\boldsymbol{L}_N$ as $N \to \infty$ can be approximated by

$$f_{\boldsymbol{\mu}(\boldsymbol{Y})}(x_1, x_2) = f_{\boldsymbol{Y}}(\boldsymbol{\mu}^{-1}(x_1, x_2)) \frac{\partial^2(\boldsymbol{\mu}^{-1}(x_1, x_2))}{\partial x_1 \partial x_2},$$

where $\boldsymbol{\mu}^{-1}(x_1, x_2)$ denotes a vector (y_1, y_2) satisfying $\boldsymbol{\mu}(y_1, y_2) = (x_1, x_2)$.

For a random variable X, the VaR of X at level $\alpha \in (0,1)$ is defined as $\mathrm{VaR}_\alpha := \inf\{x \in \mathbb{R} : \mathbb{P}(X \le x) \ge \alpha\}$. The result (19) leads to the following corollary on the VaR of $L_N^{(j)}, j = 1, 2$ as $N \to \infty$.

Corollary 1. *Suppose that Assumptions 1–4 hold. For any $\alpha \in (0,1)$ and $\varepsilon > 0$, we have*

$$\lim_{N\to\infty} \mathbb{P}(L_N^{(j)} \le \mathrm{VaR}_\alpha(\mu_j(\boldsymbol{Y})) + \varepsilon) \ge \alpha, \ j = 1, 2,$$

and

$$\lim_{N\to\infty} \mathbb{P}(L_N^{(j)} \le \mathrm{VaR}_\alpha(\mu_j(\boldsymbol{Y})) - \varepsilon) \le \alpha, \ j = 1, 2.$$

Given Corollary 1, the VaR of $\boldsymbol{\mu}(\boldsymbol{Y})$ can be applied to approximate the VaR of $L_N^{(j)}, j = 1, 2$ for large N.

3.2. The Central Limit Theorem

In the following, a central limit theorem (CLT) is provided for $\boldsymbol{L}_N$ in (3). Denote by $\mathbb{I}_n$ the $n \times n$ identity matrix. For a positive-definite matrix A, $A^{-1/2}$ is defined as its inverse square root via spectral decomposition; see the footnote for details.[3]

Theorem 1. *Suppose that Assumptions 1–3 and 5 hold.*

(i) Conditional on $\boldsymbol{Y} = \boldsymbol{y}$,

$$\boldsymbol{\Sigma}_N(\boldsymbol{y})^{-\frac{1}{2}}(\boldsymbol{L}_N - \boldsymbol{\mu}_N(\boldsymbol{y})) \xrightarrow{d} \mathcal{N}(0, \mathbb{I}_2) \tag{20}$$

as $N \to \infty$, where

$$\boldsymbol{\mu}_N(\boldsymbol{y}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{m}(\boldsymbol{x}_i, \boldsymbol{y}) \text{ and } \boldsymbol{\Sigma}_N(\boldsymbol{y}) = \frac{1}{N^2}\sum_{i=1}^{N} \boldsymbol{V}(\boldsymbol{x}_i, \boldsymbol{y}) \tag{21}$$

with $\boldsymbol{m}(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{V}(\boldsymbol{x}, \boldsymbol{y})$ given in (11).

(ii) Suppose further that Assumption 4 holds. Then, conditional on $\boldsymbol{Y} = \boldsymbol{y}$,

$$\sqrt{N}(\boldsymbol{L}_N - \boldsymbol{\mu}(\boldsymbol{y})) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{y})) \tag{22}$$

as $N \to \infty$, with $\boldsymbol{\mu}(\boldsymbol{y})$ and $\boldsymbol{\Sigma}(\boldsymbol{y})$ given in (17).

Similar to Proposition 2, $\boldsymbol{\mu}_N(\boldsymbol{y})$ and $\boldsymbol{\Sigma}_N(\boldsymbol{y})$ in Theorem 1 (i) include the information of both systematic factors $\boldsymbol{Y}$ and individual factors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, while $\boldsymbol{\mu}(\boldsymbol{y})$ and $\boldsymbol{\Sigma}(\boldsymbol{y})$ in

Theorem 1 (ii) are only determined by $\boldsymbol{Y}$ and the asymptotic distribution $F_{\boldsymbol{X}}$. The simplified expression in (ii) reduces the data requirements while enhancing the computational efficiency.

Based on Theorem 1, the following corollary further studies the CLT on the default loss $L_N^{(1)}$, the prepayment loss $L_N^{(2)}$, and the total loss $L_N^{(1)} + L_N^{(2)}$. Recall that $\boldsymbol{\Sigma}(y) = (\Sigma_{i,j}(y))_{2\times 2}$.

Corollary 2. *Suppose that Assumptions 1–5 hold. Then, conditional on $\boldsymbol{Y} = \boldsymbol{y}$, we have*

$$\sqrt{N}(L_N^{(j)} - \mu_j(\boldsymbol{y})) \xrightarrow{d} \mathcal{N}(0, \Sigma_{j,j}(\boldsymbol{y})), \quad j = 1, 2$$

and

$$\sqrt{N}(L_N^{(1)} + L_N^{(2)} - \mu_1(\boldsymbol{y}) - \mu_2(\boldsymbol{y})) \xrightarrow{d} \mathcal{N}(0, \Sigma_{1,1}(\boldsymbol{y}) + 2\Sigma_{1,2}(\boldsymbol{y}) + \Sigma_{2,2}(\boldsymbol{y})),$$

as $N \to \infty$.

Corollary 2 follows directly from Theorem 2.3 in Van der Vaart (2000) on preserving the asymptotic normality under linear transformations. The proof is therefore omitted.

Theorem 1 (i) implies that conditional on $\boldsymbol{Y} = \boldsymbol{y}$, the distribution of portfolio loss $\boldsymbol{L}_N$ with large number N can be approximated by the normal distribution $\mathcal{N}(\boldsymbol{\mu}_N(\boldsymbol{y}), \boldsymbol{\Sigma}_N(\boldsymbol{y}))$. Denote by $F_{\boldsymbol{Y}}$ the distribution function of $\boldsymbol{Y}$ and $\Phi_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}$ is the distribution function of normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, the joint distribution of $\boldsymbol{L}_N$ can be approximated by

$$\mathbb{P}(\boldsymbol{L}_N \le \boldsymbol{x}) = \mathbb{E}[\mathbb{P}(\boldsymbol{L}_N \le \boldsymbol{x}|\boldsymbol{Y})] \approx \mathbb{E}\Phi_{(\boldsymbol{\mu}_N(\boldsymbol{Y}),\boldsymbol{\Sigma}_N(\boldsymbol{Y}))}(\boldsymbol{x}) = \int \Phi_{(\boldsymbol{\mu}_N(\boldsymbol{y}),\boldsymbol{\Sigma}_N(\boldsymbol{y}))}(\boldsymbol{x}) \mathrm{d}F_{\boldsymbol{Y}}(\boldsymbol{y}). \quad (23)$$

Moreover, the error term in approximation (23) admits an explicit upper bound, as presented below.

Theorem 2. *Suppose that Assumptions 1–3 and 5 hold. The error term in (23) is bounded by*

$$\left| \mathbb{P}(\boldsymbol{L}_N \le \boldsymbol{x}) - \int \Phi_{(\boldsymbol{\mu}_N(\boldsymbol{y}),\boldsymbol{\Sigma}_N(\boldsymbol{y}))}(\boldsymbol{x}) \mathrm{d}F_{\boldsymbol{Y}}(\boldsymbol{y}) \right| \le \frac{264}{\sqrt{N\varepsilon}},$$

with ε introduced in Assumption 5.

Theorem 2 implies that the error term for the distribution function of $\boldsymbol{L}_N$ converges to zero at the rate $O(N^{-1/2})$.

3.3. Numerical Analysis for Some Models

Under certain assumptions, the limiting portfolio loss follows the SLLN in Section 3.1 and CLT in Section 3.2, respectively. Specifically, Proposition 2 states that the portfolio loss $\boldsymbol{L}_N$ can be approximated by $\boldsymbol{\mu}(\boldsymbol{Y})$ in (19). However, sometimes it is difficult to calculate $\boldsymbol{\mu}(\boldsymbol{Y})$ due to the large number of parameters in $\boldsymbol{Y}$ and the complexity of the expressions involved in $\boldsymbol{\mu}$. Hence, this part introduces two simplified models, where $\boldsymbol{\mu}(\boldsymbol{Y})$ can be directly represented.

3.3.1. The Asset Value Model

The asset value models introduced by Merton (1974) and Finger (1999) are widely used in the industry, where the loan's repayment behavior is determined by its underlying asset value (or return rate). Based on the repayment solvency considerations, we assume that the default occurs when the asset value falls below a predetermined threshold, while the prepayment is triggered when the asset value exceeds a higher threshold. For simplicity, we abstract away from temporal dynamics and consider a single-period setting ($K = 1$ and $r = 0$).

In this model, the vector $\boldsymbol{Y}$ consists of n market factors, jointly following a standard multivariate normal distribution $\mathcal{N}(0, \mathbb{I}_n)$. The vector $\boldsymbol{x}_i$ comprises the n coefficients (sensitivities) of the ith borrower to these market factors. The return rate is defined as

$$V_i = \boldsymbol{x}_i'\boldsymbol{Y} + \sqrt{1 - \boldsymbol{x}_i'\boldsymbol{x}_i} \cdot e_i \sim \mathcal{N}(0,1),$$

where $e_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ are idiosyncratic performance terms independent of $\boldsymbol{Y}$. Let c_1 and c_2 denote the critical thresholds for default and prepayment, respectively, defined in terms of the return rate. In this case, the probabilities of default and prepayment for the ith loan are given by

$$q^{(1)}(1, \boldsymbol{x}_i, \boldsymbol{y}) = \mathbb{P}(V_i < c_1 | \boldsymbol{Y} = \boldsymbol{y}) = \mathbb{P}(\boldsymbol{x}_i'\boldsymbol{y} + \sqrt{1 - \boldsymbol{x}_i'\boldsymbol{x}_i} \cdot e_i < c_1) = \Phi\left(\frac{c_1 - \boldsymbol{x}_i'\boldsymbol{y}}{\sqrt{1 - \boldsymbol{x}_i'\boldsymbol{x}_i}}\right),$$

$$q^{(2)}(1, \boldsymbol{x}_i, \boldsymbol{y}) = \mathbb{P}(V_i > c_2 | \boldsymbol{Y} = \boldsymbol{y}) = \mathbb{P}(\boldsymbol{x}_i'\boldsymbol{y} + \sqrt{1 - \boldsymbol{x}_i'\boldsymbol{x}_i} \cdot e > c_2) = \Phi\left(-\frac{c_2 - \boldsymbol{x}_i'\boldsymbol{y}}{\sqrt{1 - \boldsymbol{x}_i'\boldsymbol{x}_i}}\right).$$

For simplicity, assume the default loss is the constant $\bar{b}_1$ and the prepayment loss is the constant $\bar{b}_2$. Then, we obtain $m_j(\boldsymbol{x}, \boldsymbol{y}) = \bar{b}_j q^{(j)}(1, \boldsymbol{x}, \boldsymbol{y})$ from (12). Building on the Equation (17), the entries of $\boldsymbol{\mu}(\boldsymbol{y})$ can be expressed as

$$\mu_1(\boldsymbol{y}) = \int \bar{b}_1 q^{(1)}(1, \boldsymbol{x}, \boldsymbol{y}) dF_{\boldsymbol{X}}(\boldsymbol{x}) = \bar{b}_1 \int \Phi\left(\frac{c_1 - \boldsymbol{x}'\boldsymbol{y}}{\sqrt{1 - \boldsymbol{x}'\boldsymbol{x}}}\right) dF_{\boldsymbol{X}}(\boldsymbol{x}), \tag{24}$$

$$\mu_2(\boldsymbol{y}) = \int \bar{b}_2 q^{(2)}(1, \boldsymbol{x}, \boldsymbol{y}) dF_{\boldsymbol{X}}(\boldsymbol{x}) = \bar{b}_2 \int \Phi\left(-\frac{c_2 - \boldsymbol{x}'\boldsymbol{y}}{\sqrt{1 - \boldsymbol{x}'\boldsymbol{x}}}\right) dF_{\boldsymbol{X}}(\boldsymbol{x}). \tag{25}$$

3.3.2. The Markov Model

Consistent with the transform mechanism of the Markov model (see Jarrow et al. 1997), we assume that the probabilities of default and prepayment are fixed during the payment period, i.e., $q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y}) = \bar{q}^{(j)}(\boldsymbol{x}, \boldsymbol{y})$ for $1 \le k \le K$, $j = 1, 2$. Specifically, it follows the multinomial logistic model given by Example 1:

$$q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y}) = \bar{q}^{(j)}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\exp\{\beta_0^{(j)} + \boldsymbol{\beta}_{\boldsymbol{X}}^{(j)\prime}\boldsymbol{x} + \boldsymbol{\beta}_{\boldsymbol{Y}}^{(j)\prime}\boldsymbol{y}\}}{1 + \sum_{j=1}^{2} \exp\{\beta_0^{(j)} + \boldsymbol{\beta}_{\boldsymbol{X}}^{(j)\prime}\boldsymbol{x} + \boldsymbol{\beta}_{\boldsymbol{Y}}^{(j)\prime}\boldsymbol{y}\}}. \tag{26}$$

For simplicity, we assume the default and prepayment losses are proportional to the remaining repayment period, expressed as

$$b_i^{(j)} = b^{(j)}(k, \boldsymbol{x}, \boldsymbol{y}) = \bar{b}_j \frac{K - k + 1}{K}, \text{ given } \tau_i = k, D_i = j \tag{27}$$

for $j = 1, 2$. This linear scaling implies that losses decrease uniformly over time. We do not assume the special forms of $F_{\boldsymbol{Y}}(\boldsymbol{y})$ and $F_{\boldsymbol{X}}(\boldsymbol{x})$ in Assumption 4. Under this setup, the functions $m_j(\boldsymbol{x}, \boldsymbol{y})$, $j = 1, 2$ defined in (12) have the following specific form.

Lemma 2. *Suppose that (26) and (27) hold. Then, $m_j(\boldsymbol{x}, \boldsymbol{y})$ defined in (12) is given by*

$$m_j(\boldsymbol{x}, \boldsymbol{y}) = \frac{\bar{b}_j \bar{q}^{(j)}(\boldsymbol{x}, \boldsymbol{y})}{Ke^r} \cdot \frac{K - (K+1)a(\boldsymbol{x}, \boldsymbol{y}) + a(\boldsymbol{x}, \boldsymbol{y})^{K+1}}{(1 - a(\boldsymbol{x}, \boldsymbol{y}))^2}, j = 1, 2, \tag{28}$$

where $a(\boldsymbol{x}, \boldsymbol{y}) = (1 - \bar{q}^{(1)}(\boldsymbol{x}, \boldsymbol{y}) - \bar{q}^{(2)}(\boldsymbol{x}, \boldsymbol{y}))/e^r$.

Subsequently, the function $\boldsymbol{\mu}(\boldsymbol{y})$ is given by (21) with $\boldsymbol{m}(\boldsymbol{x}, \boldsymbol{y})$ specified in Lemma 2.

3.3.3. Simulation Results

Based on the simplified theoretical loss models, we conduct the Monte Carlo simulations to analyze the loss distributions and compare them with the log-normal distribution proposed by Moody's Investors Service (2024). For each model, we perform the simulations with a portfolio of 10,000 loans and 10,000 Monte Carlo trials. The key parameters include the payment periods K, the risk-free interest rate r, the distribution $F_{\boldsymbol{Y}}(\boldsymbol{y})$ of $\boldsymbol{Y}$ and $F_{\boldsymbol{X}}(\boldsymbol{x})$ in Assumption 4, and the functions $q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$ and $b^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$ in Assumption 2. These parameters are provided differently in the models of Sections 3.3.1 and 3.3.2. The simulation takes the following steps.

(i) The individual and systematic factors. First, we generate the individual factors $\boldsymbol{x}_i$ from the distribution $F_{\boldsymbol{X}}(\boldsymbol{x})$ and fix them in the subsequent steps. Then, in each simulation, we generate the systematic factors $\boldsymbol{Y}$ from the distribution $F_{\boldsymbol{Y}}(\boldsymbol{y})$, representing an economic scenario.

(ii) The repayment behaviors. Given the factors $\boldsymbol{x}_i$ and $\boldsymbol{Y} = \boldsymbol{y}$ in each simulation, the repayment behaviors (τ_i, D_i) are simulated as in Figure 1, with default/prepayment rates $q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$, $j = 1, 2$. For simplicity, the default/prepayment losses are fixed at their expected values $b^{(j)}(k, \boldsymbol{x}, \boldsymbol{y})$.

(iii) The portfolio losses. In each simulation, the loan losses $\boldsymbol{l}_i$, $1 \leq i \leq N$ in (2) are obtained based on the first two steps, and the portfolio loss $\boldsymbol{L}_N$ is derived by (3). By repeating this process across 10,000 economic scenarios, we generate an empirical distribution of $\boldsymbol{L}_N$.

In the following, denote the distribution function of the Beta(α, β) distribution by

$$I(x; \alpha, \beta) := \frac{1}{B(\alpha, \beta)} \int_0^x u^{\alpha-1}(1-u)^{\beta-1} \mathrm{d}u, \ 0 \leq x \leq 1, \ \alpha, \beta > 0.$$

Based on the asset value model, Figure 2 presents the simulations for the limit distribution of portfolio loss $\boldsymbol{L}_N$ under different parameter settings. The main parameters are set based on the existing literature. The default rates 8% and 5% align with the B-rated bond studies (5% in Frey and McNeil (2003), 6.25% in Gordy (2003), and 7.27% in Credit Suisse Financial Products (1997)), while the prepayment rates were assumed based on Deng et al. (2000). Since the loan recovery rate are set as 28.29%, 38.28%, and 47.54% in Credit Suisse Financial Products (1997) for different seniority levels and securities, we adopt 20% and 40%, respectively, in Figure 2. Some other parameters are set according to the asset correlations derived from Flores et al. (2010), while the rest of the parameters were selected for the purpose of simulation.

The results in Figure 2 show the clear negative correlation between the loss of default and the loss of prepayment. The empirical distributions of both default and prepayment losses in the simulated data show strong agreement with the log-normal distributions. Furthermore, the loss distribution exhibits a unimodal pattern, peaked at relatively small loss values and decreasing monotonically as the losses increase.

Under the above Markov model, Figure 3 presents the simulations illustrating the distribution of portfolio loss $\boldsymbol{L}_N$ under different parameters. The parameter selection is similar to that in Figure 2. The results show that the marginal distributions of $\boldsymbol{L}_N$ fit the log-normal distribution well and further validate the log-normal assumption in Moody's methodology (Moody's Investors Service 2024).

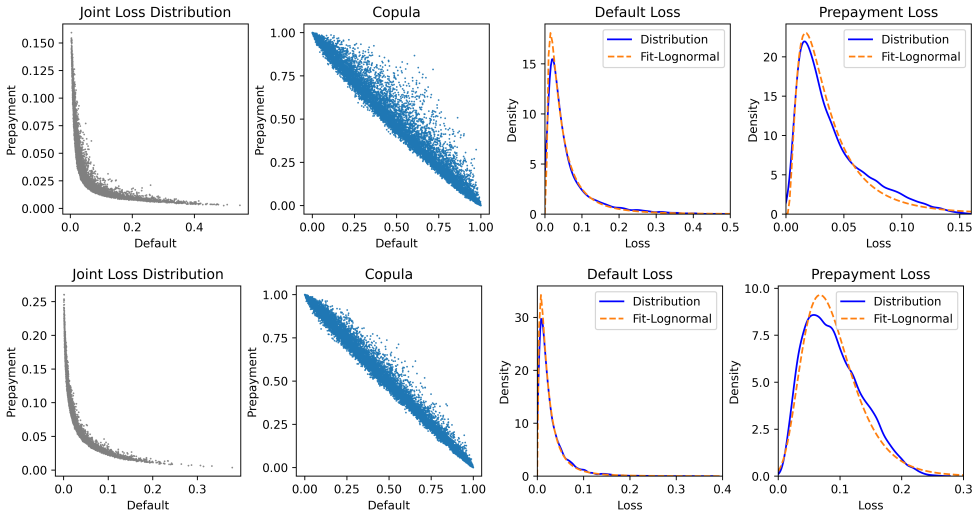


Figure 2. The joint distribution, copula, and marginal distributions of $\boldsymbol{L}_N$ for the asset value model. Note: The upper panel and lower panel exhibit the joint distribution, copula, and marginal distributions of $\boldsymbol{L}_N$ for two cases of the asset value model in (24) and (25), respectively. In the upper panel, we set $n = 2$, $\bar{b}_1 = 0.8$, $\bar{b}_2 = 0.2$, $c_1 = \Phi^{-1}(0.08) = -1.405$, $c_2 = \Phi^{-1}(0.8) = 0.842$, $F_{\boldsymbol{X}}(\boldsymbol{x}) = x_1 x_2 / 0.48, \mathbf{0} \leq \boldsymbol{x} \leq (0.8, 0.6)$, and $\boldsymbol{Y} \sim \mathcal{N}(0, \mathbb{I}_2)$. In the lower panel, we set $n = 2$, $\bar{b}_1 = 0.6$, $\bar{b}_2 = 0.3$, $c_1 = \Phi^{-1}(0.05) = -1.645$, $c_2 = \Phi^{-1}(0.7) = 0.524$, $F_{\boldsymbol{X}}(\boldsymbol{x}) = I(x_1; 2, 3) I(x_2; 4, 4)$, and $\boldsymbol{Y} \sim \mathcal{N}(0, \mathbb{I}_2)$.

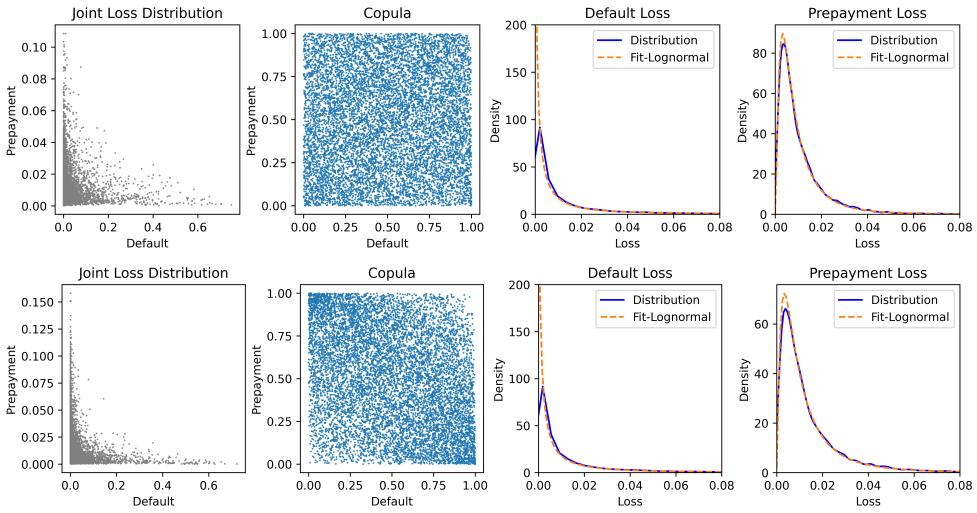


Figure 3. The joint distribution, copula, and marginal distributions of $\boldsymbol{L}_N$ for the Markov model. Note: The upper panel and lower panel exhibit the joint distribution, copula, and marginal distributions of $\boldsymbol{L}_N$ for two cases of the Markov model in (26) and (27), respectively. In both cases, we set $K = 60$, $r = 0.003$, $\bar{b}_1 = 0.8$, $\bar{b}_2 = 0.2$, and $(\beta_0^{(1)}, \boldsymbol{\beta}_X^{(1)}, \boldsymbol{\beta}_Y^{(1)}) = (-7, -2, 0.2, 1, 0.6)$. The upper panel corresponds to $(\beta_0^{(2)}, \boldsymbol{\beta}_X^{(2)}, \boldsymbol{\beta}_Y^{(2)}) = (-8, 0.2, 1, 0.8, 1)$, $F_{\boldsymbol{X}}(\boldsymbol{x}) = x_1 x_2 / 0.48, \mathbf{0} \leq \boldsymbol{x} \leq (0.8, 0.6)$, and $\boldsymbol{Y} \sim \mathcal{N}((1,1), ((1, -0.2), (-0.2, 1)))$. The lower panel corresponds to $(\beta_0^{(2)}, \boldsymbol{\beta}_X^{(2)}, \boldsymbol{\beta}_Y^{(2)}) = (-8, 0.2, 1, 1, 0.8)$, $F_{\boldsymbol{X}}(\boldsymbol{x}) = I(x_1; 2, 3) I(x_2; 4, 4)$, and $\boldsymbol{Y} \sim \mathcal{N}((1,1), ((1, 0.2), (0.2, 1)))$.

4. Empirical Studies

In this section, we present an empirical analysis of the residential mortgage-backed security (RMBS) under our setup of jointly modeling the default and prepayment. As a type of ABS, the RMBS is created by pooling numerous residential mortgage loans and converting them into tradable investment instruments. The investors can invest into the RMBS according to their personal risk preferences. Nowadays, the RMBS stands as the

largest proportion of the cash securitization market globally, especially popular in the US, continental Europe, the UK, Australia, etc.[4]

The residential mortgage loans in the RMBS typically exhibit distinct characteristics, such as fixed or adjustable interest rates, long-term maturities (e.g., 15–30 years), and amortizing payment structures. The RMBS is also exposed to the default and prepayment risks. The default is usually driven by borrower financial distress, declining property values, or macroeconomic shocks, leading to the credit losses. The prepayment is triggered by refinancing (e.g., rate declines) or home sales, altering the cash flow timing and reinvestment risk. The performances of residential mortgages in the RMBS are affected by a series of factors, including the borrower quality, underwriting guidelines, and servicer/originator quality.

As an application of our theoretical framework, the empirical study pursues two objectives. First, we incorporate both default and prepayment risks into the portfolio loss model of the RMBS and empirically investigate how the factors significantly affect the default and prepayment. Second, we use the estimated hazards to generate bottom-up Monte Carlo simulations of portfolio losses. Note that it is complex to represent the asymptotic conditional expectation $\boldsymbol{\mu}(\boldsymbol{Y}) = (\mu_1(\boldsymbol{Y}), \mu_2(\boldsymbol{Y}))$ of portfolio loss in Proposition 2. Our empirical results suggest that this conditional expectation of portfolio loss is approximately log-normally distributed, thereby enriching the theoretical model with distributional insight grounded in data.

The study employs Freddie Mac's single-family loan dataset issued in 2000 and focuses on the 15-year mortgages with monthly repayments. The dataset with full origination and monthly performance records has become a benchmark source for mortgage credit studies (see, e.g., Goodman et al. 2014; Goodman and Zhu 2015; and Bhattacharya et al. 2019). Since this standard dataset consists solely of fixed-rate loans, we restrict the analysis to this loan type. The dataset of fixed year 2000 is selected to ensure a sufficiently long repayment observation period (up to 15 years/180 months as analyzed). Robustness checks confirm that the result does not depend on the choice of the year.

4.1. Model Setup for RMBS Product

With the notation introduced in Section 2, suppose that there are N mortgages in the pool of the RMBS, with the common repayment period of K months. For each mortgage, there are three mutually exclusive states in each month: default, prepaid, or active (normal amortization). The default and prepayment risks of an individual loan are modeled by the Cox proportional hazard functions, which depend on the systematic factors $\boldsymbol{Y} = (\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(K))$ and the vector of individual factors $\boldsymbol{x}$ fixed at the issue date.

4.1.1. Systematic Factors

The systematic factors $\boldsymbol{Y} = (\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(K))$ represent a multivariate time series of macroeconomic variables. Denote the vector of systematic factors observed in the kth month by

$$\boldsymbol{Y}(k) = \big(\delta(k),\ HPC(S_1, k), \ldots, HPC(S_d, k)\big).$$

The first component $\delta(k)$ is the U.S. Treasury long-term yield rate and serves as the refinancing benchmark, which is distinct from the contractual loan rate r_i set for individual borrowers. The difference $r_i - \delta(k)$ captures the evolving interest rate spread that affects both prepayment for refinancing and default decisions. The symbols $S_1, S_2, \ldots, S_d$ represent d administrative states in the U.S., so that each mortgage in the pool is located in some state. For $1 \leq m \leq d$, $HPC(S_m, k)$ denotes the state-level house price changes (HPCs) in the state S_m at the kth month and captures the house price dynamics, defined by

$$HPC(S_m, k) = \log HPI(S_m, k) - \log HPI(S_m, k-1), \tag{29}$$

where $HPI(S_m, k)$ is the house price index (HPI) for the state S_m at the kth month. By collecting the systematic factors among $k = 1, \ldots, K$, the obtained vector $\boldsymbol{Y}$ is random with the dimension $K(d+1)$.

4.1.2. Individual Factors

For the ith loan, the individual factors are collected as

$$\boldsymbol{x}_i = (r_i,\, CS_i,\, LTV_i,\, DTI_i,\, FTB_i,\, NB_i,\, PS_i),$$

where r_i is the original loan rate, CS_i is the borrower's credit score, LTV_i is the loan-to-value ratio, DTI_i is the debt-to-income ratio, FTB_i is the indicator for first-time home-buyer (1 for yes and 0 for no), and NB_i is the number of borrowers on the loan. The final coordinate $PS_i \in \{S_1, S_2, \ldots, S_d\}$ identifies the U.S. state where the property is located. This identifier links the ith loan to the time series of state-level house price changes $HPC(PS_i, k)$.

Based on the systematic and individual factors, the competing risks (default and prepayment) are specified by the CPH model demonstrated in Table 1. The CPH framework is widely adopted in mortgage research and practice; see, for example, Deng et al. (2000), Li et al. (2023), and Moody's Mortgage Portfolio Analyzer (MPA) methodology (Stein et al. 2011). Specifically, the probabilities of default and prepayment in (4) are given by

$$\begin{aligned} q^{(j)}(k, \boldsymbol{x}_i, \boldsymbol{Y}) =& q_0^{(j)}(k) \exp\{\beta_r^{(j)}(r_i - \delta(k)) + \beta_{HPC}^{(j)} HPC(PS_i, k) + \beta_{CS}^{(j)} CS_i + \beta_{LTV}^{(j)} LTV_i \\ &+ \beta_{DTI}^{(j)} DTI_i + \beta_{FTB}^{(j)} FTB_i + \beta_{NB}^{(j)} NB_i\},\ 1 \le k \le K \end{aligned} \tag{30}$$

for $j = 1, 2$. Here, $q_0^{(j)}(k)$ is the baseline hazard function and captures the time profile of default or prepayment risk when all the explanatory factors are fixed at the neutral levels. The exponential term in (30) rescales the benchmark $q_0^{(j)}(k)$ according to the individual characteristics of mortgages and prevailing macroeconomic conditions.

Following the setup in Example 2, the default loss $b^{(1)}(k, \boldsymbol{x}_i, \boldsymbol{Y})$ and prepayment loss $b^{(2)}(k, \boldsymbol{x}_i, \boldsymbol{Y})$ of the RMBS are specified as

$$b^{(1)}(k, \boldsymbol{x}_i, \boldsymbol{y}) = M_i(k) \frac{\alpha(\boldsymbol{x}_i, \boldsymbol{y})}{\alpha(\boldsymbol{x}_i, \boldsymbol{y}) + \beta(\boldsymbol{x}_i, \boldsymbol{y})} \text{ and } b^{(2)}(k, \boldsymbol{x}_i, \boldsymbol{y}) = R_i \sum_{h=0}^{K-k} e^{-hr} - M_i(k),$$

where R_i and $M_i(k)$ are defined in the Equation (7), and the functions $\alpha(\boldsymbol{x}_i, \boldsymbol{y}), \beta(\boldsymbol{x}_i, \boldsymbol{y})$ are the parameters of Beta distribution. Consequently, the portfolio loss $\boldsymbol{L}_N$ of the RMBS is calculated by (2) and (3).

4.2. Freddie Mac Single-Family Loan Dataset

Our empirical analysis utilizes the complete Freddie Mac single-family loan dataset[5], which includes all available 15-year fixed-rate mortgages issued in 2000. We partition the data into a training set of 500,000 loans and a test set of 57,202 loans (the remaining parts outside the training set). Both sets are mutually exclusive and randomly selected, preserving the homogeneity. Crucially, our robustness checks confirm that the results also hold even when using only 500,000 randomly sampled loans (vs. the entire 557,202), which indicates that the sensitivity to dataset size or selection is small. The dataset comprises the loans from all 50 U.S. states, with the distribution among states illustrated in Figure 4. The largest share is from California (10.8%), followed by Florida (6.6%), Texas (6.4%), and Illinois (5.4%), and no other state exceeds 5%. The "Others" category aggregates smaller states, accounting for approximately half (54.2%) of the sample, highlighting that the dataset is broadly representative across the geographic regions without significant concentration. In sum, the dataset for estimation consists of $N = 500{,}000$ loans with $K = 180$ monthly repayments.

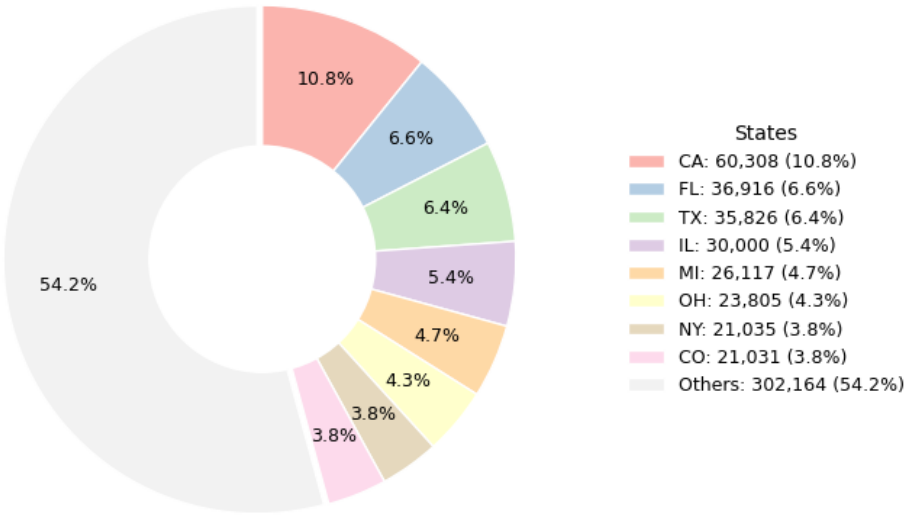


Figure 4. Distribution of mortgages by state.

The loan status can be deduced from the variables "Zero balance code" and "Defect settlement date" in the dataset. With slight modifications from the definitions outlined in Bhattacharya et al. (2019) due to data updates, the default and prepayment are now defined as follows:

- A loan is considered default if, for any month, the "Zero balance code" falls within the set $\{02, 03, 09, 15, 16, 96\}$.
- A loan is considered prepaid if, for any month, the "Zero balance code" is equal to 01 and the "Defect settlement date" is "NAN".
- A loan remains active if it does not meet the criteria for prepayment or default.

As described in Section 4.1, the systematic factors in our study are the U.S. Treasury long-term yield[6] $\delta(k)$ available at a monthly frequency, and the quarterly state-level HPI data[7], from which we compute the house price changes $HPC(S_m, k)$. The quarterly HPI data are linearly interpolated to a monthly frequency. The individual factors (i.e., loan rate, CS, LTV, DTI, FTB, and NB) are directly extracted from the dataset, which are fixed at the mortgage issue date.

To derive the loss value of the RMBS, the LGD of defaulted loans are obtained via the variables "Actual Loss Calculation" and "Current Actual UPB" from the dataset,

$$\text{LGD}_i = \frac{\text{Actual Loss Calculation}_i}{\text{Current Actual UPB}_i} \in [0, 1]. \tag{31}$$

Those observations outside the interval $[0, 1]$ are discarded.

4.3. Estimation Results

4.3.1. The Coefficients of Factors

Based on the training data of 500,000 mortgages issued at year 2000 (to year 2015) in the Freddie Mac dataset, the coefficient vectors in the joint default–prepayment hazard model (30) are estimated by the maximum partial likelihood method, where the details of the methodology are provided in Appendix B.1. The estimation results are shown in Table 2 and provide some empirical evidence to explain the effect of different factors on the default and prepayment.

Table 2. The coefficient vectors of the CPH model.

Coefficient	Default	Prepayment
interest spread	6.3437	1.1938
HPC	−0.9418	1.1428
CS	−3.3724	0.7926
LTV	3.9230	0.0332
DTI	0.0934	0.4961
FTB	−0.5673	0.0181
NB	−0.6337	0.2550

First, the borrower behavior is strongly influenced by the market-level systematic factors. Specifically, higher interest rate spreads encourage the borrowers to refinance or sell, substantially increasing the voluntary prepayment; however, the borrowers who are unable to refinance experience greater financial pressure, which leads to higher default risk. Further, the rising house prices enhance the borrower equity, improving the incentives to maintain ownership, thus lowering the default risk (reflected by the negative coefficient of HPC). Conversely, the rising house prices provide easier refinancing opportunities and profitable property sales, increasing the prepayment probabilities (reflected by the positive coefficient of HPC). These findings align with the existing literature (Campbell and Cocco 2015; Quercia and Stegman 1992).

Second, the borrowers' credit quality and financial structure play a critical role in determining the default risk. Higher credit scores, first-time buyers, and the presence of co-borrowers are associated with significantly lower default probabilities, as indicated by their negative coefficients, reflecting the stronger repayment capabilities and reduced financial distress. In contrast, the single-borrower households demonstrate greater vulnerability to financial distress, leading to elevated default rates. Additionally, higher LTV ratios significantly increase the default risk by reducing the equity buffers and intensifying the financial pressure for borrowers.

Third, the prepayment behavior is accelerated primarily by the stronger refinancing capacity of borrowers. Specifically, the borrowers with higher CS, DTI, and NB prepay more rapidly, reflecting the greater incentives and ability to refinance or sell. This is consistent with the related analysis of prepayment in Munk (2011). In contrast, the prepayment rate is insensitive to the LTV ratio and FTB status.

In summary, the estimation results show that the default and prepayment behaviors of borrowers are affected by distinct economic factors. This can improve the understanding of RMBS loan performance and leads to more accurate portfolio loss forecasts.

4.3.2. Baseline Hazard Functions

Based on the estimated coefficients in Equation (30), we estimate the baseline hazard functions $q_0^{(j)}(k)$ by matching the expected event numbers with the actual data. The details of the methodology are provided in Appendix B.2. The estimation results for default and prepayment are plotted in Figure 5. The prepayment hazard rate peaks in the first 60 months and flattens from then. The default rate is much lower than the prepayment rate and presents a smoother curve.

These patterns have important implications for modeling the borrower behavior and predicting the RMBS cash flows. The significantly higher prepayment risk in the early life of loans underscores the need to capture the time-varying characteristics of the interest rate, which has an influence on the borrowers' refinancing incentives. On the other hand, the relatively smooth and lower default hazard suggests that the borrowers' default behaviors

are relatively rare and often occur uniformly over time rather than present the clustering characteristics as the prepayment.

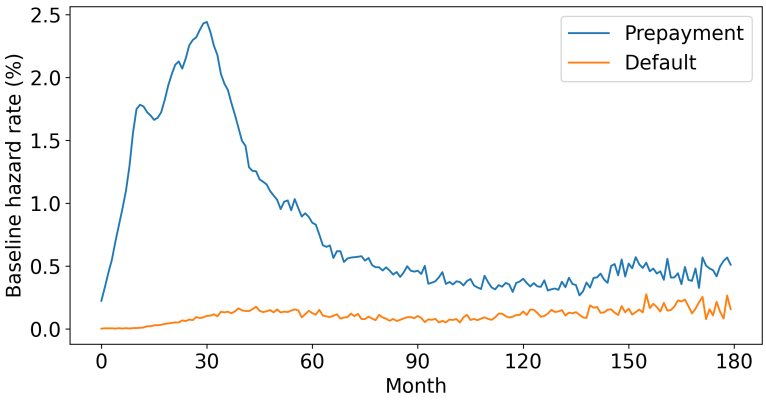


Figure 5. The baseline hazard functions.

4.3.3. The Distribution of LGD

To incorporate the borrower credit quality heterogeneity into the loss severity, we partition the loans into three categories based on the CS percentiles. Specifically, the top 40% of loans ordered by CS are labeled as the high credit score borrowers, the middle 40% as the moderate, and the bottom 20% as the low. For each group, we estimate the LGD distribution by fitting the observed loss data derived by (31) to the Beta distributions, following Moody's RMBS methodology (Moody's Investors Service 2024). Table 3 reports the mean of the LGD and the corresponding Beta distribution parameters for each category.

Table 3. The LGD distibutions for different categories of credit scores.

Category	Description	Mean	Distribution of LGD
High Credit Score	Highest 40% CS	0.3089	Beta (0.4799, 1.0739)
Moderate Credit Score	Moderate 40% CS	0.3161	Beta (0.4833, 1.0458)
Low Credit Score	Lowest 20% CS	0.3284	Beta (0.5359, 1.0962)

As shown in Table 3, the borrowers with higher credit scores tend to have lower LGD. Incorporating this classification into the simulation allows for more accurate modeling of loss severity across heterogeneous credit risk profiles.

4.3.4. Out-of-Sample Validation

To check the robustness of the estimation procedure, we conduct an out-of-sample validation using a separate test set consisting of 57,202 loans issued in 2000. Figure 6 compares the model-implied cumulative defaults and prepayments against the actual observed values in this test set.

The close alignment between the estimated and realized outcomes indicates that the estimation procedure is effective. The estimation error is relatively larger in the first few months, primarily due to the limited number of early default/prepayment behaviors, which affects the statistical precision. Nevertheless, the errors decrease quickly over time, indicating that the model captures the underlying risk dynamics reasonably well. These results suggest that our framework is useful for accurately estimating the default and prepayment probabilities in modeling the RMBS, which are essential for further evaluating the portfolio losses.

With a realized systematic factor path $\boldsymbol{Y} = \boldsymbol{y}$ fixed (see Appendix B.3 for details of the structure of $\boldsymbol{y}$), we draw 10,000 independent samples of the portfolio loss $\boldsymbol{L}_N = (L_N^{(1)}, L_N^{(2)})$ from the estimated CPH model. Figure 7 plots the empirical distributions of $L_N^{(1)}$ and

$L_N^{(2)}$ together with the fitted Gaussian densities. The close visual agreement supports the conditional CLT in Theorem 1. The bivariate normal parameters estimated from the 10,000 draws are

$$\hat{\boldsymbol{\mu}}(\boldsymbol{y}) = (0.482,\ 10.714), \quad \hat{\boldsymbol{\Sigma}}(\boldsymbol{y}) = \begin{pmatrix} 3.922 \times 10^{-4} & -9.668 \times 10^{-5} \\ -9.668 \times 10^{-5} & 1.515 \times 10^{-4} \end{pmatrix},$$

and the correlation coefficient is -0.397. The negative correlation between default and prepayment losses reflects the competing-risk nature of these two events in the RMBS portfolios, where the early prepayment naturally reduces the number of loans exposed to default risk. This finding aligns with actual market behavior and highlights the importance of jointly modeling default and prepayment risks.

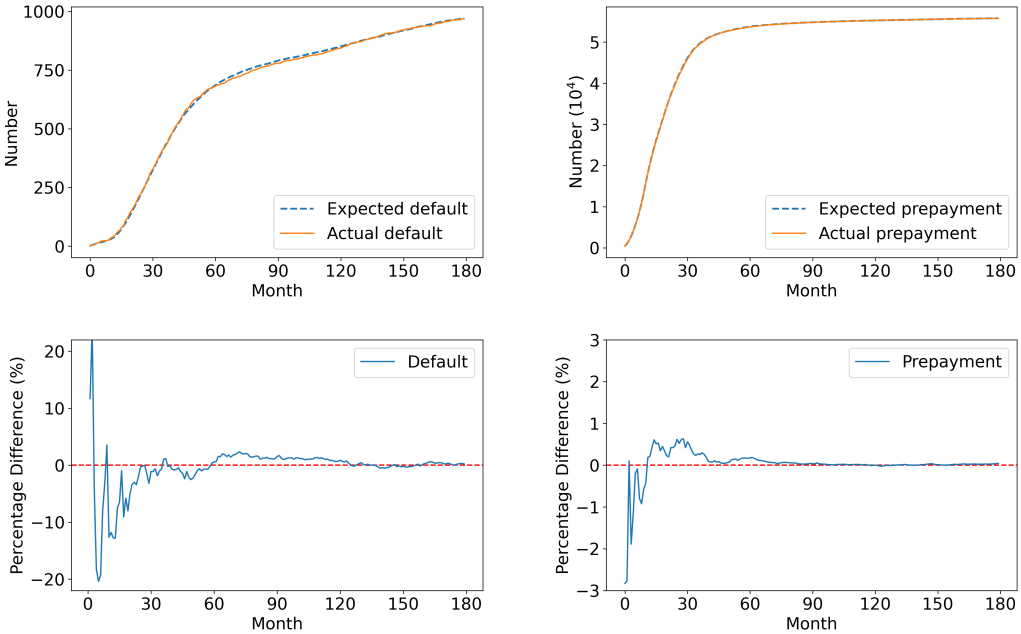


Figure 6. The estimation results of default (**left**) and prepayment (**right**) in the test set. Note: the red dotted lines in the lower panels represent zero difference between the expected and actual number of defaults, serving as a baseline for comparing the percentage difference curves.

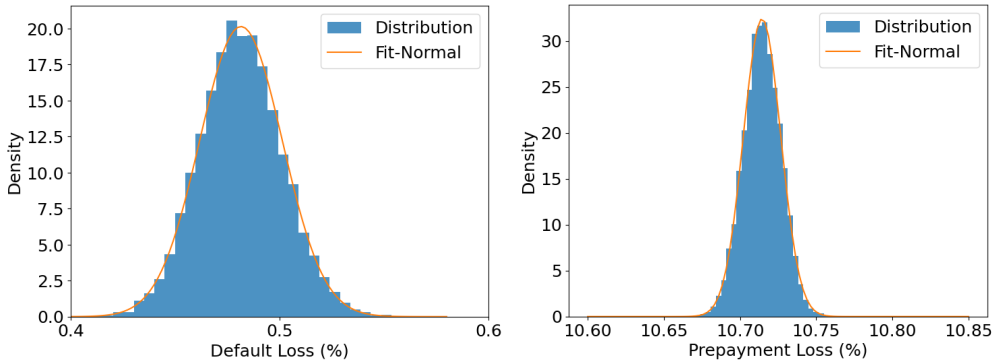


Figure 7. Empirical limit distributions of default loss (**left**) and prepayment loss (**right**).

4.4. Simulation of Portfolio Loss

Recall that in Proposition 2 (ii), the portfolio loss $\boldsymbol{L}_N$ converges to the conditional mean $\boldsymbol{\mu}(\boldsymbol{Y})$ as $N \to \infty$, but the distributions of random variables $\mu_1(\boldsymbol{Y})$ and $\mu_2(\boldsymbol{Y})$ in $\boldsymbol{\mu}(\boldsymbol{Y})$ do not admit explicit expressions. To further investigate the distribution of $\boldsymbol{\mu}(\boldsymbol{Y})$, in this part we generate the loss distribution of a representative RMBS pool using a bottom-up approach.

Given the estimated hazard functions for default and prepayment from Section 4.3, we simulate the losses of individual loans based on the simulated dynamics of systematic risk factors. Then, the portfolio loss for the RMBS is generated.

4.4.1. The Models of Systematic Factors

In our model, both the interest rate and house price changes have influence over the payment period. To simulate the evolution of interest rates, we adopt the Cox–Ingersoll–Ross (CIR) model (Cox et al. 1985), which is widely used due to its ability to capture the non-negativity and mean-reverting nature of interest rates. The discrete-time approximation of the CIR process using the Euler–Maruyama scheme is given by

$$\delta(k+1) = \delta(k) + \kappa(\theta - \delta(k))\Delta t + \sigma\sqrt{\delta(k)}\sqrt{\Delta t} \cdot Z_k,$$

where Δt is the time step (1/12 for monthly frequency), θ is the mean of the long-term interest rate, κ is the rate of average recovery, σ is the volatility, and $\{Z_k\}$ are i.i.d. $N(0,1)$ random variables.

The parameters of the CIR model are estimated using the historical U.S. Treasury long-term yield data obtained from the U.S. Department of the Treasury. The estimated results are reported in Table 4. One sample path of the simulated interest rate over a 180-month period is shown in Figure 8. Here, the actual rate curve represents the observed U.S. Treasury long-term yield data from the historical record over the same period. The downward trend in the simulated interest rate arises naturally due to the mean reversion towards the estimated long-run equilibrium rate. As shown in Table 4, the initial historical rate exceeds the equilibrium value $\hat{\theta} = 3.4545$. The declining rates widen the loan spreads, subsequently elevating the default and prepayment probabilities and thereby increasing the potential portfolio losses.

Table 4. The estimators of the CIR model.

$\hat{\kappa}$	$\hat{\theta}$	$\hat{\sigma}$
0.3210	3.4545	0.1973

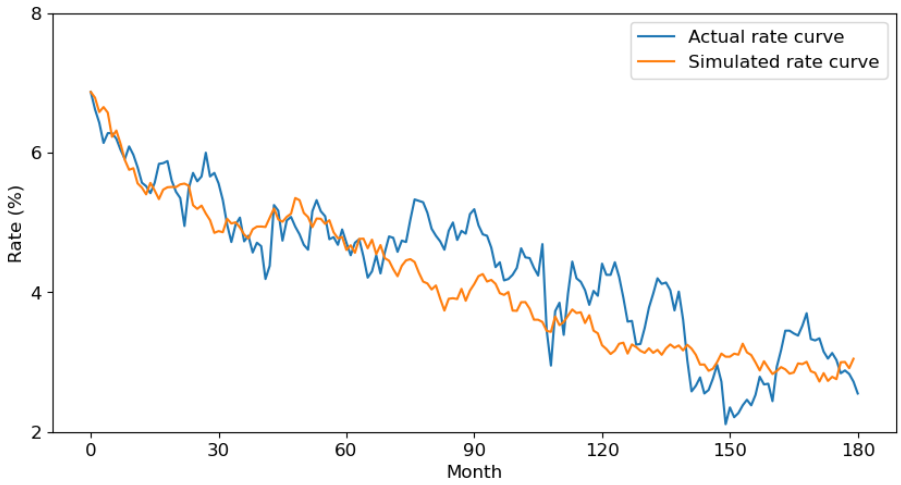


Figure 8. A simulation of the interest rate curve.

To simulate the dynamics of the house price across the U.S. states, we employ a second-order autoregressive model (AR(2)) on the log house price index series. Specifically, for each state S_m, the log house price index evolves according to

$$\log HPI(S_m, k) = \phi_0 + \phi_1 \log HPI(S_m, k-1) + \phi_2 \log HPI(S_m, k-2) + \sigma_H Z(S_m, k),$$

where $Z(S_m, k) \sim N(0,1)$, $1 \leq k \leq K$ are i.i.d. standard normal shocks in the kth month. Subsequently, the HPC is calculated by the Equation (29).

The estimated AR(2) parameters of all U.S. states are summarized in Table 5. The last row of Table 5 reports the AR(2) estimates for the national-level log house price series, which provides a benchmark for comparing the state-level dynamics. The AR(2) simulations consistently reveal a persistent and autoregressive growth pattern, despite the different initial house price levels across states.

Table 5. Summary statistics of AR(2) parameters for log house price time series across U.S. states.

Parameter	Mean	Std Deviation	Minimum	Maximum	U.S. National
ϕ_0	−0.012569	0.039243	−0.119973	0.053533	−0.008022
ϕ_1	1.596188	0.141466	1.255425	1.818609	1.708026
ϕ_2	−0.593126	0.145872	−0.823798	−0.257050	−0.706100
σ_H	0.012770	0.002779	0.008456	0.021200	0.011299

4.4.2. Simulation Results

Based on the empirical results in Section 4.3, we simulate the RMBS portfolio losses under a bottom-up framework. In each scenario, we simulate one sample of the portfolio default and prepayment loss $(L_N^{(1)}, L_N^{(2)})$ for the pool of 57,202 loans in the test set. By repeating this process over 100,000 times, we generate the distribution of $L_N^{(1)}$ and $L_N^{(2)}$, where the marginal distributions for $L_N^{(1)}$ and $L_N^{(2)}$ are presented in Figure 9.

From the right panel in Figure 9, both normal and log-normal distributions provide a good approximation for the distribution of portfolio prepayment loss. In contrast, as seen in the left panel of Figure 9, the log-normal distribution offers a substantially better fit of the portfolio default loss, compared to the normal distribution. This empirical observation provides quantitative evidence for the commonly assumed log-normal specification in the top-down models.

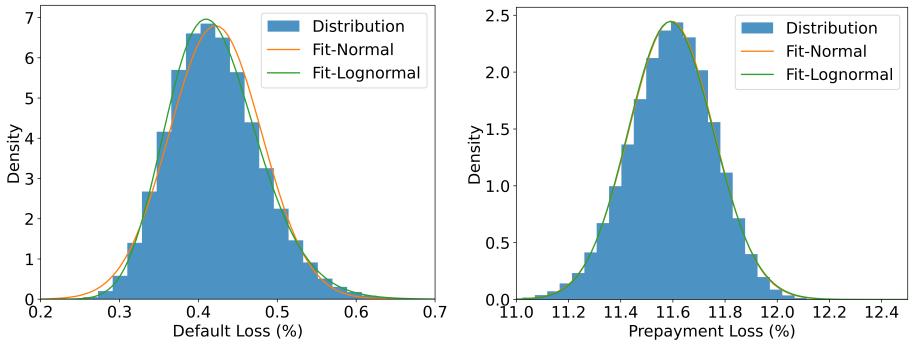


Figure 9. The marginal distributions of the simulated portfolio loss $\boldsymbol{L}_N$ of the RMBS.

To further empirically examine the distributional form of $L_N^{(1)}$ and $L_N^{(2)}$, we take the logarithm of the simulated portfolio loss and construct the quantile–quantile (QQ) plots against the normal distribution. As shown in Figure 10, the log of losses are close to the theoretical quantiles of a normal distribution, with only mild deviations in the tails. This

evidence suggests that $\mu(Y)$ is approximately log-normally distributed under the simulated macroeconomic scenarios.

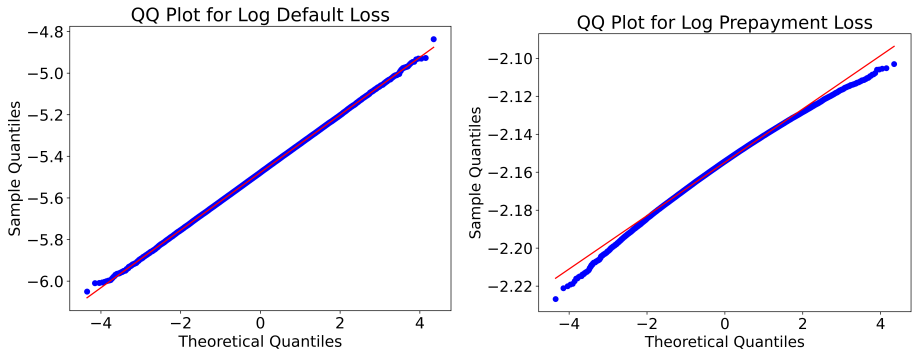


Figure 10. The QQ plots for the simulated portfolio loss of the RMBS. Note: The blue line represents the empirical quantiles of the simulated data, while the red straight line corresponds to the theoretical quantiles of a normal distribution.

Moreover, the value-at-risk (VaR) of simulated $L_N^{(1)}$ and $L_N^{(2)}$ and the corresponding fitted log-normal distributions are shown in Table 6. The small deviations between the fitted and simulated values reveal that the log-normal distribution provides a reasonable approximation of the portfolio loss distribution. This result not only supports the log-normal assumption widely adopted in the top-down credit risk models but also fills an important gap in our theoretical framework by offering a plausible empirical distribution for the conditional loss mean.

Table 6. The VaR of the simulated $L_N^{(1)}$ and $L_N^{(2)}$ of the RMBS.

	$L_N^{(1)}$ (%)			$L_N^{(2)}$ (%)		
	VaR$_{0.95}$	**VaR$_{0.975}$**	**VaR$_{0.99}$**	**VaR$_{0.95}$**	**VaR$_{0.975}$**	**VaR$_{0.99}$**
Simulated	0.5245	0.5473	0.5766	11.85	11.90	11.95
Fit-log-normal	0.5246	0.5480	0.5765	11.86	11.92	11.98

5. Conclusions

This paper discussed the joint distribution of default and prepayment losses for a large portfolio of loans, extending the previous studies that focused solely on default losses. Under the conditionally independent framework, the two-dimensional limit distributions of the portfolio default and prepayment losses were obtained, including the strong law of large numbers and the central limit theorem. Numerical studies for the portfolio losses were performed for some simplified models. Empirical analysis on the RMBS was carried out for analyzing the limit joint distribution of the default and prepayment losses. The portfolio default and prepayment losses were simulated based on the empirically estimated model parameters. The copula of default and prepayment loss distributions was exhibited. Both the simulated distributions of portfolio default and prepayment losses fitted the log-normal distributions well, which is consistent with the log-normal distribution assumptions for portfolio losses in Moody's Investors Service (2024).

Author Contributions: Conceptualization, methodology, formal analysis and investigation, C.X., X.Z., L.B., Q.D. and J.Y.; software, validation, data curation and visualization, C.X., L.B. and Q.D.; writing—original draft preparation and writing—review and editing, C.X., X.Z., Q.D. and J.Y.; supervision and funding acquisition, X.Z. and J.Y.; resources and project administration, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key R&D Program of China (Grant No. 2018YFA0703900) and the National Natural Science Foundation of China (Grant Nos. 12471445 and 12071016). The research of Zang was also supported by the National Natural Science Foundation of China (Grant No. 12301598).

Data Availability Statement: The data of the empirical study in this paper is obtained from the Freddie Mac single-family loan dataset released by the Federal Home Loan Mortgage Corporation.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Some Proofs

Appendix A.1. Proof of Lemma 1

Proof. (i) From Assumption 1, $\tau_i^{(1)} \neq \tau_i^{(2)}$ a.s. Then, for $1 \leq k \leq K-1$, we have

$$\begin{aligned}&\mathbb{P}(\tau_i > k|\tau_i > k-1, \boldsymbol{Y}=\boldsymbol{y}) = 1 - \mathbb{P}(\tau_i = k|\tau_i > k-1, \boldsymbol{Y}=\boldsymbol{y})\\ =&1 - \mathbb{P}(\tau_i^{(1)} = k|\tau_i > k-1, \boldsymbol{Y}=\boldsymbol{y}) - \mathbb{P}(\tau_i^{(2)} = k|\tau_i > k-1, \boldsymbol{Y}=\boldsymbol{y})\\ =&1 - q^{(1)}(k, \boldsymbol{x}_i, \boldsymbol{y}) - q_i^{(2)}(k, \boldsymbol{x}_i, \boldsymbol{y}),\end{aligned}$$

where the probabilities $q^{(j)}(k, \boldsymbol{x}, \boldsymbol{y}), j = 1, 2$ are defined in (4), and the last equation is due to Assumption 2. Consequently, the conditional survival function of the ith loan is given by

$$\begin{aligned}\mathbb{P}(\tau_i > k|\boldsymbol{Y}=\boldsymbol{y}) =& \prod_{h=1}^{k} \mathbb{P}(\tau_i > h|\tau_i > h-1, \boldsymbol{Y}=\boldsymbol{y}) = \prod_{h=1}^{k} \Big(1 - q^{(1)}(h, \boldsymbol{x}_i, \boldsymbol{y}) - q^{(2)}(h, \boldsymbol{x}_i, \boldsymbol{y})\Big)\\ =&S(k, \boldsymbol{x}_i, \boldsymbol{y}),\ 0 \leq k \leq K-1\end{aligned}$$

with $S(k, \boldsymbol{x}, \boldsymbol{y})$ in (9). This yields Equation (10).

(ii) Note that $D_i = 0$ implies $\tau_i = K$ by Definition (1), so

$$\mathbb{P}(\tau_i = k, D_i = 0|\boldsymbol{Y}=\boldsymbol{y}) = 0,\ 1 \leq k \leq K-1.$$

And similar to the proof in (i),

$$\begin{aligned}\mathbb{P}(\tau_i = K, D_i = 0|\boldsymbol{Y}=\boldsymbol{y}) =&\mathbb{P}(\tau_i > K-1|\boldsymbol{Y}=\boldsymbol{y})\mathbb{P}(\tau_i = K, D_i = 0|\tau_i > K-1, \boldsymbol{Y}=\boldsymbol{y})\\ =&\prod_{h=1}^{K} \Big(1 - q^{(1)}(h, \boldsymbol{x}_i, \boldsymbol{y}) - q^{(2)}(h, \boldsymbol{x}_i, \boldsymbol{y})\Big) = S(K, \boldsymbol{x}_i, \boldsymbol{y}).\end{aligned}$$

For $j = 1, 2$ and $1 \leq k \leq K$, we have

$$\begin{aligned}\mathbb{P}(\tau_i = k, D_i = j|\boldsymbol{Y}=\boldsymbol{y}) =&\mathbb{P}(\tau_i > k-1|\boldsymbol{Y}=\boldsymbol{y})\mathbb{P}(\tau_i = k, D_i = j|\tau_i > k-1, \boldsymbol{Y}=\boldsymbol{y})\\ =&\mathbb{P}(\tau_i > k-1|\boldsymbol{Y}=\boldsymbol{y})\mathbb{P}(\tau_i^{(j)} = k|\tau_i > k-1, \boldsymbol{Y}=\boldsymbol{y})\\ =&S(k-1, \boldsymbol{x}_i, \boldsymbol{y})q^{(j)}(k, \boldsymbol{x}_i, \boldsymbol{y})\end{aligned}$$

based on (4) and (10). This finishes the proof. □

Appendix A.2. Proof of Proposition 1

Proof. (i) Given $\boldsymbol{Y}=\boldsymbol{y}$, for $j = 1, 2$, the conditional expectation of $l_i^{(j)}$ in (2) is given by

$$\begin{aligned}\mathbb{E}[l_i^{(j)}|\boldsymbol{Y}=\boldsymbol{y}] =&\mathbb{E}\Big[1_{D_i=j}b_i^{(j)}e^{-r\tau_i}|\boldsymbol{Y}=\boldsymbol{y}\Big]\\ =&\sum_{k=1}^{K}\mathbb{P}(\tau_i=k,D_i=j|\boldsymbol{Y}=\boldsymbol{y})\mathbb{E}[b_i^{(j)}|\tau_i=k,D_i=j,\boldsymbol{Y}=\boldsymbol{y}]e^{-kr}\\ =&\sum_{k=1}^{K}S(k-1,\boldsymbol{x}_i,\boldsymbol{y})q^{(j)}(k,\boldsymbol{x}_i,\boldsymbol{y})b^{(j)}(k,\boldsymbol{x}_i,\boldsymbol{y})e^{-kr}=m_j(\boldsymbol{x}_i,\boldsymbol{y})\end{aligned}$$

with $m_j(\boldsymbol{x},\boldsymbol{y})$ defined in (12). Similarly, we get

$$\begin{aligned}\mathbb{E}[(l_i^{(j)})^2|\boldsymbol{Y}=\boldsymbol{y}] =&\mathbb{E}\Big[1_{D_i=j}(b_i^{(j)})^2e^{-2r\tau_i}|\boldsymbol{Y}=\boldsymbol{y}\Big]\\ =&\sum_{k=1}^{K}\mathbb{P}(\tau_i=k,D_i=j|\boldsymbol{Y}=\boldsymbol{y})\mathbb{E}[(b_i^{(j)})^2|\tau_i=k,D_i=j,\boldsymbol{Y}=\boldsymbol{y}]e^{-2kr}\\ =&\sum_{k=1}^{K}S(k-1,\boldsymbol{x}_i,\boldsymbol{y})q^{(j)}(k,\boldsymbol{x}_i,\boldsymbol{y})B^{(j)}(k,\boldsymbol{x}_i,\boldsymbol{y})e^{-2kr},\end{aligned}$$

which yields that

$$\begin{aligned}\mathrm{Var}[l_i^{(j)}|\boldsymbol{Y}=\boldsymbol{y}] =&\mathbb{E}[(l_i^{(j)})^2|\boldsymbol{Y}=\boldsymbol{y}]-\mathbb{E}[l_i^{(j)}|\boldsymbol{Y}=\boldsymbol{y}]^2\\ =&\sum_{k=1}^{K}S(k-1,\boldsymbol{x}_i,\boldsymbol{y})q^{(j)}(k,\boldsymbol{x}_i,\boldsymbol{y})B^{(j)}(k,\boldsymbol{x}_i,\boldsymbol{y})e^{-2kr}-m_j^2(\boldsymbol{x}_i,\boldsymbol{y})=v_{j,j}(\boldsymbol{x}_i,\boldsymbol{y})\end{aligned}$$

with $v_{j,j}(\boldsymbol{x},\boldsymbol{y})$ defined in (13). Moreover,

$$\begin{aligned}\mathrm{Cov}[l_i^{(1)},l_i^{(2)}|\boldsymbol{Y}=\boldsymbol{y}] =&\mathbb{E}[l_i^{(1)}l_i^{(2)}|\boldsymbol{Y}=\boldsymbol{y}]-\mathbb{E}[l_i^{(1)}|\boldsymbol{Y}=\boldsymbol{y}]\mathbb{E}[l_i^{(2)}|\boldsymbol{Y}=\boldsymbol{y}]\\ =&-m_1(\boldsymbol{x}_i,\boldsymbol{y})m_2(\boldsymbol{x}_i,\boldsymbol{y}),\end{aligned}$$

where the last equality is due to $l_i^{(1)}l_i^{(2)}=0$, and thus, $\mathbb{E}[l_i^{(1)}l_i^{(2)}|\boldsymbol{Y}=\boldsymbol{y}]=0$. So, conclusion (14) follows.

(ii) For $\boldsymbol{L}_N=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{l}_i$ in (3), we have

$$\mathbb{E}[\boldsymbol{L}_N|\boldsymbol{Y}=\boldsymbol{y}]=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\boldsymbol{l}_i|\boldsymbol{Y}=\boldsymbol{y}]=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{m}(\boldsymbol{x}_i,\boldsymbol{y}).$$

Moreover, by Assumption 3, $\boldsymbol{l}_1,\boldsymbol{l}_2,\ldots,\boldsymbol{l}_N$ are conditionally independent given $\boldsymbol{Y}=\boldsymbol{y}$. This yields

$$\mathrm{Var}[\boldsymbol{L}_N|\boldsymbol{Y}=\boldsymbol{y}]=\frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}[\boldsymbol{l}_i|\boldsymbol{Y}=\boldsymbol{y}]=\frac{1}{N^2}\sum_{i=1}^{N}\boldsymbol{V}(\boldsymbol{x}_i,\boldsymbol{y}).$$

□

Appendix A.3. Proof of Proposition 2

Proof. (i) From Assumption 3, the sequence $\{\boldsymbol{l}_i\}_{i\geq1}$ is conditionally independent given $\boldsymbol{Y}=\boldsymbol{y}$. For $j=1,2$, conditional on $\boldsymbol{Y}=\boldsymbol{y}$, the sequence

$$\frac{l_i^{(j)}-\mathbb{E}[l_i^{(j)}|\boldsymbol{Y}=\boldsymbol{y}]}{i},\ i\geq1$$

satisfies the conditions of Theorem 2.5.3 in Durrett (2019), such that they are conditionally independent with a conditional expectation of 0, and

$$\sum_{i=1}^{\infty} \mathrm{Var}\left[\frac{l_i^{(j)} - \mathbb{E}[l_i^{(j)}|\boldsymbol{Y}=\boldsymbol{y}]}{i}\middle|\boldsymbol{Y}=\boldsymbol{y}\right] = \sum_{i=1}^{\infty}\frac{1}{i^2}\mathrm{Var}[l_i^{(j)}|\boldsymbol{Y}=\boldsymbol{y}] \le \sum_{i=1}^{\infty}\frac{4}{i^2} < \infty,$$

where the first inequality follows from $l_i^{(j)} \in [0,1]$ and further $|l_i^{(j)} - \mathbb{E}[l_i^{(j)}|\boldsymbol{Y}=\boldsymbol{y}]| \le 2$. Applying this theorem, we conclude that $\sum_{i=1}^{N}(l_i^{(j)} - \mathbb{E}[l_i^{(j)}|\boldsymbol{Y}=\boldsymbol{y}])/i$ converges almost surely as $N \to \infty$ conditional on $\boldsymbol{Y}=\boldsymbol{y}$. By Kronecker's lemma (Theorem 2.5.5 in Durrett 2019), this implies

$$\mathbb{P}\left(\lim_{N\to\infty}(L_N^{(j)} - \mathbb{E}[L_N^{(j)}|\boldsymbol{Y}=\boldsymbol{y}]) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}(l_i^{(j)} - \mathbb{E}[l_i^{(j)}|\boldsymbol{Y}=\boldsymbol{y}]) = 0|\boldsymbol{Y}=\boldsymbol{y}\right) = 1.$$

That is, $\mathbb{P}(\lim_{N\to\infty}(\boldsymbol{L}_N - \mathbb{E}[\boldsymbol{L}_N|\boldsymbol{Y}=\boldsymbol{y}]) = 0|\boldsymbol{Y}=\boldsymbol{y}) = 1$ in vector notation. Then, we have

$$\mathbb{P}(\lim_{N\to\infty}(\boldsymbol{L}_N - \mathbb{E}[\boldsymbol{L}_N|\boldsymbol{Y}]) = 0) = \int \mathbb{P}(\lim_{N\to\infty}(\boldsymbol{L}_N - \mathbb{E}[\boldsymbol{L}_N|\boldsymbol{Y}=\boldsymbol{y}]) = 0|\boldsymbol{Y}=\boldsymbol{y})\mathrm{d}F_{\boldsymbol{Y}}(\boldsymbol{y}) = 1,$$

which yields Equation (18).

(ii) Since the functions $q^{(j)}(k,\boldsymbol{x},\boldsymbol{y})$ and $b^{(j)}(k,\boldsymbol{x},\boldsymbol{y})$ in (4) are continuous, the moment functions $m_1(\boldsymbol{x},\boldsymbol{y}), m_2(\boldsymbol{x},\boldsymbol{y})$ defined in (11) are consequently continuous and bounded in [0,1]. Moreover, Assumption 4 establishes the weak convergence $F_{\boldsymbol{X}}^{(N)}(\boldsymbol{x}) \xrightarrow{w} F_{\boldsymbol{X}}(\boldsymbol{x})$ as $N \to \infty$. By the weak convergence theorem (Theorem 3.2.3 in Durrett 2019), we have

$$\begin{aligned}\lim_{N\to\infty}\mathbb{E}[\boldsymbol{L}_N|\boldsymbol{Y}=\boldsymbol{y}] &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{m}(\boldsymbol{x}_i,\boldsymbol{y}) = \lim_{N\to\infty}\int \boldsymbol{m}(\boldsymbol{x},\boldsymbol{y})\mathrm{d}F_{\boldsymbol{X}}^{(N)}(\boldsymbol{x}) \\ &= \int \boldsymbol{m}(\boldsymbol{x},\boldsymbol{y})\mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x}) = \boldsymbol{\mu}(\boldsymbol{y}),\end{aligned} \tag{A1}$$

with $\boldsymbol{\mu}(\boldsymbol{y})$ defined in (17). Combining (18) and (A1), we obtain Equation (19). □

Appendix A.4. Proof of Corollary 1

Proof. From Proposition 2 (ii), $\boldsymbol{L}_N$ converges almost surely to $\boldsymbol{\mu}(\boldsymbol{Y})$ as $N \to \infty$. Consequently, for $j = 1,2$, its component $L_N^{(j)} \to \mu_j(\boldsymbol{Y})$ a.s. This implies convergence in distribution, i.e.,

$$\lim_{N\to\infty}\mathbb{P}(L_N^{(j)} \le t) = \mathbb{P}(\mu_j(\boldsymbol{Y}) \le t) \tag{A2}$$

for any continuity point t of $\mu_j(\boldsymbol{Y})$.

Fix $j \in \{1,2\}$, $\alpha \in (0,1)$, and $\varepsilon > 0$. Since the distribution function of $\mu_j(\boldsymbol{Y})$ is non-decreasing and right-continuous, we can select its continuity points t_1, t_2 satisfying

$$\mathrm{VaR}_\alpha(\mu_j(\boldsymbol{Y})) - \varepsilon < t_1 < \mathrm{VaR}_\alpha(\mu_j(\boldsymbol{Y})) < t_2 < \mathrm{VaR}_\alpha(\mu_j(\boldsymbol{Y})) + \varepsilon. \tag{A3}$$

By the definition of VaR, we have

$$\mathbb{P}(\mu_j(\boldsymbol{Y}) \le t_1) \le \alpha \le \mathbb{P}(\mu_j(\boldsymbol{Y}) \le t_2). \tag{A4}$$

Combining (A2), (A3), and (A4), it follows that

$$\lim_{N\to\infty}\mathbb{P}(L_N^{(j)} \le \mathrm{VaR}_\alpha(\mu_j(\boldsymbol{Y})) - \varepsilon) \le \lim_{N\to\infty}\mathbb{P}(L_N^{(j)} \le t_1) = \mathbb{P}(\mu_j(\boldsymbol{Y}) \le t_1) \le \alpha$$

and

$$\lim_{N\to\infty}\mathbb{P}(L_N^{(j)} \le \mathrm{VaR}_\alpha(\mu_j(\boldsymbol{Y})) + \varepsilon) \ge \lim_{N\to\infty}\mathbb{P}(L_N^{(j)} \le t_2) = \mathbb{P}(\mu_j(\boldsymbol{Y}) \le t_2) \ge \alpha.$$

This finishes the proof. □

Appendix A.5. Proof of Theorem 1

For the proof of this theorem, we first propose two technical lemmas.

Lemma A1. *For $1 \leq i \leq N$, let A_i be a 2×2 symmetric matrix of the form*

$$A_i = \begin{pmatrix} a_i & b_i \\ b_i & c_i \end{pmatrix}, \text{ where } 0 \leq a_i, c_i \leq 1, -1 \leq b_i \leq 1 \text{ and } \det(A_i) > \varepsilon \tag{A5}$$

for some $\varepsilon > 0$. Define the sum matrix $A = \sum_{i=1}^{N} A_i$. Then, each entry of the inverse matrix A^{-1} is in the interval $[-(N\varepsilon)^{-1}, (N\varepsilon)^{-1}]$.

Proof. For each A_i, the determinant condition $\det(A_i) = a_i c_i - b_i^2 > \varepsilon$ implies $c_i > (b_i^2 + \varepsilon)/a_i$. Then, it follows that

$$\sum_{i=1}^{N} a_i \sum_{i=1}^{N} c_i > \sum_{i=1}^{N} a_i \sum_{i=1}^{N} \frac{b_i^2 + \varepsilon}{a_i} = \sum_{i=1}^{N} a_i \sum_{i=1}^{N} \frac{b_i^2}{a_i} + \varepsilon \sum_{i=1}^{N} a_i \sum_{i=1}^{N} \frac{1}{a_i} \geq (\sum_{i=1}^{N} b_i)^2 + N^2\varepsilon,$$

where the last inequality is due to Cauchy–Schwarz's inequality (see Theorem 4.1 in Rudin (1987)). Thus,

$$\det(A) = \sum_{i=1}^{N} a_i \sum_{i=1}^{N} c_i - (\sum_{i=1}^{N} b_i)^2 > N^2\varepsilon$$

and $0 < \det^{-1}(A) < (N^2\varepsilon)^{-1}$. Note that

$$A = \begin{pmatrix} \sum_{i=1}^{N} a_i & \sum_{i=1}^{N} b_i \\ \sum_{i=1}^{N} b_i & \sum_{i=1}^{N} c_i \end{pmatrix} \text{ and } A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} \sum_{i=1}^{N} c_i & -\sum_{i=1}^{N} b_i \\ -\sum_{i=1}^{N} b_i & \sum_{i=1}^{N} a_i \end{pmatrix}.$$

As all entries of matrix A, i.e., $\sum_{i=1}^{N} a_i, \sum_{i=1}^{N} b_i$, and $\sum_{i=1}^{N} c_i$, are bounded in $[-N, N]$, each entry of matrix A^{-1} is bounded in $[-(N\varepsilon)^{-1}, (N\varepsilon)^{-1}]$, which finishes the proof. □

Lemma A2. *Suppose that Assumptions 1–3 and 5 hold. Given vector $\boldsymbol{y}$, define random variables*

$$\boldsymbol{Z}_{N,i} := \mathrm{Var}\left[\sum_{i=1}^{N} \boldsymbol{l}_i \middle| \boldsymbol{Y} = \boldsymbol{y}\right]^{-\frac{1}{2}} (\boldsymbol{l}_i - \mathbb{E}[\boldsymbol{l}_i | \boldsymbol{Y} = \boldsymbol{y}]), \quad 1 \leq i \leq N. \tag{A6}$$

Then, $\|\boldsymbol{Z}_{N,i}\|^2 \leq 4(N\varepsilon)^{-1}$, with ε in Assumption 5.

Proof. Note that each entry of $\boldsymbol{l}_i$ is in $[0, 1]$ by definition (2). Further, from Assumption 5, the matrices $\mathrm{Var}[\boldsymbol{l}_i | \boldsymbol{Y} = \boldsymbol{y}]$, $1 \leq i \leq N$ satisfy condition (A5). By Lemma A1, each entry of

$$\mathrm{Var}\left[\sum_{i=1}^{N} \boldsymbol{l}_i \middle| \boldsymbol{Y} = \boldsymbol{y}\right]^{-1} = \left[\sum_{i=1}^{N} \mathrm{Var}[\boldsymbol{l}_i | \boldsymbol{Y} = \boldsymbol{y}]\right]^{-1}$$

is bounded in $[-(N\varepsilon)^{-1}, (N\varepsilon)^{-1}]$, where the equality is due to the conditional independence in Assumption 3. Moreover, since each entry of $\boldsymbol{l}_i - \mathbb{E}[\boldsymbol{l}_i | \boldsymbol{Y} = \boldsymbol{y}]$ is bounded in $[-1, 1]$, we have

$$\|\boldsymbol{Z}_{N,i}\|^2 = \boldsymbol{Z}'_{N,i}\boldsymbol{Z}_{N,i} = (\boldsymbol{l}_i - \mathbb{E}[\boldsymbol{l}_i | \boldsymbol{Y} = \boldsymbol{y}])' \mathrm{Var}\left[\sum_{i=1}^{N} \boldsymbol{l}_i \middle| \boldsymbol{Y} = \boldsymbol{y}\right]^{-1} (\boldsymbol{l}_i - \mathbb{E}[\boldsymbol{l}_i | \boldsymbol{Y} = \boldsymbol{y}]) \leq \frac{4}{N\varepsilon}$$

for $1 \leq i \leq N$. This finishes the proof. □

Now, we prove Theorem 1.

(i) Given $\boldsymbol{Y} = \boldsymbol{y}$, the variables $\boldsymbol{Z}_{N,i}, 1 \leq i \leq N$ defined in (A6) are conditionally independent by Assumption 3 and satisfy $\mathbb{E}[\boldsymbol{Z}_{N,i}|\boldsymbol{Y} = \boldsymbol{y}] = 0, 1 \leq i \leq N$. Moreover,

$$\begin{aligned}\sum_{i=1}^{N} \boldsymbol{Z}_{N,i} =& \mathrm{Var}\left[\sum_{i=1}^{N} \boldsymbol{l}_i \middle| \boldsymbol{Y} = \boldsymbol{y}\right]^{-\frac{1}{2}} \sum_{i=1}^{N} (\boldsymbol{l}_i - \mathbb{E}[\boldsymbol{l}_i|\boldsymbol{Y} = \boldsymbol{y}]) \quad \text{(A7)}\\ =& \mathrm{Var}[\boldsymbol{L}_N|\boldsymbol{Y} = \boldsymbol{y}]^{-\frac{1}{2}} (\boldsymbol{L}_N - \mathbb{E}[\boldsymbol{L}_N|\boldsymbol{Y} = \boldsymbol{y}])\\ =& \boldsymbol{\Sigma}_N(\boldsymbol{y})^{-\frac{1}{2}} (\boldsymbol{L}_N - \boldsymbol{\mu}_N(\boldsymbol{y})), \quad \text{(A8)}\end{aligned}$$

where the second equality is due to (3) and the last equality follows by combining (15) and (21).

Note that

$$\begin{aligned}\sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{Z}_{N,i}\boldsymbol{Z}'_{N,i}|\boldsymbol{Y} = \boldsymbol{y}\right] =& \sum_{i=1}^{N} \mathrm{Var}[\boldsymbol{Z}_{N,i}|\boldsymbol{Y} = \boldsymbol{y}] = \mathrm{Var}\left[\sum_{i=1}^{N} \boldsymbol{Z}_{N,i} \middle| \boldsymbol{Y} = \boldsymbol{y}\right]\\ =& \boldsymbol{\Sigma}_N(\boldsymbol{y})^{-\frac{1}{2}} \mathrm{Var}[\boldsymbol{L}_N|\boldsymbol{Y} = \boldsymbol{y}] \boldsymbol{\Sigma}_N(\boldsymbol{y})^{-\frac{1}{2}} = \mathbb{I}_2, \quad \text{(A9)}\end{aligned}$$

where the first equality is due to $\mathbb{E}[\boldsymbol{Z}_{N,i}|\boldsymbol{Y} = \boldsymbol{y}] = 0$ and the third equality is due to (A8). And for any $\delta > 0$,

$$\begin{aligned}&\lim_{N\to\infty} \sum_{i=1}^{N} \mathbb{E}[\|\boldsymbol{Z}_{N,i}\|^2 1_{\{\|\boldsymbol{Z}_{N,i}\| > \delta\}} |\boldsymbol{Y} = \boldsymbol{y}] \quad \text{(A10)}\\ \leq& \lim_{N\to\infty} \sum_{i=1}^{N} \mathbb{E}[\|\boldsymbol{Z}_{N,i}\|^2 1_{\{4(N\varepsilon)^{-1} > \delta^2\}} |\boldsymbol{Y} = \boldsymbol{y}] = 0, \quad \text{(A11)}\end{aligned}$$

where the inequality follows from Lemma A2. Equations (A9) and (A11) satisfy the Lindeberg CLT conditions, see Theorem 11.1.6 in Athreya and Lahiri (2010). By this theorem, we achieve

$$\boldsymbol{\Sigma}_N(\boldsymbol{y})^{-\frac{1}{2}} \cdot (\boldsymbol{L}_N - \boldsymbol{\mu}_N(\boldsymbol{y})) = \sum_{i=1}^{N} \boldsymbol{Z}_{N,i} \xrightarrow{d} \mathcal{N}(0, \mathbb{I}_2),$$

which finishes the proof.

(ii) As discussed in the proof of Proposition 2 (ii), the moment functions $m_1(\boldsymbol{x}, \boldsymbol{y}), m_2(\boldsymbol{x}, \boldsymbol{y})$ in (11) are continuous and bounded in [0,1], and the weak convergence (16) holds. By the weak convergence theorem (Theorem 3.2.3 in Durrett 2019), we have

$$\begin{aligned}\lim_{N\to\infty} N\boldsymbol{\Sigma}_N(\boldsymbol{y}) =& \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{V}(\boldsymbol{x}_i, \boldsymbol{y}) = \lim_{N\to\infty} \int \boldsymbol{V}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}F_{\boldsymbol{X}}^{(N)}(\boldsymbol{x})\\ =& \int \boldsymbol{V}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}F_{\boldsymbol{X}}(\boldsymbol{x}) = \boldsymbol{\Sigma}(\boldsymbol{y}), \quad \text{(A12)}\end{aligned}$$

with $\boldsymbol{\Sigma}(\boldsymbol{y})$ and $\boldsymbol{\Sigma}_N(\boldsymbol{y})$ defined in (17) and (21). Combining (20) and (A12), we obtain Equation (22).

Appendix A.6. Proof of Theorem 2

For the proof of this theorem, a technical lemma is proposed.

Lemma A3. *Suppose that Assumptions 1–3 and 5 hold. Given* $\boldsymbol{Y} = \boldsymbol{y}$*, the random variables* $\boldsymbol{Z}_{N,i}, 1 \le i \le N$ *defined in (A6) satisfy*

$$\sum_{i=1}^{N} \mathbb{E}[\|\boldsymbol{Z}_{N,i}\|^3 | \boldsymbol{Y} = \boldsymbol{y}] \le 4(N\varepsilon)^{-\frac{1}{2}}, \tag{A13}$$

with ε *in Assumption 5.*

Proof. Given $\boldsymbol{Y} = \boldsymbol{y}$, the variables $\boldsymbol{Z}_{N,i}, 1 \le i \le N$ in (A6) are conditionally independent and satisfy $\mathbb{E}[\boldsymbol{Z}_{N,i} | \boldsymbol{Y} = \boldsymbol{y}] = 0, 1 \le i \le N$. Denote by $\mathrm{tr}(A)$ the trace of matrix A. It follows from (A9) in the proof of Theorem 1 that

$$\begin{aligned}\sum_{1}^{N} \mathbb{E}[\|\boldsymbol{Z}_{N,i}\|^2 | \boldsymbol{Y} = \boldsymbol{y}] =& \sum_{1}^{N} \mathbb{E}[\mathrm{tr}(\boldsymbol{Z}_{N,i}\boldsymbol{Z}'_{N,i}) | \boldsymbol{Y} = \boldsymbol{y}] = \sum_{1}^{N} \mathrm{tr}(\mathbb{E}[\boldsymbol{Z}_{N,i}\boldsymbol{Z}'_{N,i} | \boldsymbol{Y} = \boldsymbol{y}]) \\ =& \mathrm{tr}(\mathbb{I}_2) = 2.\end{aligned}$$

Then

$$\begin{aligned}\sum_{1}^{N} \mathbb{E}[\|\boldsymbol{Z}_{N,i}\|^3 | \boldsymbol{Y} = \boldsymbol{y}] \le& \sup_{1 \le i \le N} \{\mathbb{E}[\|\boldsymbol{Z}_{N,i}\| | \boldsymbol{Y} = \boldsymbol{y}]\} \sum_{1}^{N} \mathbb{E}[\|\boldsymbol{Z}_{N,i}\|^2 | \boldsymbol{Y} = \boldsymbol{y}] \\ \le& \sqrt{\frac{4}{N\varepsilon}} \cdot 2 = \frac{4}{\sqrt{N\varepsilon}},\end{aligned}$$

where the second inequality follows from Lemma A2. This finishes the proof. □

Now, we prove Theorem 2. Given $\boldsymbol{Y} = \boldsymbol{y}$, the variables $\boldsymbol{Z}_{N,i}, 1 \le i \le N$ defined in (A6) are conditionally independent by Assumption 3 and satisfy $\mathbb{E}[\boldsymbol{Z}_{N,i} | \boldsymbol{Y} = \boldsymbol{y}] = 0$, $1 \le i \le N$. Further, from Equation (A13) in Lemma A3, the variables $\boldsymbol{Z}_{N,i}, 1 \le i \le N$ satisfy the conditions of the multivariate Berry–Esseen theorem; see Theorem 1.1 in Raič (2019). This theorem implies that

$$\begin{aligned}&\sup_{A \in \mathcal{A}} |\mathbb{P}(\boldsymbol{\Sigma}_N(\boldsymbol{y})^{-\frac{1}{2}}(\boldsymbol{L}_N - \boldsymbol{\mu}_N(\boldsymbol{y})) \in A | \boldsymbol{Y} = \boldsymbol{y}) - \Phi_{(0,\mathbb{I}_2)}(A)| \\ =& \sup_{A \in \mathcal{A}} |\mathbb{P}(\sum_{i=1}^{N} \boldsymbol{Z}_{N,i} \in A | \boldsymbol{Y} = \boldsymbol{y}) - \Phi_{(0,\mathbb{I}_2)}(A)| \\ \le& K \cdot \sum_{i=1}^{N} \mathbb{E}[\|\boldsymbol{Z}_{N,i}\|^3 | \boldsymbol{Y} = \boldsymbol{y}] \le \frac{4K}{\sqrt{N\varepsilon}} \le \frac{264}{\sqrt{N\varepsilon}},\end{aligned}$$

where the first equality is due to (A8), $K = 16 + 42\sqrt[4]{2}$, and $\mathcal{A}$ is a suitable class of subsets of $\boldsymbol{R}^2$.

Define function $h(\boldsymbol{x}) := \boldsymbol{\Sigma}_N(\boldsymbol{y})^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\mu}_N(\boldsymbol{y}))$ and set $A_{\boldsymbol{x}} := \{\boldsymbol{X} : \boldsymbol{X} \le \boldsymbol{x}\}$ for vector $\boldsymbol{x}$. Let random vector $\xi \sim \mathcal{N}(\boldsymbol{\mu}_N(\boldsymbol{y}), \boldsymbol{\Sigma}_N(\boldsymbol{y}))$. Then, it follows that

$$\begin{aligned}&|\mathbb{P}(\boldsymbol{L}_N \le \boldsymbol{x} | \boldsymbol{Y} = \boldsymbol{y}) - \Phi_{(\boldsymbol{\mu}_N(\boldsymbol{y}), \boldsymbol{\Sigma}_N(\boldsymbol{y}))}(\boldsymbol{x})| \\ =& |\mathbb{P}(\boldsymbol{L}_N \in A_{\boldsymbol{x}} | \boldsymbol{Y} = \boldsymbol{y}) - \mathbb{P}(\xi \in A_{\boldsymbol{x}})| \\ =& |\mathbb{P}(h(\boldsymbol{L}_N) \in h(A_{\boldsymbol{x}})) - \mathbb{P}(h(\xi) \in h(A_{\boldsymbol{x}}))| \\ =& |\mathbb{P}(h(\boldsymbol{L}_N) \in h(A_{\boldsymbol{x}})) - \Phi_{(0,\mathbb{I}_2)}(h(A_{\boldsymbol{x}}))| \le \frac{264}{\sqrt{N\varepsilon}}.\end{aligned}$$

Thus,

$$
\begin{aligned}
&|\mathbb{P}(\boldsymbol{L}_N \le \boldsymbol{x}) - \int \Phi_{(\boldsymbol{\mu}_N(\boldsymbol{y}),\boldsymbol{\Sigma}_N(\boldsymbol{y}))}(x)\mathrm{d}F_{\boldsymbol{Y}}(\boldsymbol{y})| \\
&= \left| \int \mathbb{P}(\boldsymbol{L}_N \le \boldsymbol{x}|\boldsymbol{Y} = \boldsymbol{y})\mathrm{d}F_{\boldsymbol{Y}}(\boldsymbol{y}) - \int \Phi_{(\boldsymbol{\mu}_N(\boldsymbol{y}),\boldsymbol{\Sigma}_N(\boldsymbol{y}))}(x)\mathrm{d}F_{\boldsymbol{Y}}(\boldsymbol{y}) \right| \\
&\le \int \left| \mathbb{P}(\boldsymbol{L}_N \le \boldsymbol{x}|\boldsymbol{Y} = \boldsymbol{y}) - \Phi_{(\boldsymbol{\mu}_N(\boldsymbol{y}),\boldsymbol{\Sigma}_N(\boldsymbol{y}))}(x) \right| \mathrm{d}F_{\boldsymbol{Y}}(\boldsymbol{y}) \le \frac{264}{\sqrt{N\varepsilon}}.
\end{aligned}
$$

This finishes the proof.

Appendix A.7. Proof of Lemma 2

Proof. Based on (26) and (27), the functions $m_j(\boldsymbol{x},\boldsymbol{y})$ for $j = 1,2$ in (11) have a specific form

$$
\begin{aligned}
m_j(\boldsymbol{x},\boldsymbol{y}) =& \sum_{k=1}^{K} \left(1 - \bar{q}^{(1)}(\boldsymbol{x},\boldsymbol{y}) - \bar{q}^{(2)}(\boldsymbol{x},\boldsymbol{y})\right)^{k-1} \bar{q}^{(j)}(\boldsymbol{x},\boldsymbol{y}) \cdot \bar{b}_j \frac{K-k+1}{K} e^{-kr} \\
=& \frac{1}{Ke^r} \bar{q}^{(j)}(\boldsymbol{x},\boldsymbol{y}) \bar{b}_j \cdot \sum_{k=1}^{K} (K-k+1) \left(\frac{1 - \bar{q}^{(1)}(\boldsymbol{x},\boldsymbol{y}) - \bar{q}^{(2)}(\boldsymbol{x},\boldsymbol{y})}{e^r} \right)^{k-1} \\
=& \frac{1}{Ke^r} \bar{q}^{(j)}(\boldsymbol{x},\boldsymbol{y}) \bar{b}_j \cdot \frac{K - (K+1)a(\boldsymbol{x},\boldsymbol{y}) + a(\boldsymbol{x},\boldsymbol{y})^{K+1}}{(1 - a(\boldsymbol{x},\boldsymbol{y}))^2},
\end{aligned}
$$

where $a(\boldsymbol{x},\boldsymbol{y}) = (1 - \bar{q}^{(1)}(\boldsymbol{x},\boldsymbol{y}) - \bar{q}^{(2)}(\boldsymbol{x},\boldsymbol{y}))/e^r$. The third equation holds because

$$
\begin{aligned}
\sum_{k=1}^{K} (K-k+1)a^{k-1} =& \frac{1}{1-a} \left[\sum_{k=1}^{K} (K-k+1)a^{k-1} - \sum_{k=1}^{K} (K-k+1)a^k \right] \\
=& \frac{1}{1-a} \left[\sum_{k=0}^{K-1} (K-k)a^k - \sum_{k=1}^{K} (K-k+1)a^k \right] \\
=& \frac{1}{1-a} \left[K - a - \cdots - a^K \right] = \frac{K}{1-a} - \frac{a(1-a^K)}{(1-a)^2} \\
=& \frac{K - (K+1)a + a^{K+1}}{(1-a)^2}.
\end{aligned}
$$

□

Appendix B. Estimation Methodology

Appendix B.1. Maximum Partial Likelihood Estimator of CPH Model

In the following, we detail the estimation procedure for the joint default–prepayment hazard model, whose hazard function is given in the main text by Equation (30).

For each loan i, we define the risk set $R_i = \{\ell : \tau_\ell \ge \tau_i\}$ representing all loans that remain active in the repayment pool prior to the exit time τ_i. To formalize event observation, we specify for each discrete-time period $k \in \{1,\ldots,K\}$ and event type $j \in \{1,2\}$ (where 1 = default and 2 = prepayment), the event-specific cohort:

$$
\mathcal{E}_k^{(j)} = \{\, i : \tau_i = k,\ D_i = j\}.
$$

Let $\boldsymbol{\beta}^{(j)} = (\beta_r^{(j)}, \beta_{HPC}^{(j)}, \beta_{CS}^{(j)}, \beta_{LTV}^{(j)}, \beta_{DTI}^{(j)}, \beta_{FTB}^{(j)}, \beta_{NB}^{(j)})$ denote the parameter vector for event type j. The event-specific partial likelihood incorporates the explicit hazard structure:

$$
L_j(\boldsymbol{\beta}^{(j)}) = \prod_{k=1}^{K} \prod_{i \in \mathcal{E}_k^{(j)}} \frac{q^{(j)}(k, \boldsymbol{x}_i, \boldsymbol{y})}{\sum_{\ell \in R_i} q^{(j)}(k, \boldsymbol{x}_\ell, \boldsymbol{y})}.
$$

This likelihood factorizes into the product of conditional probabilities where each term compares an observed event to all at-risk loans at that event time.

Therefore, for event type j, we estimate $\boldsymbol{\beta}^{(j)}$ by maximizing the log-likelihood

$$l_j(\boldsymbol{\beta}^{(j)}) = \sum_{k=1}^{K} \sum_{i \in \mathcal{E}_k^{(j)}} \left[\eta_i^{(j)}(k) - \log\left(\sum_{\ell \in R_i} e^{\eta_\ell^{(j)}(k)} \right) \right],$$

where the linear predictor $\eta_i^{(j)}(k)$ contains all specified terms:

$$\begin{aligned} \eta_i^{(j)}(k) = \beta_r^{(j)}(r_i - \delta(k)) + \beta_{HPC}^{(j)} HPC(PS_i, k) + \beta_{CS}^{(j)} CS_i \\ + \beta_{LTV}^{(j)} LTV_i + \beta_{DTI}^{(j)} DTI_i + \beta_{FTB}^{(j)} FTB_i + \beta_{NB}^{(j)} NB_i. \end{aligned} \tag{A14}$$

The implementation uses Python's scipy.optimize.minimize with trust-region constraints.

Appendix B.2. Baseline Hazard Iterative Estimator of CPH Model

After obtaining the MPLE $\hat{\boldsymbol{\beta}}^{(j)}$ in Equation (30), the fitted linear predictors $\hat{\eta}_i^{(j)}(k)$ can be obtained by Equation (A14). Initialize survival weights by $s_i(0) = 1$ for every loan i. For each month $k = 1, \ldots, K$ the two baseline hazards $q_0^{(j)}(k)$ are obtained by matching the expected and observed event counts:

$$\begin{cases} \hat{q}_0^{(1)}(k) = \dfrac{\#\mathcal{E}_k^{(1)}}{\sum_{\ell=1}^{N} s_\ell(k-1)\, e^{\hat{\eta}_\ell^{(1)}(k)}}, \\ \hat{q}_0^{(2)}(k) = \dfrac{\#\mathcal{E}_k^{(2)}}{\sum_{\ell=1}^{N} s_\ell(k-1)\, e^{\hat{\eta}_\ell^{(2)}(k)}}, \\ s_i(k) = s_i(k-1)\left[1 - \hat{q}_0^{(1)}(k)\, e^{\hat{\eta}_i^{(1)}(k)} - \hat{q}_0^{(2)}(k)\, e^{\hat{\eta}_i^{(2)}(k)}\right], \quad i = 1, \ldots, N. \end{cases} \tag{A15}$$

Here, $\#\mathcal{E}_k^{(j)}$ denotes the number of loans that exit in month k by event type $j \in \{1, 2\}$. According to (A15), the baseline hazards $\hat{q}_0^{(j)}(k)$ and the survival weights $s_i(k)$ can be directly obtained through an iterative procedure. With the initial condition $s_i(0) = 1$, they are updated month by month, providing a complete estimation of the survival process over the study period.

Appendix B.3. Structure of Realized Systematic Factors

In our empirical analysis, the realized systematic factors $\boldsymbol{Y} = \boldsymbol{y}$ are represented as a multivariate time series that captures the monthly evolution of macroeconomic conditions over the 180-month out-of-sample period from January 2000 to January 2015. Specifically, for each month $k \in \{1, \ldots, K\}$ ($K = 180$), we observe the U.S. Treasury long-term yield $\delta(k)$ and the state-level house price changes $HPC(S_m, k)$ for all d ($d = 50$) U.S. states, as defined in Section 4.1. As a result, the realized $\boldsymbol{y}$ is expressed as a $(d+1)K$ dimension vector obtained by concatenating these monthly systematic factors across all $k = 1, \ldots, K$. Conditional on this realized $\boldsymbol{y}$, the loan losses across individual loans are independent.

Notes

1 Available online: https://www.bis.org/bcbs/basel3.htm?m=76 (accessed on 1 July 2023).

2 The periodical payment R_i satisfies $1 = R_i \sum_{k=1}^{K} e^{-kr_i}$, which yields $R_i = e^{r_i}(1 - e^{-r_i})/(1 - e^{-Kr_i})$. Consequently, the outstanding balance $M_i(k)$ is given by $M_i(k) = R_i \sum_{h=0}^{K-k} e^{-hr_i} = R_i(1 - e^{-(K-k+1)r_i})/(1 - e^{-r_i}) = e^{r_i}(1 - e^{-(K-k+1)r_i})/(1 - e^{-Kr_i})$.

[3] For a positive-definite matrix A with eigen-decomposition $A = Q\Lambda Q'$, define $A^{-1/2} := Q\Lambda^{-1/2}Q'$, where Q is an orthogonal matrix and Λ is a diagonal matrix; see Chapter 2 in Van der Vaart (2000).

[4] Available online: https://www.assetmanagement.hsbc.com.hk/en/intermediary/news-and-insights/residential-mortgage-backed-securities (accessed on 23 March 2025).

[5] The dataset is released by the Federal Home Loan Mortgage Corporation (FHLMC). Available online: https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset (accessed on 1 May 2022).

[6] Available online: https://home.treasury.gov/ (accessed on 1 March 2024).

[7] Available online: https://fred.stlouisfed.org/ (accessed on 9 September 2024).

References

Athreya, Krishna B., and Soumendra N. Lahiri. 2010. *Measure Theory and Probability Theory*. New York: Springer.

Banasik, John, Jonathan N. Crook, and Lyn C. Thomas. 1999. Not if but when will borrowers default. *Journal of the Operational Research Society* 50: 1185–90. [CrossRef]

Bhattacharya, Arnab, Simon P. Wilson, and Refik Soyer. 2019. A bayesian approach to modeling mortgage default and prepayment. *European Journal of Operational Research* 274: 1112–24. [CrossRef]

Brennan, Michael J., and Eduardo S. Schwartz. 1985. Determinants of GNMA mortgage prices. *Real Estate Economics* 13: 209–28. [CrossRef]

Buser, Stephen A., and Patric H. Hendershott. 1984. Pricing default-free fixed-rate mortgages. *Housing Finance Review* 3: 405–29.

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

Campbell, John Y., and Joao F. Cocco. 2015. A model of mortgage default. *The Journal of Finance* 70: 1495–554. [CrossRef]

Cox, John C., Jonathan E. Ingersoll, and Stephen A. Ross. 1985. A theory of the term structure of interest rates. *Econometrica* 53: 385–407. [CrossRef]

Credit Suisse Financial Products. 1997. *CreditRisk+: A Credit Risk Management Framework*. Credit Suisse First Boston. Available online: https://globalriskguard.com/resources/credit/creditrisk.pdf (accessed on 1 October 2023).

Cunningham, Donald F., and Patric H. Hendershott. 1984. Pricing FHA mortgage default insurance. *Housing Finance Review 3*: 373–92.

Deng, Yongheng, John M. Quigley, and Robert Van Order. 2000. Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica* 68: 275–307. [CrossRef]

Deng, Yongheng, John M. Quigley, Robert Van Order, and Freddie Mac. 1996. Mortgage default and low downpayment loans: The costs of public subsidy. *Regional Science and Urban Economics* 26: 263–85. [CrossRef]

Deshmukh, Shailaja. 2012. *Multiple Decrement Models in Insurance: An Introduction Using R*. New Delhi: Springer.

Dunn, Kenneth B., and John J. McConnell. 1981. Valuation of GNMA mortgage-backed securities. *The Journal of Finance* 36: 599–616. [CrossRef]

Durrett, Rick. 2019. *Probability: Theory and Examples*, 5th ed. New York: Cambridge University Press.

Epperson, James F., James B. Kau, Donald C. Keenan, and Walter J. Muller, III. 1985. Pricing default risk in mortgages. *Real Estate Economics* 13: 261–72. [CrossRef]

Finger, Christopher C. 1999. Conditional approaches for CreditMetrics portfolio distributions. *Credit Metrics Monitor* 2: 14–33.

Flores, Jesús Alan Elizondo, Valeria Álvarez Navarro, and Israel Sergio Valladares Cedillo. 2010. An actuarial approach to pricing Mortgage Insurance considering simultaneously mortgage default and prepayment. Paper Presented at International Congress of Actuaries, Cape Town, South Africa, March 7–12; Available online: https://actuaries.org/resources-post/an-actuarial-approach-to-pricing-mortgage-insurance-considering-simultaneously-mortgage-default-andprepayment/ (accessed on 8 May 2025).

Foster, Chester, and Robert Van Order. 1984. An option-based model of mortgage default. *Housing Finance Review* 3: 351–68.

Frey, Rüdiger, and Alexander J. McNeil. 2003. Dependent defaults in models of portfolio credit risk. *Journal of Risk* 6: 59–92. [CrossRef]

Giesecke, Kay, Konstantinos Spiliopoulos, and Richard B. Sowers. 2013. Default clustering in large portfolios: Typical events. *The Annals of Applied Probability* 23: 348–85. [CrossRef]

Giesecke, Kay, Konstantinos Spiliopoulos, Richard B. Sowers, and Justin A. Sirignano. 2015. Large portfolio asymptotics for loss from default. *Mathematical Finance* 25: 77–114. [CrossRef]

Goodman, Laurie, and Jun Zhu. 2015. Loss Severity on Residential Mortgages: Evidence from Freddie Mac's Newest Data. *The Journal of Fixed Income* 25: 48–57. [CrossRef]

Goodman, Laurie S., Brian Landy, Roger Ashworth, and Lidan Yang. 2014. A look at freddie mac's loan-level credit performance data. *Journal of Structured Finance* 19: 52–61. [CrossRef]

Gordy, Michael B. 2003. A risk-factor model foundation for ratings-based bank capital rules. *Journal of Financial Intermediation* 12: 199–232. [CrossRef]

Jarrow, Robert A., David Lando, and Stuart M. Turnbull. 1997. A markov model for the term structure of credit risk spreads. *The Review of Financial Studies* 10: 481–523. [CrossRef]

Jones, Chris, and Xinfu Chen. 2016. Optimal mortgage prepayment under the Cox–Ingersoll–Ross model. *SIAM Journal on Financial Mathematics* 7: 552–66. [CrossRef]

Li, Zhiyong, Aimin Li, Anthony Bellotti, and Xiao Yao. 2023. The profitability of online loans: A competing risks analysis on default and prepayment. *European Journal of Operational Research* 306: 968–85. [CrossRef]

Merton, Robert C. 1974. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance* 29: 449–70.

Moody's Investors Service. 2024. Moody's Approach to Rating US RMBS Using the MILAN Framework. Rating methodology Residential MBS, Moody's Investors Service. Available online: https://www.moodys.com/research/Rating-Methodology-Moodys-Approach-to-Rating-US-RMBS-Using-the-MILAN-Rating-Methodology--PBS_1411650 (accessed on 3 August 2025).

Munk, Claus. 2011. *Fixed Income Modelling*. New York: Oxford University Press.

Quercia, Roberto, and Jonathan Spader. 2008. Does homeownership counseling affect the prepayment and default behavior of affordable mortgage borrowers? *Journal of Policy Analysis and Management* 27: 304–25. [CrossRef]

Quercia, Roberto G., and Michael A. Stegman. 1992. Residential mortgage default: A review of the literature. *Journal of Housing Research* 3: 341–79.

Quigley, John M., and Robert Van Order. 1990. Efficiency in the mortgage market: The borrower's perspective. *Real Estate Economics* 18: 237–52. [CrossRef]

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

Raič, Martin. 2019. A multivariate berry-esseen theorem with explicit constants. *Bernoulli* 25: 2824–53. [CrossRef]

Richard, Scott F., and Richard Roll. 1989. Prepayments of fixed-rate mortgage-backed securities. *Journal of Portfolio Management* 15: 73–82. [CrossRef]

Rotar, Vladimir I. 2014. *Actuarial Models: The Mathematics of Insurance*, 2nd ed. Boca Raton: CRC Press.

Rudin, Walter. 1987. *Real and Complex Analysis*, 3rd ed. New York: McGraw-Hill.

Schwartz, Eduardo S., and Walter N. Torous. 1989. Prepayment and the valuation of mortgage-backed securities. *The Journal of Finance* 44: 375–92. [CrossRef]

Sirignano, Justin, and Kay Giesecke. 2019. Risk analysis for large pools of loans. *Management Science* 65: 107–21. [CrossRef]

Sirignano, Justin A., Gerry Tsoukalas, and Kay Giesecke. 2016. Large-scale loan portfolio selection. *Operations Research* 64: 1239–55. [CrossRef]

Stein, Roger M., Ashish Das, Yufeng Ding, and Shirish Chinchalkar. 2011. *Mortgage Portfolio Analyzer: A Quasi-Structural Model of Mortgage Portfolio Losses*. Working Paper. New York: Moody's Research Labs.

Steinbuks, Jevgenijs. 2015. Effects of prepayment regulations on termination of subprime mortgages. *Journal of Banking & Finance* 59: 445–56. [CrossRef]

Stepanova, Maria, and Lyn Thomas. 2002. Survival analysis methods for personal loan data. *Operations Research* 50: 277–89. [CrossRef]

Thackham, Mark, and Jun Ma. 2022. On maximum likelihood estimation of competing risks using the cause-specific semi-parametric cox model with time-varying covariates–an application to credit risk. *Journal of the Operational Research Society* 73: 5–14. [CrossRef]

Van der Vaart, Aad W. 2000. *Asymptotic Statistics*. New York: Cambridge University Press.

Vasicek, Oldrich. 1991. *Limiting Loan Loss Probability Distribution*. San Francisco: KMV Corporation.

Zhang, Nailong, Qingyu Yang, Aidan Kelleher, and Wujun Si. 2019. A new mixture cure model under competing risks to score online consumer loans. *Quantitative Finance* 19: 1243–53. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.