EFFICIENT RISK ANALYSIS OF MORTGAGE POOLS

JUSTIN A. SIRIGNANO AND KAY GIESECKE

Abstract. Financial institutions, government-sponsored enterprises such as Freddie Mac, and mortgage-backed security investors are often exposed to default and prepayment risk from large numbers of mortgages. Due to the size of mortgage pools and the long maturities of their cashflows, the measurement and management of these exposures is computationally expensive. This paper develops and tests efficient numerical methods for the risk analysis of large mortgage pools. For a broad class of dynamic default and prepayment models, we develop a law of large numbers and a central limit theorem for the loss and prepayment processes in a pool. The asymptotics are then used to construct efficient Monte Carlo approximations of the default and prepayment distributions for a large pool. The approximations aggregate the full loan-level dynamics, making it possible to take advantage of the detailed loan-level data often available. To demonstrate the effectiveness of our approach, we implement it on a data set of over 25 million actual subprime and agency mortgages. The results show the accuracy and speed of the approximation in comparison to brute-force simulation of a pool, for a variety of pools with different risk characteristics. Computational cost is often several orders of magnitude less than brute-force simulation of the actual pool with a similar level of accuracy. Furthermore, the computational expense of our efficient Monte Carlo approximation is constant no matter the dimension of the loan-level features; this is key since loan-level data is often high-dimensional. The approach is also directly applicable to many other types of loans which are commonly pooled together (for instance, credit card, auto, and commercial loans).

1. Introduction

The unprecedented level of subprime mortgage defaults beginning in 2007 was largely responsible for the ensuing financial crisis. Today, there is widespread recognition amongst both market participants and regulators that better models of mortgage default and prepayment behavior are needed. In particular, more accurate risk analysis can better inform capital requirements for the banking system, improve mortgage-backed security (MBS) ratings, and help financial institutions that hold or trade mortgages (as well as many other types of loans) better manage risk. The main challenges include developing more accurate models, methods to fit these models to the massive amounts of available data, and ways to efficiently simulate the very large pools of loans common in practice.

Mortgage pools are computationally challenging to analyze due to their size and the long maturities of their cashflows. Mortgage-backed securities typically have thousands to hundreds of thousands of mortgages, many of which might have 30 year terms. The MBS market is also relatively liquid; for instance, the daily volume for the agency MBS market is, on average, 300 billion dollars. There is often a need for computational speed; a typical mortgage trading desk at a major bank will on a daily basis need to price thousands of mortgage-backed securities and hundreds of collateralized mortgage obligations (CMOs) backed by mortgage pools. The government-sponsored enterprises (GSEs) Freddie Mac and Fannie Mae have credit exposure to roughly 25 million mortgages, either through directly owning the mortgages or providing credit guarantees against default for the mortgages. Major US banks can service up to ten million mortgages and, even though they do not directly own the loans, mortgage servicing cashflows are still strongly affected by default and prepayment risk. Moreover, a major US bank might directly own a million mortgages. These institutions need risk management tools for their large loan portfolios and investors require methods to price and analyze risk for MBSs and CMOs.

Date: March 22, 2015. The authors gratefully acknowledge support from the National Science Foundation through Methodology, Measurement, and Statistics Grant No. 1325031. We thank seminar participants at the International Monetary Fund for comments. Special thanks are due to Kostas Spiliopoulos for insightful comments. We would also like to thank participants at the 2014 Annual INFORMS Meeting and the 2014 SIAM Conference on Financial Mathematics and Engineering, where this paper won the 2014 SIAM Financial Mathematics and Engineering Conference Paper Prize.

This paper develops and tests efficient numerical methods for the analysis of large, heterogeneous mortgage pools. We focus on a broad class of dynamic, discrete-time models of loan-by-loan default and prepayment timing. The functions for the transition probabilities are allowed to be very general and could come from a range of statistical or machine learning models such as generalized linear models (for example, logistic regression), support vector machines, neural networks, and decision trees. For this important class of models, we develop a law of large numbers and a central limit theorem for the pool loss and prepayment processes. The convergence results are then used to construct efficient Monte Carlo approximations of the default and prepayment distributions for a large pool. The approximations account for the full loan-level dynamics, making it possible to take advantage of the detailed loan-level data often available for mortgage pools. Very importantly, the cost of our approximation remains constant no matter the dimension of the loan-level features. This is essential since loan-level data is often high-dimensional (the dimension could be in the hundreds). Furthermore, since the law of large numbers and central limit theorem are dynamic, the approximation provides the default and prepayment distribution across all time horizons at no extra computational cost.

In order to demonstrate the accuracy and low computational expense of our approximations, we numerically test our approach on a large, loan-level mortgage data set which includes over 10 million subprime mortgages and 16 million agency mortgages. We compare the approximate distribution with the true distribution (obtained by brute-force simulation) for various pools drawn from this data set. The comparison is performed using model parameters fitted to the data set. Therefore, the numerical studies reflect the actual performance of the approximation method with real data from actual mortgage pools. The approximation's computational cost is often several orders of magnitude less than the cost of brute-force simulation of the actual pool and has a similar level of accuracy. Although the approximation's cost savings grow with the size of the pool, it is highly accurate even for a pool having as little as 500 loans.

The methods that we propose work generally for any type of pooled loan. Such large pools are common for auto loans, student loans, credit cards, commercial loans, and wholesale loans. These pools can be very large in size and may either be held by the lender or securitized and then held by an investor. For instance, a major US bank might easily have on the order of 20,000 wholesale loans and 100,000 mid-market and commercial loans. There is roughly 3 trillion dollars of outstanding consumer credit in the United States and a major credit card company can have hundreds of millions of credit card accounts. Over half of all consumer credit is eventually securitized, and each deal can consist of tens of millions of credit card accounts. Like mortgages, all of these loans have loan-level features carrying important information on their default risk. Our approach can tractably handle such large pools while also taking full advantage of the detailed loan-level information available.

1.1. **Literature.**

The residential MBS market is massive, with over 50 million outstanding mortgage loans at any one time and a total value of 14 trillion dollars. For comparison, the S&P 500 has a market value of 15 trillion dollars and the corporate credit market has a value of 7 trillion dollars. Despite the clear importance of mortgages, relatively little attention has historically been paid in the academic community to mortgages. [1] proposes a loan-by-loan prepayment model which treats prepayment as an American option and uses this model to price agency mortgage-backed securities by assuming a homogeneous pool. This option-based approach is further developed in [2], [3] and [4], amongst others. Prepayment and default have also been modeled at the loan-level using a reduced-form approach, see [5], [6], [7] and [8].

The computational expense associated with loan-by-loan models for large pools is widely recognized. Prior research has analyzed several approaches to tackle this issue. [9], [10], and [11] propose and implement distributed simulation (i.e., parallel computing) of mortgage-backed securities using clusters of computers. [12], [13], [14], [15] and [16] develop top-down models of mortgage pool behavior. The models are formulated without reference to the constituent mortgages and only model the pool in aggregate. Although top-down models can incorporate pool-level characteristics (such as the average coupon rate), they ignore much of the information available from the features of each loan such as credit score, loan-to-value, debt-to-income, interest rate, size of the loan, type of property, and many more. [17] finds that top-down models can be significantly inaccurate due to ignoring the loan-level constitution of a pool. The Monte Carlo approximation that we develop for loan-by-loan models is as tractable as a top-down model while taking full advantage of the loan-level feature information.

Laws of large numbers and central limit theorems have previously been proven for pools of loans with default risk in other model frameworks; see [18], [19], [20], [21], [22], [23], [24], and others. In contrast to this literature, we consider a discrete-time formulation that is natural given the data structure common in practice, where events are reported on a monthly or quarterly basis. Moreover, our formulation is well-adapted to settings with high-dimensional loan-level feature data, which are typically hard to treat using earlier approaches. Finally, unlike the aforementioned papers, we allow for "competing risks," i.e., multiple types of events (e.g., default, prepayment, and potentially others).

1.2. **Structure of the paper.** The class of models we consider is described in Section 2. The law of large numbers and central limit theorem for this class of models is presented in Section 3. These limiting laws are used to develop an efficient Monte Carlo approximation for the pool loss and prepayment processes. Section 4 provides details on how to estimate the parameters for the class of models considered and presents certain theoretical properties for these estimators. Then, the model is fitted and the efficient Monte Carlo approximation is tested on actual mortgage data from our data set. Section 5 describes the mortgage data we use and Section 6 contains the various numerical studies. All proofs can be found in Appendix A. Some additional tables that may be useful for reference are placed in Appendix B.

2. Model Framework

We analyze a broad family of dynamic loan-by-loan models for mortgage default and prepayment in a pool at times $t \in I = \{0, 1, \ldots, T\}$. We fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an information filtration $(\mathcal{F}_t)_{t\in I}$. $\mathbb{P}$ is the actual probability measure. The total number of mortgages initially in the pool is N, spread across a finite number of geographic locations.[1] The process $U^n = (U^n_t)_{t\in I}$ prescribes the state of the n-th mortgage. The variable U^n_t takes values in a finite discrete space $\mathcal{U}$. A typical model would include states for when the mortgage is still outstanding, has been prepaid, or is in default.

We allow for idiosyncratic, systematic, and contagion factors to influence the dynamics of U^n. Each mortgage has an $\mathcal{F}_0$-measurable loan-level covariate vector $Y^n \in \mathcal{Y} \subseteq \mathbb{R}^{d_Y}$,[2] which can contain static variables such as loan-to-value (LTV) ratio, FICO score,[3] geographic location, type of loan, and historical loan performance up until the initial time of interest $t = 0$ (for instance, how many days behind payment the mortgage is or whether it is in foreclosure). These loan-level factors are specific to each mortgage or loan and are sources of idiosyncratic risk. We also consider exogenous, systematic risk factors V which will have a common influence across many mortgages in the pool. The vector process $V = (V_t)_{t\in I}$, where $V_t \in \mathbb{R}^{d_V}$, might represent the behavior of local and national economic conditions such as the unemployment rate, housing prices, and mortgage rates.[4] Finally, we allow for contagion risk factors. Define the "mean-field" process $H^N = (H^N_t)_{t\in I}$ as:

$$H^N_t = \frac{1}{N}\sum_{n=1}^{N} f(U^n_t, Y^n), \tag{1}$$

where $f = \big(f_1, \ldots, f_K\big)$ and $f_k : \mathcal{U} \times \mathbb{R}^{d_Y} \mapsto \mathbb{R}$. A specific application for the mean field term would be to model a contagion effect where past defaults of mortgages in areas geographically close to the n-th mortgage increase the likelihood of the n-th mortgage defaulting. The process H^N would then keep track of the number of defaults at each geographic location. (See Example 2.1 for a concrete formulation.) In light of the mortgage meltdown, such a feedback mechanism has been supported by several recent empirical papers; see [25], [26], [27], [28], and [29].

The dynamics of U^n are prescribed by the transition function:

$$\mathbb{P}_\theta[U^n_t = u|\mathcal{F}_{t-1}] = h_\theta(u, U^n_{t-1}, Y^n, V_{t'<t}, H^N_{t'<t}), \quad t \geq 1, \tag{2}$$

where $V_{t'<t} = (V_0, V_1, \ldots, V_{t-1})$ is the path of the systematic factor, and $H^N_{t'<t} = (H^N_0, H^N_1, \ldots, H^N_{t-1})$ is the path of the mean-field term. The function h_θ is specified by a parameter θ, which takes values in the

[1] Geographic locations could be state, ZIP code, or census tract.

[2] Y^n includes both continuous and categorical variables. Categorical variables are encoded as a vector whose elements are each in $\{0, 1\}$. Of course, $\{0, 1\}$ is a subset of the real line, so $\mathcal{Y} \subseteq \mathbb{R}^{d_Y}$.

[3] The widely used credit score from Fair, Isaac, and Company.

[4] The time t can also be included in the systematic factor vector V.

compact Euclidean space Θ, and which must be estimated from data on mortgage performance. Equation (2) gives the marginal probability for the transitions of the mortgages from their state at time $t-1$ to time t. Furthermore, we stipulate that conditional on $\mathcal{F}_{t-1}$, the states $U_t^1, \ldots, U_t^N$ are independent. This fully specifies the dynamics of the mortgages $n = 1, \ldots, N$.

The dynamics of $U^1, \ldots, U^N$ can potentially depend upon the entire history of the mean field term H^N and the systematic risk V. For instance, a well-known phenomenon for mortgage prepayments is the burnout effect where the prepayment rate is lower the second time mortgage rates fall. The mortgage holders who pay close attention to refinancing opportunities will prepay the first time mortgage rates fall. Therefore, the second time mortgage rates fall, the remaining pool will have a lower prepayment rate than the first time. This effect can be modeled through the inclusion of the history of V in h_θ.

In addition, it is worthwhile noting that the model allows for correlation between defaults and prepayment. Their joint modeling can become essential during times of financial duress. For instance, during the Financial Crisis, even though interest rates were relatively low, prepayment levels were very low. Default rates will typically be higher when prepayment rates are lower, and vice versa.

A typical choice for h_θ might be a generalized linear model (GLM), although many other choices are available. An example of a GLM is logistic regression. The results in the paper require only very mild assumptions on the form of the function h_θ; these assumptions are satisfied by many standard models such as generalized linear models and neural networks.

Example 2.1 (**Logistic regression model**). *Consider a single zip code (i.e., one geographic location) and let $\mathcal{U}$ be the state space $\{Outstanding, Default, Prepaid\}$. Let Y^n be a vector containing the FICO score, LTV ratio, and earliest interest rate for the n-th mortgage and let V_t be the unemployment rate, housing price index, and mortgage rate in the zip code at time t. Finally, let L^N_{t-1} be the fraction of mortgages which defaulted in the $(t-1)$-th period in the zip code and specify $U^n_t = \{outstanding\}$ for $t \leq 0$. Note that L^N_t is a function of $H^N_{t'<t}$ since:*

$$L^N_{t-1} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{1}_{U^n_{t-1}=\{default\}} - \mathbf{1}_{U^n_{t-2}=\{default\}}). \tag{3}$$

We choose logistic function for h_θ. Let the parameter vector $\theta = (\theta_{default}, \theta_{prepay})$ and $X^n_{t-1} = (Y^n, V_{t-1}, L^N_{t-1})$. Then, the dynamics for the n-th mortgage are

$$\begin{aligned}
h_\theta(default, outstanding, Y^n, V_{t-1}, H^N_{t'<t}) &= \frac{e^{\theta_{default}\cdot X^n_{t-1}}}{1 + e^{\theta_{default}\cdot X^n_{t-1}} + e^{\theta_{prepay}\cdot X^n_{t-1}}}, \\
h_\theta(prepaid, outstanding, Y^n, V_{t-1}, H^N_{t'<t}) &= \frac{e^{\theta_{prepay}\cdot X^n_{t-1}}}{1 + e^{\theta_{default}\cdot X^n_{t-1}} + e^{\theta_{prepay}\cdot X^n_{t-1}}}, \\
h_\theta(outstanding, outstanding, Y^n, V_{t-1}, H^N_{t'<t}) &= \frac{1}{1 + e^{\theta_{default}\cdot X^n_{t-1}} + e^{\theta_{prepay}\cdot X^n_{t-1}}}.
\end{aligned} \tag{4}$$

Of course, $h_\theta(outstanding, default, y, v, H) = h_\theta(outstanding, prepaid, y, v, H) = h_\theta(prepaid, default, y, v, H) = h_\theta(default, prepaid, y, v, H) = 0$.

Example 2.2 (**Neural network model**). *Let the elements of $\mathcal{U}$ be $u_1, \ldots, u_K$. If a mortgage is in state u_k, denote the states to which it can transition as $u_k^1, \ldots, u_k^{N_k}$. Then,*

$$\begin{aligned}
h_\theta(u_n, u_k, Y^n, V_{t'<t}, H^N_{t'<t}) &= \frac{e^{\sigma_\theta(u_n, u_k, Y^n, V_{t'<t}, H^N_{t'<t})}}{1 + \sum_{n'=1}^{N_k-1} e^{\sigma_\theta(u_k^{n'}, u_k, Y^n, V_{t'<t}, H^N_{t'<t})}}, \quad n < N_k, \\
h_\theta(u_{N_k}, u_k, Y^n, V_{t'<t}, H^N_{t'<t}) &= 1 - \sum_{n'=1}^{N_k-1} h_\theta(u_k^{n'}, u_k, Y^n, V_{t'<t}, H^N_{t'<t}).
\end{aligned} \tag{5}$$

The function σ_θ is a neural network with a single hidden layer:

$$\sigma_\theta(u_n, u_k, Y^n, V_{t'<t}, H^N_{t'<t}) = W^2_\theta q(W^1_\theta Q^n), \tag{6}$$

where $Q^n = (u_n, u_k, Y^n, V_{t'<t}, H^N_{t'<t}) \in \mathbb{R}^{d_Q \times 1}$, and $W^1_\theta \in \mathbb{R}^{d_h \times d_Q}$ and $W^2_\theta \in \mathbb{R}^{1 \times d_h}$ are weights on the input Q^n and the output of the hidden layer q, respectively. The hidden layer $q : \mathbb{R}^{d_h \times 1} \mapsto \mathbb{R}^{d_h \times 1}$ is the function

$q(x) = (q_1(x_1), q_2(x_2), \ldots, q_{d_h}(x_{d_h}))^\top$. Typical choices for $q_1, \ldots, q_{d_h}$ are sigmoid functions and d_h is the number of "neurons" in the hidden layer. See [30] for more details on neural networks.

Example 2.3 (Decision tree model). *Divide the space $\mathcal{U} \times \mathcal{Y} \times \mathbb{R}^{d_V} \times \mathbb{R}^K$ into the regions $\{R_\theta^m\}_{m=1}^M$, with the condition that the regions $\{R_\theta^m\}_{m=1}^M$ are disjoint and their union covers the entire space. The transition probability h_θ then is*

$$h_\theta(u, u', Y^n, V_{t'<t}, H_{t'<t}^N) = \sum_{m=1}^M p_\theta^m(u, u') \mathbf{1}_{(u', Y^n, V_{t'<t}, H_{t'<t}^N) \in R_\theta^m}, \tag{7}$$

where $\sum_u p_\theta^m(u, u') = 1$ for $m = 1, \ldots, M$. The regions $\{R_\theta^m\}_{m=1}^M$ are chosen by sequentially splitting the space via recursive binary splitting. See [30] for more details on decision trees.

Both Examples 2.1 and 2.2 are implemented in our empirical paper [31] and fitted to actual mortgage data. We find that the choice of a logistic regression model for h_θ performs as well as the more complicated neural network.

3. Limiting Laws and an Efficient Monte Carlo Approximation

Risk analysis for the large pools of mortgages common in practice is challenging. Due to their large size, brute-force simulation of entire pools is computationally expensive. We develop an efficient Monte Carlo approximation for the loss and prepayment levels in a pool. The approximation is based on a law of large numbers (LLN) and a central limit theorem (CLT).

3.1. Weak convergence results.

We introduce some notation. Define $\mu_t^N \in B = \mathcal{P}(\mathcal{U} \times \mathbb{R}^{d_Y})$ as the empirical measure of the processes $(U_t^1, Y^1), \ldots, (U_t^N, Y^N)$ at time t, where $\mathcal{P}$ is the space of probability measures. Formally,

$$\mu_t^N = \frac{1}{N} \sum_{n=1}^N \delta_{(U_t^n, Y^n)}. \tag{8}$$

The vector process $H^N \in \mathbb{R}^K$ can be expressed in terms of the empirical measure:

$$H_t^N = \sum_{u \in \mathcal{U}} \int_{\mathbb{R}^{d_Y}} f(u, y) \mu_t^N(u, dy). \tag{9}$$

With this notation out of the way, it is now possible to prove a law of large numbers for the empirical measure μ_t^N. The proofs for this and the other theorems in the paper are given in Appendix A.

Assumption 3.1. *Suppose that μ_0^N weakly converges to $\bar{\mu}_0$, where $\bar{\mu}_0$ is deterministic, and that the distribution of Y^n has compact support on $\mathbb{R}^{d_Y}$. Also, the functions $h, f_1, \ldots, f_K$ are continuous and bounded.*[5]

Theorem 3.2. *Provided Assumption 3.1, the empirical measure μ^N weakly converges to $\bar{\mu}$ in B^{T+1} as $N \longrightarrow \infty$, where $\bar{\mu}$ satisfies the equation:*

$$\bar{\mu}_t(u, dy) \quad = \quad \sum_{u' \in \mathcal{U}} h_\theta(u, u', y, V_{t'<t}, \bar{H}_{t'<t}) \bar{\mu}_{t-1}(u', dy), \tag{10}$$

and $\bar{H}_t = \sum_{u \in \mathcal{U}} \int_{\mathbb{R}^{d_Y}} f(u, y) \bar{\mu}_t(u, dy)$.

It is important to note that the law of large numbers is dynamic and is also a random equation; randomness enters through the covariate V. The law of large numbers has a natural link with the original model (2). The function h_θ from (2) appears in the law of large numbers. If the model includes dependence on the past pool dynamics through $\bar{H}_{t'<t}$, the law of large numbers is nonlinear.

The law of large numbers can also be supplemented with a central limit theorem. Define the empirical fluctuation process $\Xi_t^N = \sqrt{N}(\mu_t^N - \bar{\mu}_t) \in W = D'(\mathcal{U} \times \mathbb{R}^{d_Y})$.[6] The central limit theorem satisfies an equation

[5] The assumption that the functions $f_k \in C_b(\mathbb{R}^{d_Y})$ still allows H_t^N to keep track of the fraction of the pool in any combination of performance states in $\mathcal{U}$ and categorical states in $\mathcal{Y}$. For example, H_t^N could track the fraction of the pool which has defaulted at each geographic location as well as the fraction of the pool which has defaulted for each mortgage product type (fixed rate, adjustable-rate, interest-only, etc.). The loss in the last time period in each of these categories would simply be $H_{t-1}^N - H_{t-2}^N$.

[6] D' is the space of distributions.

linearized around the nonlinear dynamics of the law of large numbers. Randomness enters both through V and a martingale term $\bar{\mathcal{M}}$. Like the law of large numbers, the central limit theorem is also dynamic.

Assumption 3.3. *Assume that $\sqrt{N}(\mu_0^N - \bar{\mu}_0)$ weakly converges to $\bar{\Xi}_0$. In addition, assume that $h_\theta(u, u', v, y, H)$ is twice-differentiable in H and its second derivative is bounded.*[7]

Theorem 3.4. *Provided Assumptions 3.1 and 3.3, Ξ^N weakly converges to $\bar{\Xi}$ in W^{T+1} as $N \longrightarrow \infty$, where $\bar{\Xi}$ satisfies the equation:*

$$\begin{aligned} \bar{\Xi}_t(u, dy) &= \sum_{u' \in \mathcal{U}} h_\theta(u, u', y, V_{t'<t}, \bar{H}_{t'<t}) \bar{\Xi}_{t-1}(u', dy) \\ &+ \sum_{u' \in \mathcal{U}} \big(\frac{\partial}{\partial H} h_\theta(u, u', y, V_{t'<t}, \bar{H}_{t'<t}) \cdot \bar{E}_{t'<t} \big) \bar{\mu}_{t-1}(u', dy) + \bar{\mathcal{M}}_t(u, dy), \end{aligned} \tag{11}$$

where $E_t^N = \sum_{u \in \mathcal{U}} \int_{\mathbb{R}^{d_Y}} f(u, y) \Xi_t^N(u, dy)$ and $\bar{E}_t = \sum_{u \in \mathcal{U}} \int_{\mathbb{R}^{d_Y}} f(u, y) \bar{\Xi}_t(u, dy)$. Given V, $\bar{\mathcal{M}}(u, dy)$ is a conditionally Gaussian process with zero mean and covariance:

$$Cov\Big[\bar{\mathcal{M}}_t(u_1, dy), \bar{\mathcal{M}}_t(u_2, dy)\Big] = -\sum_{u' \in \mathcal{U}} h_\theta(u_1, u', y, V_{t'<t}, \bar{H}_{t'<t}) h_\theta(u_2, u', y, V_{t'<t}, \bar{H}_{t'<t}) \bar{\mu}_{t-1}(u', dy),$$

$$Var\Big[\bar{\mathcal{M}}_t(u, dy)\Big] = \sum_{u' \in \mathcal{U}} h_\theta(u, u', y, V_{t'<t}, \bar{H}_{t'<t})(1 - h_\theta(u, u', y, V_{t'<t}, \bar{H}_{t'<t}) \bar{\mu}_{t-1}(u', dy),$$

where $u_1 \neq u_2$.

The law of large numbers and central limit theorem can be combined to form an approximation for a finite pool of N mortgages:[8]

$$\mu^N = \bar{\mu} + \frac{1}{\sqrt{N}} \Xi^N \overset{d}{\approx} \bar{\mu}^N = \bar{\mu} + \frac{1}{\sqrt{N}} \bar{\Xi}. \tag{12}$$

The LLN $\bar{\mu}$ is a first-order approximation of μ^N while the CLT $\bar{\Xi}$ provides a second-order correction. The large pool approximation (12) is conditionally Gaussian given V and can be utilized to simulate large pools of mortgages for the purposes of pricing MBSs and CMOs,[9] forecasting default conditions in different geographic areas, and risk management for banks and other financial institutions which hold large quantities of mortgages (or any other type of loan).

Simulation of the approximation (12) is straightforward. First, many paths are simulated from the systematic process V. Conditional on each path of V, $\bar{\mu}$ can be calculated deterministically and the conditional covariance of $\bar{\Xi}$ can be evaluated in closed-form. (For details, see Section 3.2 below.) Since $\bar{\Xi}$ is conditionally Gaussian, this directly yields the conditional distribution of $\bar{\Xi}$. Then, the unconditional distribution for the approximation $\bar{\mu}^N$ can easily be found by averaging the conditional distributions across the paths of V since $\mathbb{P}_\theta[\bar{\mu}^N \in A] = \mathbb{E}_\theta[\mathbb{P}_\theta[\bar{\mu}^N \in A | V]]$.

The ideas and results in this paper, which are developed for discrete time, can be easily extended to a continuous time framework where default and prepayment events are modeled in continuous time as counting processes with intensities governed by h_θ. It is easy to find the law of large numbers and central limit theorem; they satisfy a random ODE and an SDE, respectively. Their forms are very similar to the forms given above for the discrete-time model considered in this paper. In fact, the discrete-time limiting laws presented here are the Euler discretization for the continuous time limiting laws.

3.2. **Simulating the Approximation.** The numerical evaluation of $\bar{\mu}$ and $\bar{\Xi}$ conditional on each path of V requires discretizing $\mathbb{R}^{d_Y}$ and then calculating $\bar{\mu}_t(u, dy)$ and $\bar{\Xi}_t(u, dy)$ at each grid point. Let $\mathcal{R} \in \mathbb{R}^{d_Y}$ be the set of grid points. The initial measures $\bar{\mu}_0$ would be calibrated to the data of the pool and the second-order correction $\bar{\Xi}_0$ would be set to zero. For each V, $\bar{\Xi}_t$'s distribution on the set of grid points can be computed in closed-form (since it is conditionally Gaussian). The approach is outlined below:

[7] This is not a stringent assumption and is satisfied by a wide range of models used in practice. For instance, logistic regression and neural networks both satisfy this assumption.

[8] The addition of $\bar{\mu}$ and $\bar{\Xi}$ for the approximation is rigorous since it is proven in Appendix A that conditionally on each V $\mu^N \xrightarrow{p} \bar{\mu}$ and $\Xi^N \xrightarrow{d} \bar{\Xi}$.

[9] In this application setting, the reference measure $\mathbb{P}$ would be a risk-neutral pricing measure.

- Simulate paths $V^1, \dots, V^L$ from the random variable V. For each V^l:
  - Given V, $\bar{\mu}$ is deterministic. Calculate $\bar{\mu}^l_t(u, dy_i)$ for all $y_i \in \mathcal{R}$ at the times $t = 1, \dots, T$. Approximate $\bar{H}_{t'<t}$ by quadrature.
  - Given V, $\bar{\Xi}$ is a zero mean Gaussian; therefore, its conditional distribution is completely described by its conditional covariance. Approximate $\bar{E}_{t'<t-1}$ by quadrature. Then, provided $Cov[\bar{\Xi}_{t-1}(u, dy_i), \bar{\Xi}_{t-1}(u', dy_j)|V^l]$ and $Cov[\bar{\Xi}_{t-1}(u, dy_i), \bar{E}_{t'<t-1}|V^l]$ for all $y_i, y_j \in \mathcal{R}$ and $u, u' \in \mathcal{U}$, it is easy to calculate $Cov[\bar{\Xi}_t(u, dy_i), \bar{\Xi}_t(u', dy_j)|V^l]$ and $Cov[\bar{\Xi}_t(u, dy_i), \bar{E}_{t'<t}|V^l]$ *in closed-form*. If there is no mean field term, it is only necessary to calculate at each time t the covariance $Cov[\bar{\Xi}_t(u, dy_i), \bar{\Xi}_t(u', dy_i)|V^l]$. Therefore, one can march forward through time, calculating the (closed-form) covariance of $\bar{\Xi}^l_t$ across the grid points at each time step. For both cases, the initial covariance $Cov[\bar{\Xi}_0(u, dy_i), \bar{\Xi}_0(u', dy_j)|V^l] = 0$ for all $y_i, y_j \in \mathcal{R}$.
  - For each V^l and any function f, a numerical approximation can be made:

$$\left\langle f, \mu^{N,l}_t \right\rangle \overset{d}{\approx} \sum_{i,u} h_{i,u} \big(\bar{\mu}^l_t(u, dy_i) + \frac{1}{\sqrt{N}} \bar{\Xi}^l_t(u, dy_i) \big),$$

    where $h_{i,u}$ is a quadrature rule. Since $\bar{\mu}^l$ is conditionally deterministic and $\bar{\Xi}^l$ is conditionally Gaussian with covariance known in closed-form, $\langle f, \mu^{N,l}_t \rangle$ is conditionally $\mathcal{N}(m^l, \Sigma^l)$, where m^l and Σ^l can again be easily calculated in closed-form.
- Collecting the conditional distributions for each V^l, and letting $\phi(\cdot, m, \Sigma)$ be the density of Gaussian random variable with mean m and covariance Σ, the density of $\left\langle f, \mu^N_t \right\rangle$ can be directly computed:

$$p_\theta^{\left\langle f, \mu^N_t \right\rangle}(z) \overset{d}{\approx} \frac{1}{L} \sum_{l=1}^{L} \phi(z, m^l, \Sigma^l).$$

Alternatively, $\bar{\Xi}_t$ could simply be simulated instead of finding its closed-form covariance conditional on each V^l. In this latter approach, one would simply simulate $\bar{\Xi}_t$ for each V^l. Such simulation is straightforward since $\bar{\Xi}_t$ is conditionally Gaussian given V.

As can be seen, simulation of the approximation (12) is simple in principle. However, in many practical applications, the loan-level feature space $\mathcal{Y}$ will be high-dimensional. The number of grid points in $\mathcal{R}$ will grow exponentially with the dimension d_Y of the loan-level feature space. This is commonly referred to as the curse of dimensionality. The next section presents a transformation which reduces the dimension d_Y of the LLN and CLT to a constant dimension d_W, *no matter how large the original dimension d_Y is*, paving the way for tractable computations. We believe the subclass of models covered by this transformation is not statistically outperformed by more complicated models which the transformation does not apply to. However, in general, one can still compute (12) for high-dimensions (without any transformation) via sparse non-uniform grids. We describe and numerically implement this latter approach in Section 6.6, which also strongly outperforms brute-force simulation of the actual pool.

3.3. Efficient Low-dimensional Approximation.

A potential drawback of the law of large numbers in equation (10) is that if the loan-level feature space $\mathcal{Y}$ is very high-dimensional, the law of large numbers will be high-dimensional and computations using traditional *uniform grids* can become expensive due to the curse of dimensionality. Fortunately, for a subclass of models, one can perform a transformation of the LLN in equation (10) which converts the law of large numbers into a low-dimensional problem. Alternatively, one can still compute the high-dimensional law of large numbers for the full class of models via a sparse non-uniform grid scheme, which we demonstrate in Section 6.6.

Suppose that there is a function g_θ such that $h_\theta(u, u', y, v, H) = g_\theta(u, u', f(y), v, H)$ where $f : \mathcal{Y} \mapsto \mathbb{R}^{d_W}$. Then, h_θ is invariant under the coordinate transformation $w = f(y)$ and one can reduce the high-dimensional equation for $\bar{\mu}_t(u, dy)$ to the low-dimensional equation:

$$\bar{\mu}_t(u, dw) \quad = \quad \sum_{u' \in \mathcal{U}} g_\theta(u, u', w, V_{t'<t}, \bar{H}_{t'<t}) \bar{\mu}_{t-1}(u', dw), \qquad w \in \mathbb{R}^{d_W}. \tag{13}$$

For instance, a function h_θ which satisfies this requirement and reduces to the low-dimensional representation in equation (13) is any generalized linear model (GLM), such as logistic regression. As is described in our empirical work [31], we find no increased predictive capability from more complicated or more nonlinear models for h_θ (such as a neural network). In fact, such increased complexity may cause overfitting. Moreover, we emphasize that the transformation allows arbitrary complexity with respect to the input factors y and v; the sole restriction is that v can only interact with y through the function $f(y)$. For instance, one can always make a GLM arbitrarily nonlinear in the input space y by adding features which are nonlinear functions of the initial set of features. A common choice is to use basis functions; a very simple example might be polynomials while a more complicated choice might be wavelets. This can greatly increase the dimension d_Y. However, such an expansion of the feature space does not increase the dimension d_W of the low-dimensional LLN (13) and its computational cost remains the same. In fact, *the dimension d_W does not depend on how large the original dimension d_Y is.*

The initial distribution $\bar{\mu}_0(u, dw)$ can be easily calibrated directly from the data available for the pool of loans. Simply transform the features $Y = (Y^1, \dots, Y^N)$ to $W = (W^1, \dots, W^N)$ via the map f and then calculate the empirical distribution of W. Due to its broad applicability and computational tractability, the low-dimensional LLN (13) is of great practical use.

Similarly, one can use the exact same transformation to arrive at a low-dimensional CLT, which has similar computational tractability to the low-dimensional LLN:

$$\begin{aligned}\bar{\Xi}_t(u,dw) &= \sum_{u'\in\mathcal{U}} g_\theta(u,u',w,V_{t'<t},\bar{H}_{t'<t})\bar{\Xi}_{t-1}(u',dw)\\ &+ \sum_{u'\in\mathcal{U}} \big(\frac{\partial}{\partial H} g_\theta(u,u',w,V_{t'<t},\bar{H}_{t'<t})\cdot \bar{E}_{t'<t}\big)\bar{\mu}_{t-1}(u',dw) + \bar{\mathcal{M}}_t(u,dw), \quad w\in\mathbb{R}^{d_W}. \end{aligned}\tag{14}$$

Given V, $\bar{\mathcal{M}}(u, dw)$ is a conditionally Gaussian process with zero mean and covariance:

$$\operatorname{Cov}\Big[\bar{\mathcal{M}}_t(u_1,dw),\bar{\mathcal{M}}_t(u_2,dw)\Big] = -\sum_{u'\in\mathcal{U}} g_\theta(u_1,u',w,V_{t'<t},\bar{H}_{t'<t})g_\theta(u_2,u',w,V_{t'<t},\bar{H}_{t'<t})\bar{\mu}_{t-1}(u',dw),$$
$$\operatorname{Var}\Big[\bar{\mathcal{M}}_t(u,dw)\Big] = \sum_{u'\in\mathcal{U}} g_\theta(u,u',w,V_{t'<t},\bar{H}_{t'<t})(1-g_\theta(u,u',w,V_{t'<t},\bar{H}_{t'<t})\bar{\mu}_{t-1}(u',dw),$$

where $u_1 \neq u_2$. Some examples of the low-dimensional approximation are presented below.

Example 3.5 (One-dimensional Efficient Approximation). *Let $\mathcal{U}$ be {Outstanding, Prepaid }. Let Y^n be a vector containing loan-level features for the n-th mortgage and V_t be a vector of systematic factors (such as the national unemployment rate and mortgage rate) at time t. For this example, h_θ does not depend upon the mean field term H^N. A logistic function is chosen for h_θ. Let the parameter vector $\theta = (\theta^Y_{prepay}, \theta^X_{prepay})$ where $X_t = (1, V_t)$. Then, the dynamics for the n-th mortgage are*

$$\begin{aligned} h_\theta(\textit{prepaid}, \textit{outstanding}, Y^n, V_{t-1}) &= \frac{e^{\theta^X_{prepay}\cdot X_{t-1}+\theta^Y_{prepay}\cdot Y^n}}{1+e^{\theta^X_{prepay}\cdot X_{t-1}+\theta^Y_{prepay}\cdot Y^n}},\\ h_\theta(\textit{outstanding}, \textit{outstanding}, Y^n, V_{t-1}) &= \frac{1}{1+e^{\theta^X_{prepay}\cdot X_{t-1}+\theta^Y_{prepay}\cdot Y^n}}. \end{aligned}\tag{15}$$

Let $w = \theta^Y_{prepay} \cdot y$. Then, both the low-dimensional LLN (13) and the low-dimensional CLT (14) only have one spatial dimension.

Example 3.6 (Two-dimensional Efficient Approximation). *Let $\mathcal{U}$ be {Outstanding, Default, Prepaid }. Let Y^n be a vector containing loan-level features for the n-th mortgage and V_t be a vector of systematic factors at time t. Again, there is no dependence upon the mean field term H^N. We choose a logistic function for h_θ. Let the parameter vector $\theta = (\theta^Y_{default}, \theta^X_{default}, \theta^Y_{prepay}, \theta^X_{prepay})$ where $X_t = (1, V_t)$. Then, the dynamics for the n-th mortgage are*

$$\begin{aligned} h_\theta(\textit{default}, \textit{outstanding}, Y^n, V_{t-1}) &= \frac{e^{\theta^X_{default}\cdot X_{t-1}+\theta^Y_{default}\cdot Y^n}}{1+e^{\theta^X_{default}\cdot X_{t-1}+\theta^Y_{default}\cdot Y^n}+e^{\theta^X_{prepay}\cdot X_{t-1}+\theta^Y_{prepay}\cdot Y^n}},\\ h_\theta(\textit{prepaid}, \textit{outstanding}, Y^n, V_{t-1}) &= \frac{e^{\theta^X_{prepay}\cdot X_{t-1}+\theta^Y_{prepay}\cdot Y^n}}{1+e^{\theta^X_{default}\cdot X_{t-1}+\theta^Y_{default}\cdot Y^n}+e^{\theta^X_{prepay}\cdot X_{t-1}+\theta^Y_{prepay}\cdot Y^n}},\\ h_\theta(\textit{outstanding}, \textit{outstanding}, Y^n, V_{t-1}) &= \frac{1}{1+e^{\theta^X_{default}\cdot X_{t-1}+\theta^Y_{default}\cdot Y^n}+e^{\theta^X_{prepay}\cdot X_{t-1}+\theta^Y_{prepay}\cdot Y^n}}. \end{aligned}\tag{16}$$

Let $w = (w_1, w_2) = (\theta^Y_{default} \cdot y, \theta^Y_{prepay} \cdot y)$. The low-dimensional LLN (13) and the low-dimensional CLT (14) then have two spatial dimensions.

We emphasize that the low-dimensional approximations for these two examples only have one and two spatial dimensions, respectively, *no matter how large the number of loan-level features d_Y is.* Therefore, the computational expense of the approximation remains constant even as d_Y grows very large.

Although both cases provide much greater computational efficiency than brute-force simulation of the pool, the one-dimensional approximation will always be faster than the two-dimensional approximation since it has one less dimension. For subprime mortgages, both prepayment and default must be modeled, so the two-dimensional approximation is necessary. However, the one-dimensional approximation is applicable for a wide range of cases which we list below:

- Pools of loans only subject to default risk. (Although mortgages don't fall into this category, many other types of loans do. An example is corporate bonds.)
- High quality mortgage pools with very low default rates. For such mortgage pools, it may be reasonable to only model prepayment and ignore defaults.
- Losses from defaults in agency mortgage pools are insured by the government-sponsored enterprises (GSEs) Freddie Mac and Fannie Mae. This means that a default is no different than a prepayment in an agency mortgage pool; both are just the early arrivals of cashflows from a mortgage. Therefore, prepayments and defaults can be modeled as a single type of event in agency mortgage pools.

The approximation using the low-dimensional LLN and CLT is now very computationally tractable for simulating the loss and prepayment levels in a pool. The computational performance of the low-dimensional LLN and CLT can be further enhanced by the use of a non-uniform grid for discretizing $\mathbb{R}^{d_W}$. Using a non-uniform grid, more points would be placed where $\bar{\mu}_0(u, dw)$ is large and less points would be placed where $\bar{\mu}_0(u, dw)$ is small. Section 3.4 provides an example of a particular non-uniform grid well-adapted to our problem. The non-uniform grid proposed in Section 3.4 is highly accurate even with only a small number of grid points. A sparse grid can also be used in order to further decrease computational time. Section 3.5 describes this approach. Using a sparse grid, simulation is performed at only a few points and then the solution is evaluated on a finer grid via interpolation. A final advantage of the approximation is its reusability; one set of simulations on a pre-chosen grid can be used to calculate the distribution for many different pools by re-weighting the simulated CLT and LLN (see Section 3.5). The low-dimensional approximation combined with these computational methods provides *an efficient Monte Carlo approximation* for mortgage pools.

In many typical applications, the efficient Monte Carlo approximation will be several orders of magnitude faster than brute-force Monte Carlo simulation of the original pool with a similar level of accuracy. In fact, just the LLN (without the second-order correction provided by the CLT) is sufficiently accurate for many typically-sized pools of interest. Section 6 studies the numerical performance of the efficient Monte Carlo approximation using actual mortgage data.

3.4. Non-uniform Grids.

To increase computational efficiency for the low-dimensional LLN and CLT, we recommend non-uniform grids. In a non-uniform grid, more points would be placed where $\bar{\mu}_0(u, dw)$ is large and less points would be placed where $\bar{\mu}_0(u, dw)$ is small. In the case where $h_\theta(y) = g_\theta(w)$ is a logistic function, we propose a non-uniform grid for $\mathbb{R}^{d_W}$ below:

- Divide $\mathbb{R}^{\mathbb{W}}$ into K boxes, each with equal mass $\int_{\text{box } k} \bar{\nu}(dw) = 1/K$ where we take $\bar{\nu}(dw) = \bar{\mu}_0(\{\text{outstanding}\}, dw)$. In one dimension, this can be easily done by finding the quantiles of the distribution $\bar{\nu}$. It is assumed that $\int_{\mathbb{R}^{d_W}} \bar{\mu}_0(\{outstanding\}, dw) = 1$.
- In the k-th box, choose the grid point $w_k = \log K \int_{\text{box } k} e^w \bar{\nu}(dw)$.
- Evaluate the solution $\bar{\mu}_t(w)$ at the grid points $w_1, \ldots, w_K$.

If g_θ is locally linear (at least within the k-th box) in e^w, the grid points y_k can make this scheme highly accurate. We demonstrate for one time-step to explain the choice of the points y_k. Define the function q_θ such that $q_\theta(u, e^w) = g_\theta(u, w)$. The exact mass within the k-th box at $t = 1$ is $\int_{\text{box } k} \bar{\mu}_1(u, dw) = \int_{\text{box } k} g_\theta(u, w)\bar{\nu}(dw) = \int_{\text{box } k} q_\theta(u, e^w)\bar{\nu}(dw)$ where we have suppressed the other arguments of g for notational convenience. If g_θ is approximately locally linear in e^w (i.e., q is approximately linear) in the k-th box, one has that

$$\int_{\text{box } k} \bar{\mu}_1(u, dw) = \int_{\text{box } k} q_\theta(u, e^w)\bar{\nu}(dw) \approx \frac{1}{K} q(u, K \int_{\text{box } k} e^w \bar{\nu}(dw))$$

$$= \frac{1}{K} q(u, w_k) = q(u, w_k) \int_{\text{box } k} \bar{\mu}_0(\{\text{outstanding}\}, dw).$$

Then, if g is close to locally linear in e^w, the choice of the grid point w_k will lead to a very accurate solution for the total mass in the k-th box. In the end, the quantity of interest is the total mass in each state u (i.e., what fraction of mortgages are still alive, what fraction have defaulted, and what fraction have prepaid), so this is highly useful. One can simply sum up the mass in each box to find the total mass in state u. Although this scheme has been specifically tailored to the case where h_θ is a logistic function, generalizations can easily be made to other function choices.

3.5. **Pre-computation for Financial Institutions.** Even for the risk analysis of smaller, individual mortgage-backed securities, the efficient Monte Carlo approximation can provide considerably faster computations. For instance, for a single MBS of $1,000$ mortgages, although the approximation is accurate, it does not offer as large computational cost savings as for very large pools. However, a typical financial institution will deal with thousands of mortgage-backed securities. As mentioned earlier, a mortgage trading desk at a major bank will on a daily basis analyze thousands of MBSs and hundreds of CMOs.

Assuming there is no mean field dependence in equation (2), one can pre-simulate the LLN and CLT at a set of grid points $\mathcal{R} \in \mathbb{R}^{d_W}$. This pre-simulation occurs only once. Then, one can find the distribution for the k-th pool by taking a weighted combination of the pre-simulated approximation $\bar{\mu}^N$ across the grid points $\mathcal{R}$, where the weights are chosen to match the k-th pool's loan-level feature distribution.

If the series of pools have sizes $N_1, \ldots, N_K$ with $N = N_1 + \cdots + N^K$, then the computational cost of the efficient Monte Carlo approximation compared with brute-force Monte Carlo simulation of the actual pool is N_g/N where $N_g = |\mathcal{R}|$ is the number of grid points. Furthermore, the method immediately yields the correlation between the different pools, which is essential for risk management purposes. The approach is summarized below:

- Pre-simulate the LLN $\bar{\mu}$ and CLT $\bar{\Xi}$ on the grid $\mathcal{R}$ with initial condition $\bar{\mu}_0(\{outstanding\}, dw_i) = 1$ and $\bar{\mu}_0(u, dw_i) = 0$ for $u \neq \{outstanding\}$.
- For each pool $1, \ldots, K$: Find the k-th pool's distribution in the w-space and approximate it at the grid points $\mathcal{R}$; let h_i be the fraction at the i-th grid point. Then, the k-th pool's distribution is

$$\mu_t^{N_k}(u, dw_i) = h_i \bar{\mu}_t(u, dw_i) + \frac{\sqrt{h_i}}{\sqrt{N_k}} \bar{\Xi}(u, dw_i). \tag{17}$$

The method can be further improved by taking a sparse grid $\mathcal{R}$ in order to reduce the number of calculations and then, after the pre-simulation, interpolating on a finer grid. Due to the smoothness of $\bar{\mu}^N$ for typical functions h_θ, only a few grid points are usually needed in order to get an accurate interpolated solution. Using this approach, the efficient Monte Carlo approximation can be highly useful even for small mortgage-backed securities as long as the financial institution is dealing with many such mortgage-backed securities in aggregate. The approach is implemented using actual mortgage data in Section 6.5.

4. Parameter Estimation

Before one analyzes loan pools using the efficient Monte Carlo approximation, the parameter θ specifying the model (2) must be fitted to historical data. The weak convergence results of Section 3.1 turn out to be useful for that purpose also. More specifically, we wish to estimate θ given observations of both $(U^1, \ldots, U^N)$ and V at each time $t = 1, \ldots, T$. The observations of U are generated by the true parameter $\theta_0 \in \Theta$. Collectively, the observations of the particles U up to time T are $\mathsf{D}_{T,N} = (Z_1^N, \ldots, Z_T^N)$ where $Z_t^N = (U_t^1, \ldots, U_t^N)$. The log-likelihood function for the data is

$$\begin{aligned}\mathcal{L}_{T,N}(\theta) = \log \mathbb{P}_\theta(\mathsf{D}_{T,N}) \propto \frac{1}{N} \log \mathbb{P}_\theta[Z_1^N, \ldots, Z_T^N | V] &= \frac{1}{N} \log \prod_{t=1}^{T} \mathbb{P}_\theta[Z_t^N | Z_0^N, \ldots, Z_{t-1}^N, V] \\ = \frac{1}{N} \log \prod_{t=1}^{T} \mathbb{P}_\theta[Z_t^N | Z_{t-1}^N, H_{t'<t}^N, V] &= \frac{1}{N} \log \prod_{t=1}^{T} \prod_{n=1}^{N} h_\theta(U_t^n, U_{t-1}^n, Y^n, V_{t'<t}, H_{t'<t}^N) \\ = \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} \log h_\theta(U_t^n, U_{t-1}^n, Y^n, V_{t'<t}, H_{t'<t}^N).\end{aligned} \tag{18}$$

Note that we have used the conditional independence of $U_t^1, \dots, U_t^N$ with respect to $\mathcal{F}_{t-1}$ on the second line in equation (18). Typically, one will also choose a separate model for the systematic factors V with its own parameters. These parameters can be estimated separately from θ using standard methods; note that the likelihood for θ depends only on the observed values of V and is independent of V's exact form or parameterization since V is an exogenous process.

A maximum likelihood estimator (MLE) $\theta_{T,N}$ for the parameter θ maximizes (18). Using the law of large numbers and central limit theorem for μ_t^N, it can be shown that $\theta_{T,N}$ converges in probability to θ_0 as $N \longrightarrow \infty$. In addition, a central limit theorem for the MLE can be proven. These two theoretical results can be useful in practice and can be shown to hold for finite samples (i.e., $T < \infty$). The first result suggests that the MLE will be accurate if N is large, even for a finite time. The second result can be used to find standard errors. This is important for determining the statistical significance of the estimators. A brute-force approach might be to estimate standard errors via bootstrapping; this requires estimating the parameters many times. For the large data sets typical in practice, a single estimation already takes considerable time. The central limit theorem provides a quick and tractable way to calculate these standard errors.

Assumption 4.1. *There exists a $T_0 < \infty$ such that for $T \geq T_0$, the map $\bar{\mu}_{t \leq T} : \Theta \mapsto B^{T+1}$ is almost surely one-to-one.*

Theorem 4.2. *Let Assumption 4.1 hold and $T > T_0$. In addition, suppose Assumption 3.1 holds and h_θ is continuously differentiable. Then, the estimator $\theta_{T,N}$ converges in probability to θ_0 as $N \longrightarrow \infty$.*

Assumption 4.1 is actually very mild and can be shown to be satisfied by many models in practice. It is equivalent to saying that the function h_θ is identifiable from the limiting data. For instance, one can easily see that a logistic function (which is the model we implement in the numerical portion of the paper) satisfies Assumption 4.1. To demonstrate concretely, consider the multinomial logistic function in Example 3.6. Suppose one observes h_θ at points $y = y_1, \dots, y_M$ and times $t = 1, \dots, T_0$. For each y_m and time t, one has the system of linear equations:

$$\begin{aligned} b_t^m &= \theta^X_{\text{default}} \cdot V_{t-1} + \theta^Y_{\text{default}} \cdot y_m, \\ c_t^m &= \theta^X_{\text{prepay}} \cdot V_{t-1} + \theta^Y_{\text{prepay}} \cdot y_m, \end{aligned} \tag{19}$$

where b_t^m and c_t^m are some constants. Assume that V has a density. Note that since $\bar{\mu}$ is observed everywhere in $\mathcal{Y}$, one can choose any set of points $\{y_m\}_{m=1}^M$ for equation (19). Choose points $\{y_m\}_{m=1}^M$ such that the corresponding coefficient matrix is full rank and $M = d_Y$. Then, Assumption 4.1 is clearly satisfied for $T_0 = d_V + 1$ if for each y_m either $\bar{\mu}_0(\{\text{outstanding}\}, y_m) > 0$ or $\bar{\mu}_0(\{\text{outstanding}\}, dy)$ admits a positive density with support on $(y_m - \epsilon, y_m + \epsilon)$ for $\epsilon > 0$. Typically, Assumption 4.1 will hold for a model h_θ if the limiting data is observed at more points than h_θ has degrees of freedom.

Theorem 4.3. *Let Assumption 4.1 hold with $T > T_0$ and let h_θ be twice continuously differentiable. In addition, suppose that Assumptions 3.1 and 3.3 hold. Define $\mu^N_{t,t-1} = \frac{1}{N}\sum_{n=1}^N \delta_{(U_t^n, U_{t-1}^n, Y^n)}$ and $\Xi^N_{t,t-1} = \sqrt{N}(\mu^N_{t,t-1} - \bar{\mu}_{t,t-1})$ where $\mu^N_{t,t-1} \Rightarrow \bar{\mu}_{t,t-1}$ and $\Xi^N_{t,t-1} \Rightarrow \bar{\Xi}_{t,t-1}$. As $N \to \infty$, the quantity $\sqrt{N}(\theta_{T,N} - \theta_0)$ converges in distribution to*

$$-\left(\frac{\partial^2}{\partial \theta^2}\mathcal{L}_{T,\infty}(\theta_0)\right)^{-1} \sum_{t=1}^T \left[\left\langle \frac{\partial \log h_{\theta_0}}{\partial \theta}(\cdot, \bar{H}_{t'<t}), \bar{\Xi}_{t,t-1} \right\rangle + \left\langle \frac{\partial^2 \log h_{\theta_0}}{\partial \theta \partial H}(\cdot, \bar{H}_{t'<t}) \cdot \bar{E}_{t'<t}, \bar{\mu}_{t,t-1} \right\rangle \right].$$

5. Mortgage Data

We will use actual mortgage data to test the performance of the efficient Monte Carlo approximation. Our data is comprised of two loan-level data sets. The first is a subprime data set during the time period 1995 – 2013, obtained from the Trust Company of the West (TCW). In total, there are over 10 million subprime mortgages in the data set. The second data set, obtained from Freddie Mac, contains agency mortgages over the time period 1999 – 2014 and consists of 16 million mortgages. These data sets will be used to statistically estimate the model (2) and then numerically test the performance of the efficient Monte Carlo approximation on real mortgage pools drawn from the data set. Detailed statistical analysis of these data sets can be found in our other paper, [31]. The data sets do not overlap since the first is non-agency

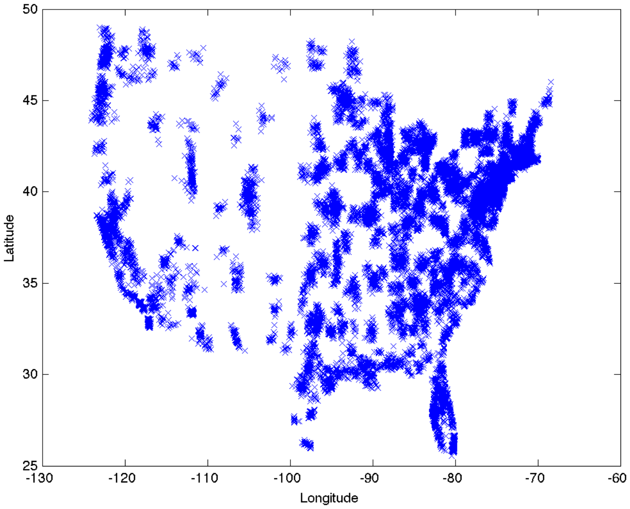

FIGURE 1. Map showing geographic locations of mortgages in subprime data set.

(specifically, subprime) and the second is agency. A summary of the two data sets is given below and a comparison is also provided in Table 4 in Appendix B.

5.1. **Subprime Mortgage Data.** There are over 10 million mortgages in the subprime data set. The mortgages are spread across the entire United States, covering over $36,000$ different zip codes (Figure 1 displays the geographic locations of the mortgages in the data set). The data includes zip code, FICO score, LTV, initial interest rate, initial balance, type of mortgage, deal ID, and time of origination. Default, foreclosure, modification, real estate owned (REO) and prepayment events, if they occur, are also recorded in the data set. Default is defined in this data set as when the property is sold to a third-party under financial distress. Such an event is distinct from a foreclosure. Often, the bank will foreclose on a property as a threat in order to force the mortgage holder to become current on their payments. Therefore, a foreclosure may not necessarily end in a default if the mortgage holder is able to become current on their payments. REO properties are foreclosed properties which failed to be sold at auction and return to the bank. The bank will then seek to sell these properties to third parties via a realtor. Default events as defined for this data set include REO properties sold to third parties, foreclosed properties sold at auction to third parties, and short sales. The servicer of the mortgage will typically keep advancing principle and interest to the MBS investor until the mortgage is finally sold to a third party, even if the property is not current on payments, foreclosed, or REO. Therefore, the definition of default used in this data set is the one pertinent to the MBS investor.

5.2. **Agency Mortgage Data.** In total, there are over 16 million agency mortgages in the data set. The data includes FICO score, first time homebuyer indicator, type of mortgage, maturity date, number of units, occupancy status, combined loan-to-value (CLTV), original debt-to-income (DTI) ratio, loan-to-value (LTV) ratio, initial interest rate, prepayment penalty mortgage (PPM) indicator, loan purpose (purchase, cash-out refinance, or no cash-out refinance), original loan term, and number of borrowers. In addition, there is monthly performance data for each mortgage. This monthly performance data includes the current loan delinquency status (current, 30-59 days delinquent, 60-89 days delinquent, 90-119 days delinquent, 120-149 days delinquent, 150-179 days delinquent, and 180+ days delinquent), modification indicator, and current interest rate for each month of the mortgage until the present time (or until a termination event when the mortgage leaves the data set). If the mortgage is terminated from the data set, the time and reason is also recorded. Terminations events include: prepayment, third party sale prior to 180 days delinquent, short sale or short payoff prior to 180 days delinquent, deed-in-lieu of foreclosure prior to 180 days delinquent, repurchases prior to 180 days delinquent, 180 days delinquent, and REO acquisition prior to 180 days delinquent. The zip codes have been partially anonymized in this data set, but the metropolitan statistical area (MSA) is reported for each mortgage. There are roughly 430 different MSAs reported in the data set.

6. Computational Performance of the Efficient Monte Carlo Approximation

We compare the accuracy and computational cost of the efficient Monte Carlo approximation with brute-force Monte Carlo simulation using actual mortgage data. The efficient Monte Carlo approximation has very high accuracy even for small pools in the hundreds of mortgages. Mortgage-backed securities can range from a few thousand mortgages to hundreds of thousands of mortgages. Banks and other financial institutions often have credit exposure to hundreds of thousands or millions of mortgages. GSEs such as Fannie Mae and Freddie Mac have credit exposure to tens of millions of mortgages. The computational cost of the efficient Monte Carlo approximation is typically orders of magnitude lower than brute-force Monte Carlo simulation of the actual pool.

6.1. **Numerical Performance of LLN and CLT.** First, we demonstrate the accuracy of the law of large numbers by comparing the LLN distribution with the actual distribution for the pool. The actual (or "true") distribution is found via brute-force Monte Carlo simulation of the actual pool. The pool is drawn at random from the subprime data set, allowing us to assess the performance of the LLN with actual mortgage data.

A multinomial logistic regression model is used for h_θ and there is no dependence on the mean field term H^N (e.g., Example 3.6).[10] The loan-level factors used for both default and prepayment are FICO score, LTV, initial balance of the mortgage, and initial interest rate for the mortgage. The national unemployment rate is used as the systematic factor for default. Both the national unemployment rate and national mortgage rate are used as systematic factors for prepayments. We perform $25,000$ Monte Carlo simulations for both the brute-force simulation of the actual pool as well as the simulation of the LLN. We simulate the pool for a one-year time horizon, with monthly discretization. The mortgage rate and unemployment rate are simulated as discrete-random walks with standard deviations fitted to their historical values prior to January 1, 2012. The parameters θ are also fitted using the entire subprime data set prior to January 1, 2012. The parameter fits are reported in Table 1. The variables have been normalized; so, the magnitudes of the parameters are directly comparable with each other. The numerical studies in Sections 6.1 and 6.2 are performed using the parameter fits from Table 1. We model both default and prepayment, so $d_W = 2$ for the efficient Monte Carlo approximation.

Factor	Default	Prepayment
Constant	-5.906	-4.363
National unemployment rate	.7593	-.9782
National mortgage rate	NA	-.1517
LTV ratio	.3072	-.0657
Initial Balance	.1001	.0656
Initial interest rate	.2859	.0503
FICO score	.0706	-.0402

Table 1. Parameter fits for default and prepayment model.

Figure 2 compares the actual distribution with the LLN distribution for the loss from default for pools of sizes $N = 5,000$, $10,000$, $25,000$ and $100,000$, respectively. The pools are drawn at random from the data set. The LLN is very accurate, especially in the right tail of the distribution. The right tail is essential for risk management purposes, such as calculating the value at risk (VaR). The LLN can be combined with the CLT to create a second-order accurate approximation. The approximation is accurate even for very small pools in the hundreds of mortgages. Figure 3 compares the approximate distribution (using both the LLN and CLT) with the actual distribution for pools with sizes $N = 500, 1000, 2500$, and 5000.

[10] Model parameters are fitted to the data set using stochastic gradient descent. We find in [31] that more complicated models for h_θ (including logistic functions with polynomial features and neural networks) do not outperform standard multinomial logistic regression.

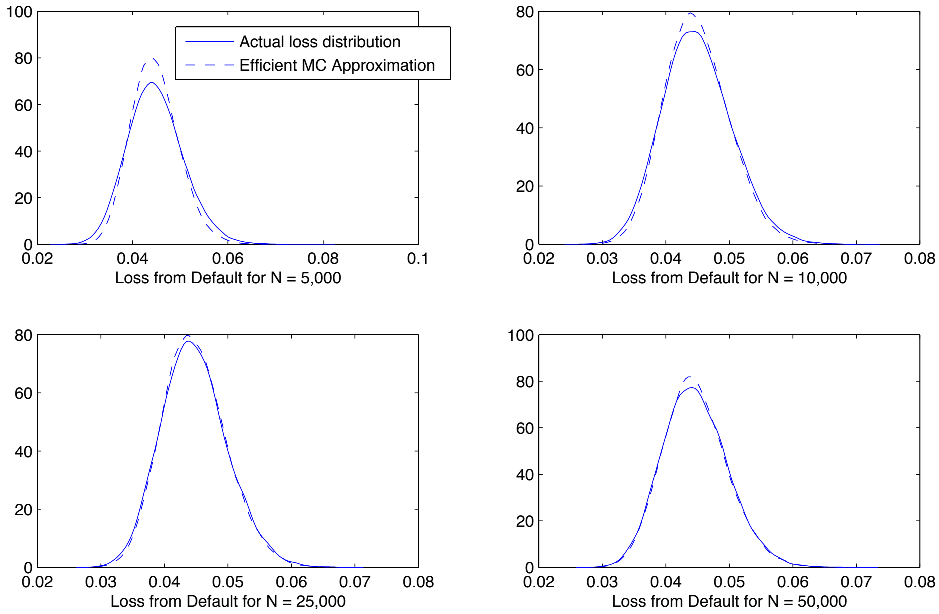

FIGURE 2. Comparison of actual loss distribution with LLN loss distribution (does not include CLT in approximation). Loss reported as fraction of pool which defaulted. The horizon is 12 months.

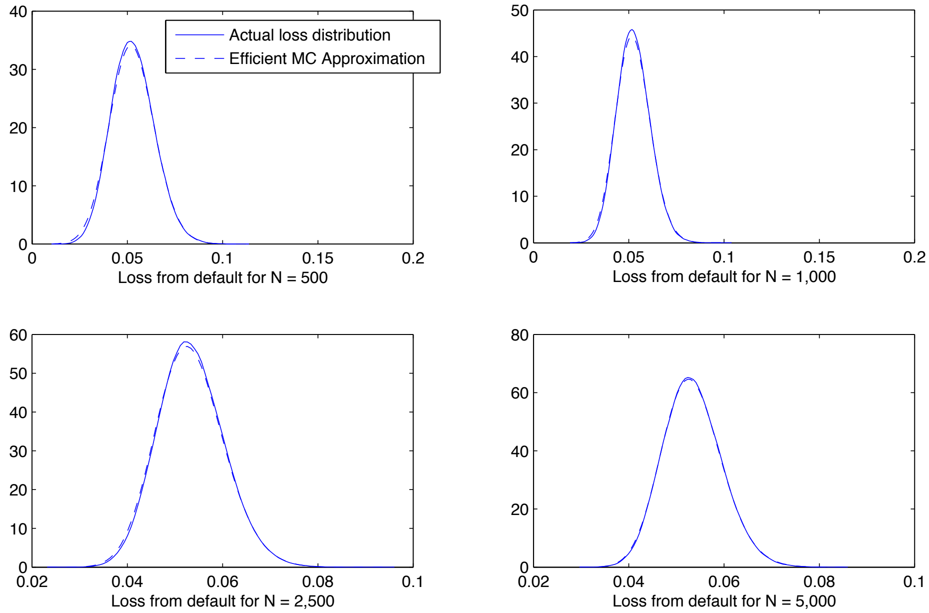

FIGURE 3. Comparison of actual distribution with approximate distribution (using both LLN and CLT). Loss reported as fraction of pool which defaulted. The horizon is 12 months.

Computational costs are reported in Table 2 for different sized pools N. In general, a rough approximation of the ratio of the computational costs is

$$\frac{\text{Cost of LLN}}{\text{Cost of Simulation of Actual Pool}} = \frac{N_g}{N}, \tag{20}$$

where N_g is the number of grid points needed for the numerical solution of the LLN and CLT equations across the space $\mathbb{R}^{d_W}$ and N is the number of mortgages in the actual pool. For instance, if one only needs 25 grid points, the ratio of computational costs is $\frac{25}{N}$. For a million mortgages, that leads to a reduction in computational time of well over 4 orders in magnitude. Computational times listed in Table 2 for brute-force simulation of the pool, simulation of the LLN, and simulation of the full approximation (LLN combined with the CLT) are for $d_W = 2$ (model includes both prepayment and default). These computational times are for a twelve-month time horizon.

N	Time for Brute-force Simulation	Time for LLN	Time for LLN and CLT
1,000	44.34	1.03	2.67
5,000	153.78	1.03	2.67
10,000	273.39	1.03	2.67
25,000	608.94	1.03	2.67
100,000	2,847.43	1.03	2.67
1,000,000	28,563.68	1.03	2.67

Table 2. Comparison of computational times (seconds) for efficient Monte Carlo approximation and brute-force Monte Carlo simulation of the pool.

6.2. **Numerical Performance across Actual Deals.** The subprime data set covers over 6,000 deals.[11] As mentioned earlier, for each mortgage, a deal ID is available. One can therefore reconstruct the actual pools for deals. We will perform some numerical tests to study the efficient Monte Carlo approximation's accuracy across the wide diversity of deals in the data universe.

Recall that we transformed the loan-level features y into the two-dimensional vector $w = (w_1, w_2) = (\theta^Y_{\text{default}} \cdot y, \theta^Y_{\text{prepay}} \cdot y)$. One can think of w_1 as the level of default risk. The distribution of a pool over the w_1 space indicates the distribution of risk for the pool. Figure 4 compares the distribution of risk for the deals in the subprime data set. One can see that some deals are very risky, while others are less risky.

Figure 5 compares the approximate distribution from the efficient Monte Carlo approximation with the true distribution (found via brute-force Monte Carlo approximation) at a time horizon of 12 months for ten actual deals in the data set. The deals are chosen at random and each contains between $5,000$ and $10,000$ mortgages. It is interesting that the default rate can vary considerably between deals as a consequence of the quality of the underlying mortgages, demonstrating how important it is for a model to consider the loan-level characteristics of the mortgages in the pool.

To further assess the accuracy of the efficient Monte Carlo approximation, a set of deals is selected at random from the data set and the efficient Monte Carlo approximation's 99 % value at risk (VaR) is compared with the true 99 % value at risk in Table 3. In total, we look at 185 deals and report the average error of just the LLN by itself as well as the average error for the full efficient Monte Carlo approximation (LLN and CLT combined). In addition, Figure 6 shows the distribution of the efficient Monte Carlo approximation's error across the set of deals. $50,000$ Monte Carlo simulations are performed and the time horizon is again twelve months.

[11] "Deal" refers to the securitization of a pool of mortgages into an MBS. An MBS' structure may vary widely: it could be anything from a pass-through to a collateralized mortgage obligation (CMO). A CMO has a tranched structure, with different payment rules for the different tranches.

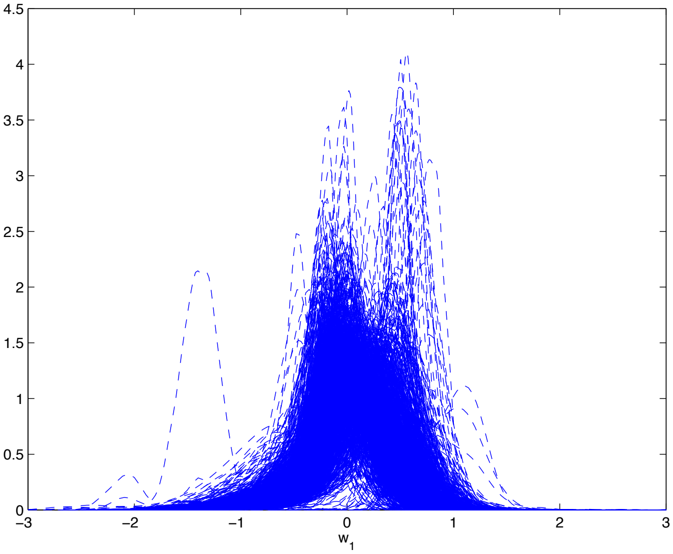

FIGURE 4. Comparison of risk for the different deals in the subprime data set.

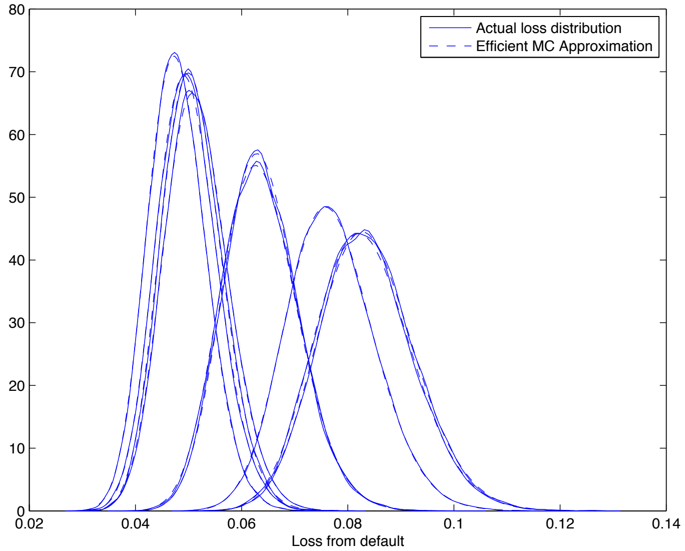

FIGURE 5. Comparison of actual distribution with approximate distribution for ten actual deals with $5,000 < N < 10,000$. Loss reported as fraction of pool which defaulted. The horizon is 12 months.

6.3. **Extremely high-dimensional loan-level feature space.** The spatial dimension d_W for the efficient Monte Carlo approximation does not depend on d_Y. Consequently, the large pool approximation's computational expense remains constant even as d_Y becomes very large. We demonstrate the tractability of the approximation for a high-dimensional loan-level feature space. In addition to the features previously included, we now include (for both prepayment and default) the mortgage type and the zip code. Each mortgage type

Size of Deal	Average Error, LLN	Average Error, LLN and CLT
$5,000 < N < 10,000$	2.25 %	0.22 %
$10,000 \leq N$	1.25 %	0.18 %

Table 3. Comparison of 99 % VaR from the efficient Monte Carlo approximation with the actual 99 % VaR (found via brute-force simulation of the pool).

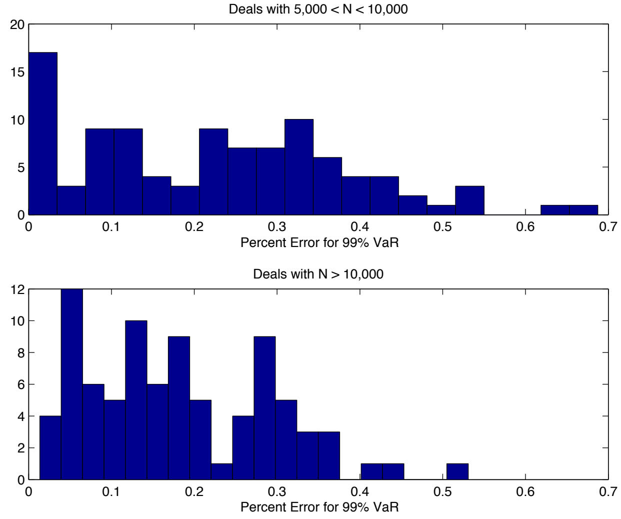

Figure 6. Distribution across deals of error for 99 % VaR from the efficient Monte Carlo approximation. The time horizon is 12 months.

and each zip code is encoded as a $\{0, 1\}$ element in the vector $y \in \mathcal{Y}$. Since there are over 38 mortgage types and the data set includes mortgages from over $36,000$ zip codes, $d_Y > 36,000$. For each zip code and mortgage type, there is a corresponding parameter that must be fitted.[12] Despite d_Y being huge, d_W is still equal to just 2 and the efficient Monte Carlo approximation's computational expense does not increase.

Figure 7 shows the efficient Monte Carlo approximation compared to the actual loss distribution for $N = 5,000$ as well as the distribution of the pool in the "risk space" w_1. The inclusion of the product types and zip codes has produced a different shape for the w_1-distribution (compared to Figure 4). This again highlights that the loan-level constitution of a portfolio can directly impact the risk of a pool as a whole, arguing in favor of a model framework that incorporates loan-level information.

6.4. **One-dimensional Efficient Monte Carlo Approximation.** So far, we have focused on the case where $d_W = 2$. In this case, one models both default and prepayment. However, as mentioned previously, many types of loans only have default risk and do not have prepayment risk. In addition, for very high quality mortgage pools, default risk is small and it may be reasonable to consider only prepayment risk. Finally, for agency mortgage pools, the GSEs insure against any default losses, so prepayment and default can be treated as the same event. To demonstrate the one-dimensional approach, we fit a model only including prepayment to the agency mortgage data.[13]

[12] In practice, to avoid overfitting, early stopping or an ℓ^2 penalty can be imposed on the zip code parameters. The penalty parameter and early stopping rules are chosen via cross-validation. Since our goal here is solely to demonstrate the effectiveness of the approach in high dimensions, we do not include such precautions against overfitting.

[13] The agency mortgage data set is particularly challenging to fit. It has over 500 million rows of data; therefore, one cannot read the data into memory all at once. We read chunks of the data one by one. In order to remove any bias from the ordering of the data, we randomly shuffle the data from the original files into a new set of files. Then, we use stochastic gradient descent to fit the model.

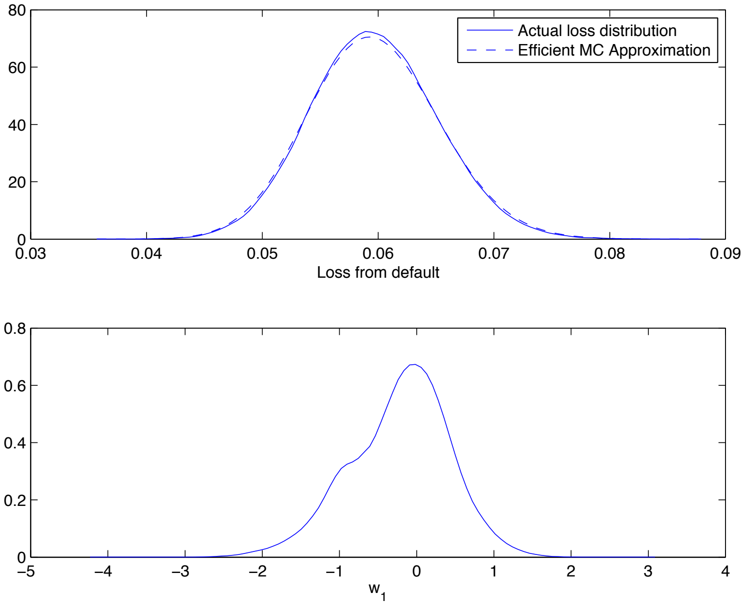

Figure 7. Top plot compares the actual loss distribution with the efficient Monte Carlo approximation for $N = 5,000$. The time horizon is 12 months. Bottom plot shows the distribution of the pool in the risk space.

For prepayment, we consider a number of loan-level features: FICO score, whether a first time homebuyer, the number of units, occupancy status (owner occupied, investment property, or second home), combined loan-to-value, loan-to-value, initial interest rate for the mortgage, debt-to-income ratio, whether there is a prepayment penalty in the mortgage contract, property type (condo, leasehold, PUD, manufactured housing, $1-4$ fee simple, Co-op), loan purpose (purchase, cash-out refinance, or no cash-out refinance), and number of borrowers. These features amount to 22 dimensions in the feature space. We also include the metropolitan statistical area (MSA); there are 430 metropolitan statistical areas in the data set. Therefore, in total, $d_Y = 452$. We emphasize again that, even though $d_Y = 452$, d_W only equals 1. The systematic factors are the national unemployment rate and the national mortgage rate. Table 5 in Appendix B reports the parameter fits for the first 22 loan-level features as well as the systematic variables; the parameter fits for the geographic locations are not included due to space constraints (they are available upon request). Figure 8 compares the efficient Monte Carlo approximation with the actual prepayment distribution for a randomly drawn pool of agency mortgages for a time horizon of 12 months.

6.5. **Precomputation for Large Financial Institutions.** In Section 3.5, we proposed to pre-simulate $\bar{\mu}^N$ on a pre-chosen grid and then use this one set of simulations on the single grid in order to find the distribution for many different mortgage-backed securities. Using the methodology proposed in Section 3.5, the efficient Monte Carlo approximation can provide great computational cost savings even for very small mortgage-backed securities, as long as a financial institution is dealing with many of these small mortgage-backed securities in aggregate.

We now implement this approach using the parameter fits from Section 6.4. 400 pools, each of size $N = 2,500$, are drawn at random from the agency mortgage data set. Each of these pools is simulated $50,000$ times using brute-force Monte Carlo simulation. Using the efficient Monte Carlo approximation, we also pre-simulate on a pre-chosen grid (as described in Section 6.4). $50,000$ Monte Carlo paths are also used for the efficient Monte Carlo simulation on this grid. The pre-chosen grid only has 20 grid points, placed at uniform intervals. The distribution of $\bar{\mu}^N$ is smooth in w, so it is easy to numerically interpolate from the sparse grid points to get a finer solution. We use piecewise cubic spline interpolation. This saves

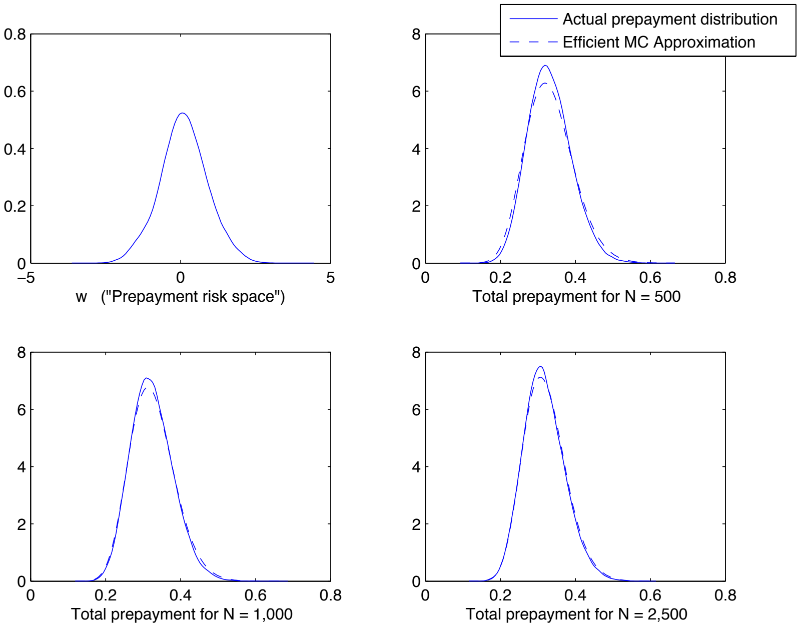

FIGURE 8. Top right plot and bottom plots compare the actual prepayment distribution with the efficient Monte Carlo approximation for $N = 500$, $N = 1,000$, and $N = 2,500$. The total prepayment is reported as the fraction of the pool which prepays. The time horizon is 12 months. The top left plot shows the distribution of the pool in the "prepayment risk space".

computational time since it allows one to simulate on a very sparse grid but still ultimately achieve a very accurate solution on a fine grid. Figure 9 is a histogram of the percent error over the 400 pools for the 99 % VaR from the efficient Monte Carlo approximation using pre-simulation. The time horizon is 12 months. The efficient Monte Carlo approximation is very accurate; the average percent error across the 400 pools for the 99 % VaR from the approximation is only 0.25%. Brute-force simulation of the 400 pools takes $36,905.86$ seconds while the efficient Monte Carlo approximation of the 400 pools takes 4.72 seconds. The approximation provides cost savings of nearly 4 orders of magnitude versus brute-force simulation.

6.6. **Numerical Evaluation of Large Pool Approximation without Low-dimensional Transformation.** The previous numerical solutions of the law of large numbers and central limit theorem rely upon a low-dimensional transformation, which requires a restriction on the interaction between the factors Y^n and X. We believe that this restriction results in no loss of statistical performance in practice. However, for completeness, we now present a method for the computation of the law of large numbers and central limit theorem for the full class of models. First, we cluster the pool into K clusters using k-means clustering. The K centroids are chosen as the grid points. This is a sparse non-uniform grid in a high-dimensional space.

We implement this approach for a pool of size $N = 50,000$ and $K = 50$. $25,000$ Monte Carlo trials are performed for both the brute-force simulation and the law of large numbers. The function h_θ is again a multinomial logistic regression model, with loan-level feature space $\mathcal{Y} = (\text{FICO score}, \text{LTV ratio}, \text{original balance}, \text{initial interest rate})$ and systematic factors X including the national unemployment rate and the national mortgage rate. For the default inputs, we include all of the Y^n factors, the national unemployment rate, and all first order polynomials of Y^n and the national unemployment rate (e.g., FICO score multiplied by the national unemployment rate). For the prepayment inputs, we include both unemployment rate and national mortgage rate, all of the Y^n factors, and the nonlinear factor $\max(\text{initial interest rate} - \text{national mortgage rate}, 0)$. Figure 10 shows a comparison between the Monte Carlo approximation with the sparse non-uniform grid and brute force simulation of the pool. The Monte Carlo approximation takes 22 seconds while brute force simulation takes 3908 seconds.

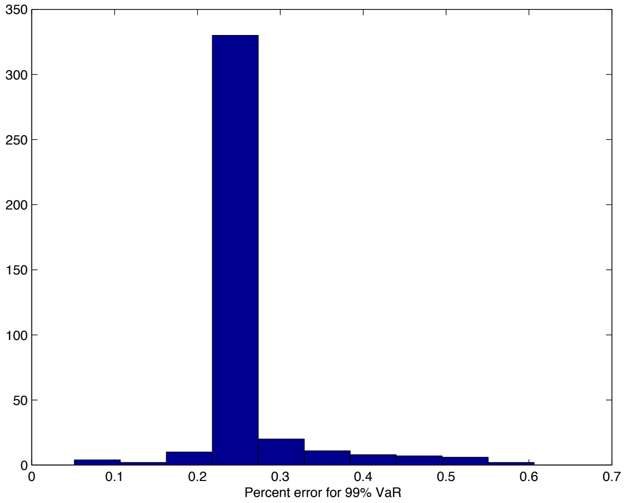

FIGURE 9. Distribution across deals of error for 99 % VaR from the efficient Monte Carlo approximation using pre-simulation. The time horizon is 12 months.

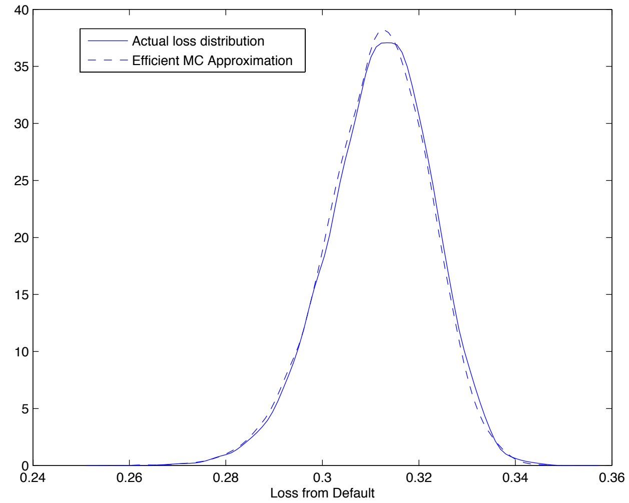

FIGURE 10. Comparison of actual loss distribution with the law of large numbers using a sparse non-uniform grid for $N = 50,000$ at a time horizon of 12 months.

7. Conclusion

This paper develops an efficient Monte Carlo approximation for a general class of loan-by-loan default and prepayment models for pools of mortgages. The approximation is based upon a law of large numbers and a central limit theorem for the pool default and prepayment processes. We extensively test our approach on actual mortgage data. The approximation is highly accurate even for very small pools (as little as 500 loans). In practice, pools commonly range from a few thousand to millions of loans. Brute-force simulation for large pools can be computationally expensive; our approximation can save several orders of magnitude in computational time. The efficient Monte Carlo approximation accounts for the full loan-level dynamics, taking advantage of the detailed loan-level information typically available (such as FICO score, loan-to-value ratio, initial interest rate, and type of mortgage). A key feature of our approximation is that its computational cost is constant no matter the dimension of the loan-level data. This feature is desirable since loan-level data can be high-dimensional. Although we have largely focused on mortgage pools, our approach is broadly applicable to other types of consumer and commercial loans.

Appendix A. Proofs

Proof of Theorem 3.2. First, note that the sequence μ_t^N is trivially tight in the space $\mathcal{P}(\mathcal{U} \times \mathbb{R}^{d_Y})$. Since the Y^n's live on a compact subset of $\mathbb{R}^{d_Y}$, this implies that μ_t^N lies in a compact subset of $\mathcal{P}(\mathcal{U} \times \mathbb{R}^{d_Y})$.

For each $u \in \mathcal{U}$ and any $\phi \in C_b(\mathcal{U}, \mathbb{R}^{d_Y})$, we have that

$$\left\langle \phi, \mu_t^N(u, dy) \right\rangle = \sum_{u' \in \mathcal{U}} \left\langle \phi(y) h(u, u', y, V_{t'<t}, H^N_{t'<t}), \mu^N_{t-1}(u', dy) \right\rangle + \mathcal{M}_t^{1,N}(u, dy). \tag{21}$$

where $\mathcal{M}^{1,N}$ is the martingale

$$\mathcal{M}^{1,N} = \frac{1}{N} \sum_{n=1}^{N} \phi(Y^n)(\mathbf{1}_{U_t^n = u} - h(u, U^n_{t-1}, Y^n, V_{t'<t}, H^N_{t'<t})). \tag{22}$$

Clearly, the variance of $\left\langle \phi, \mathcal{M}^{1,N} \right\rangle$ converges to zero. Therefore, assuming $\{\mu^N_{t'}\}_{t'<t}$ converges in probability to $\{\bar{\mu}_{t'}\}_{t'<t}$ for each V, and using the fact that h is bounded and continuous, we have that $\left\langle \phi, \mu_t^N \right\rangle$ converges to $\langle \phi, \bar{\mu}_t \rangle$, where any accumulation point $\bar{\mu}_t$ satisfies

$$\langle \phi, \bar{\mu}_t(u, dy) \rangle = \sum_{u' \in \mathcal{U}} \left\langle \phi(y) h(u, u', y, V_{t'<t}, \bar{H}_{t'<t}), \bar{\mu}_{t-1}(u', dy) \right\rangle, \quad \text{a.s.} \tag{23}$$

Note that equation (3.2) satisfies equation (23) for every ϕ. It is easy to see that $\bar{\mu}_t$ must be unique. (Suppose there are different solutions $\bar{\mu}_t^1$ and $\bar{\mu}_t^2$, and let $\nu = \bar{\mu}_t^1 - \bar{\mu}_t^2$. Then, note that $\langle \phi, \nu \rangle = 0$ for every ϕ.)

Let $\mathcal{S} : \mathcal{P}(\mathcal{U} \times \mathbb{R}^{d_Y}) \longrightarrow \mathbb{R}$ be the collection of functions of the form $\Phi(\mu) = \phi_1(\langle f_1, \mu \rangle, \ldots, \langle f_M, \mu \rangle)$ for $\phi_1 \in C^\infty(\mathbb{R}^M), f_m \in C^\infty(\mathcal{U} \times \mathbb{R}^{d_Y})$. Then, $\mathcal{S}$ separates $\mathcal{P}(\mathcal{U} \times \mathbb{R}^{d_Y})$ (see [32]). $\mathbb{E}[\Phi(\mu_t^N)|V] \to \mathbb{E}[\Phi(\bar{\mu}_t)|V]$ follows from equation (23), the continuous mapping theorem, and Y^n's distribution having compact support. We then have that $\mu_t^N \xrightarrow{d} \bar{\mu}_t$ and, since $\bar{\mu}_t$ is deterministic given V, μ_t^N also converges in probability to $\bar{\mu}_t$ for each V. This also means that $\{\mu^N_{t'}\}_{t'<t+1} \xrightarrow{p} \{\bar{\mu}_{t'}\}_{t'<t+1}$ for each V. Note that in Assumption 3.1, μ_0^N weakly converges to $\bar{\mu}_0$, where $\bar{\mu}_0$ is deterministic. By Assumption 3.1 and induction, it follows that μ^N converges in probability to $\bar{\mu}$ for each V. Since the convergence in probability holds for every V, we certainly have that μ^N weakly converges to $\bar{\mu}$. □

Proof of Theorem 3.4. The sequence Ξ_t^N is tight in D' if $\left\langle \phi, \Xi_t^N \right\rangle$ is tight for every $\phi \in C_0(\mathcal{U} \times \mathbb{R}^{d_Y})$ (for instance, see [33]). Since each ϕ is bounded, tightness of the sequence Ξ_t^N follows trivially. For each $u \in \mathcal{U}$ and any $\phi \in C_0(\mathbb{R}^{d_Y})$, we have that

$$\begin{aligned} \left\langle \phi(y), \Xi_t^N(u, dy) \right\rangle = \sum_{u' \in \mathcal{U}} \Bigg[& \left\langle \phi(y), h(u, u', y, V_{t'<t}, H^N_{t'<t}) \mu^N_{t-1}(u', dy) \right\rangle \\ - & \left\langle \phi(y), h(u, u', y, V_{t'<t}, H^N_{t'<t}) \bar{\mu}_{t-1}(u', dy) \right\rangle \Bigg] + \left\langle \phi, \mathcal{M}_t^{2,N}(u) \right\rangle, \end{aligned} \tag{24}$$

where $\mathcal{M}_t^{2,N}$ is the martingale:

$$\left\langle \phi, \mathcal{M}_t^{2,N}(u) \right\rangle = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \phi(Y^n)(\mathbf{1}_{U_t^n = u} - h(u, U^n_{t-1}, Y^n, V_{t'<t}, H^N_{t'<t}). \tag{25}$$

Using a Taylor expansion, one can obtain

$$\begin{aligned} \left\langle \phi(y), \Xi_t^N(u, dy) \right\rangle = & \sum_{u' \in \mathcal{U}} \left\langle \phi(y) h(u, u', y, V_{t'<t}, \bar{H}_{t'<t}), \Xi^N_{t-1}(u', dy) \right\rangle \\ + & \sum_{u' \in \mathcal{U}} \left\langle \phi(y) h_H(u, u', y, V_{t'<t}, \bar{H}_{t'<t}) \cdot E^N_{t'<t}, \mu^N_{t-1}(u', dy) \right\rangle + \frac{1}{\sqrt{N}} \mathcal{R}^N + \left\langle \phi, \mathcal{M}^{2,N}(u, dy) \right\rangle, \end{aligned} \tag{26}$$

where $\mathcal{R}^N$ is the Taylor remainder term which can be bound using Assumption 3.3.

$$\mathcal{R}^N \leq C ||E^N_{t'<t}||^2. \tag{27}$$

Conditional on each V, assume that $\{\Xi^N_{t'}\}_{t'<t}$ weakly converges to $\{\bar{\Xi}_{t'}\}_{t'<t}$. By the continuous mapping theorem and Slutsky's theorem, we have that $\frac{1}{\sqrt{N}} \mathcal{R}^N$ converges in probability to zero. By the martingale

central limit theorem, the martingale term $\left\langle \phi, \mathcal{M}^{2,N}(u,dy) \right\rangle$ converges (conditional on V) to a mean-zero Gaussian $\left\langle \phi, \bar{\mathcal{M}} \right\rangle$ with covariance:

$$Cov\Big[\left\langle \phi_1, \bar{\mathcal{M}}_t(u_1,dy) \right\rangle, \left\langle \phi_2, \bar{\mathcal{M}}_t(u_2,dy) \right\rangle \Big] = -\sum_{u' \in \mathcal{U}} \left\langle \phi_1(y)\phi_2(y) h(u_1,u',y,V_{t'<t},\bar{H}_{t'<t}) h(u_2,u',y,V_{t'<t},\bar{H}_{t'<t}), \bar{\mu}_{t-1} \right\rangle,$$

$$Var\Big[\left\langle \phi, \bar{\mathcal{M}}_t(u,dy) \right\rangle \Big] = \sum_{u' \in \mathcal{U}} \left\langle \phi(y)^2 h(u,u',y,V_{t-1},\bar{H}_{t-1})(1 - h(u,u',y,V_{t'<t},\bar{H}_{t'<t}), \bar{\mu}_{t-1} \right\rangle,$$

where $u_1 \neq u_2$. Furthermore, it is easy to see that conditional on each V, $\left\langle \phi, \bar{\mathcal{M}}_{t_1} \right\rangle$ is independent of $\left\langle \phi, \bar{\mathcal{M}}_{t_2} \right\rangle$ for $t_1 \neq t_2$:

$$\begin{aligned}
&\mathbb{E}\big[\exp\big(i\alpha_1 \left\langle \phi, \mathcal{M}^{2,N}_{t_1}(u) \right\rangle + i\alpha_2 \left\langle \phi, \mathcal{M}^{2,N}_{t_2}(u') \right\rangle \big) | V \big] \\
&= \mathbb{E}\big[\exp\big(i\alpha_1 \left\langle \phi, \mathcal{M}^{2,N}_{t_1}(u) \right\rangle \big) \mathbb{E}[\exp\big(i\alpha_2 \left\langle \phi, \mathcal{M}^{2,N}_{t_2}(u') \right\rangle \big) | \{\mu^N_{t'}\}_{t'<t_2}, V] | V \big] \\
&\underset{N \longrightarrow \infty}{\longrightarrow} \mathbb{E}\big[\exp\big(i\alpha_1 \left\langle \phi, \bar{\mathcal{M}}_{t_1}(u) \right\rangle \big) | V \big] \mathbb{E}[\exp\big(i\alpha_2 \left\langle \phi, \bar{\mathcal{M}}_{t_2}(u') \right\rangle \big) | V \big],
\end{aligned}$$

where the last equality follows from the boundedness of the characteristic functions, the weak convergence of $\exp\big(i\alpha_1 \left\langle \phi, \mathcal{M}^{2,N}_{t_1}(u) \right\rangle \big)$, and the convergence in probability of $\mathbb{E}[\exp\big(i\alpha_2 \left\langle \phi, \mathcal{M}^{2,N}_{t_2}(u') \right\rangle \big) | \{\mu^N_{t'}\}_{t'<t_2}, V]$.

The second term in 26 also weakly converges since μ^N_{t-1} converges in probability to $\bar{\mu}_{t-1}$ for each V. Then, for each V, assuming $\{\Xi^N_{t'}\}_{t'<t}$ weakly converges to $\{\bar{\Xi}_{t'}\}_{t'<t}$, any accumulation point of Ξ^N_t satisfies the evolution equation:

$$\begin{aligned}
\left\langle \phi(y), \bar{\Xi}_t(u,dy) \right\rangle &= \sum_{u' \in \mathcal{U}} \left\langle \phi(y) h(u,u',y,V_{t'<t},\bar{H}_{t'<t}), \bar{\Xi}_{t-1}(u',dy) \right\rangle \\
&+ \sum_{u' \in \mathcal{U}} \left\langle \phi(y) h_H(u,u',y,V_{t'<t},\bar{H}_{t'<t}) \cdot \bar{E}_{t'<t}, \bar{\mu}_{t-1} \right\rangle + \left\langle \phi, \bar{\mathcal{M}}_t(u,dy) \right\rangle, \quad \phi \in C_0(\mathbb{R}^{d_Y}).
\end{aligned} \tag{28}$$

Note that there is a unique distribution $\bar{\Xi}_t$ which satisfies this equation. (Again, suppose there are two solutions $\bar{\Xi}^1_t$ and $\bar{\Xi}^2_t$; let their difference be $\nu = \bar{\Xi}^1_t - \bar{\Xi}^2_t$. Substituting into the above equation yields that $\langle \phi, \nu \rangle = 0$ for every ϕ, which implies uniqueness.) Therefore, for each V, $\Xi^N_t \overset{d}{\to} \bar{\Xi}_t$ and, moreover, $\{\Xi^N_{t'}\}_{t' \leq t} \overset{d}{\to} \{\bar{\Xi}_{t'}\}_{t' \leq t}$. Since $\Xi^N_0 \overset{d}{\to} \bar{\Xi}_0$, $\Xi^N \overset{d}{\to} \bar{\Xi}$ for each V by induction. The weak convergence holds for every V and, consequently, the general theorem follows. □

Proof of Theorem 4.2. Define the empirical measure $\mu^N_{t,t-1} = \frac{1}{N} \sum_{n=1}^N \delta_{U^n_{t-1}, U^n_t, Y^n}$ and recall that the likelihood is

$$\begin{aligned}
\mathcal{L}_{T,N} &\propto \frac{1}{N} \sum_{t=1}^T \sum_{n=1}^N \log h_\theta(U^n_t, U^n_{t-1}, Y^n, V_{t'<t}, H^N_{t'<t}) \\
&= \sum_{t=1}^T \left\langle \log h_\theta(u_t, u_{t-1}, y, V_{t'<t}, H^N_{t'<t}), \mu^N_{t,t-1} \right\rangle,
\end{aligned} \tag{29}$$

and

$$\theta_{T,N} = \arg\max_{\theta \in \Theta} \mathcal{L}_{T,N}. \tag{30}$$

For every $\theta \in \Theta$ and each V, we have that $\mathcal{L}_{T,N}(\theta) \overset{p}{\to} \mathcal{L}_{T,\infty}$ where $\mathcal{L}_{T,\infty} = \left\langle \log h_\theta(u_t, u_{t-1}, y, V_{t'<t}, H^N_{t'<t}), \bar{\mu}_{t,t-1} \right\rangle$. The weak convergence of $\mu^N_{t,t-1}$ to $\bar{\mu}_{t,t-1}$ can be proven by the same approach as for $\mu^N_t \overset{d}{\to} \bar{\mu}_t$ and $\bar{\mu}_{t,t-1}(u,u',dy) = h_{\theta_0}(u,u',y,V_{t'<t},\bar{H}_{t'<t}) \bar{\mu}_{t-1}(u',dy)$. It is sufficient to prove that $\theta_{T,N} \overset{p}{\to} \theta_0$ for each V; since θ_0 is a constant, this implies that $\theta_{T,N} \overset{p}{\to} \theta_0$ by Slutsky's theorem. For each V, we prove stochastic equicontinuity of the likelihood. By the mean value theorem,

$$\lim_{N \longrightarrow \infty} \mathbb{P}_{\theta_0} \big[\sup_{|\theta_1 - \theta_2| \leq \delta} |\mathcal{L}_{T,N}(\theta_1) - \mathcal{L}_{T,N}(\theta_2)| > \epsilon | V \big]$$

$$\begin{aligned} &= \lim_{N\longrightarrow\infty} \mathbb{P}_{\theta_0}\big[\delta \sup_{\theta\in\Theta} \big|\sum_{t=1}^{T} \Big\langle \frac{\partial}{\partial\theta} \log h_\theta(u_t, u_{t-1}, y, V_{t'<t}, H^N_{t'<t}), \mu^N_{t,t-1} \Big\rangle \big| > \epsilon | V\big] \\ &\leq \mathbb{P}_{\theta_0}\big[\delta \sup_{K_\eta} \big|\frac{\partial}{\partial\theta} \log h_\theta\big| > \epsilon|V\big](1-\eta) + \eta. \end{aligned} \tag{31}$$

We have used the fact that μ^N is tight, so there exists a compact set K_η in which μ^N lies with probability $1-\eta$ for all N. Under the assumption that h_θ and $\frac{\partial}{\partial\theta} h_\theta$ are continuous, $\frac{\partial}{\partial\theta}\log h_\theta$ takes a maximum on the compact set $K_\eta \times \Theta$. Pointwise convergence and stochastic equicontinuity imply that, for each V,

$$\theta_{T,\infty} \in \arg\max_{\theta\in\Theta} \mathcal{L}_{T,\infty}. \tag{32}$$

We now must show that the limiting estimator is unique. Clearly, $\theta_{T,\infty}$ must be a $\theta \in \Theta$ such that $h_\theta(u, u', y, V_{t'<t}, \bar{H}_{t'<t}) = h_{\theta_0}(u, u', y, V_{t'<t}, \bar{H}_{t'<t})$ over every set of (u', y) which has positive measure under $\bar{\mu}_t$ since

$$\mathcal{L}_{T,\infty} = \sum_{t=1}^{T} \int_{\mathbb{R}^{d_Y}} \sum_{u,u'\in\mathcal{U}} h_{\theta_0}(u, u', y, V_{t-1}, \bar{H}_{t'<t}) \log\big(h_\theta(u, u', y, V_{t-1}, \bar{H}_{t'<t})\big)\bar{\mu}_{t-1}(u', dy), \tag{33}$$

whose integrand is uniquely maximized by $h_\theta(u, u', y, v, H) = h_{\theta_0}(u, u', y, v, H)$. This can always be achieved for every point by choosing $\theta = \theta_0$. The question remains whether there are other $\theta \in \Theta$ such that $h_\theta(u, u', y, v, H) = h_{\theta_0}(u, u', y, v, H)$ holds over every set of (u', y) for which $\bar{\mu}_{t-1}$ has positive measure. Assumption 4.1 implies that, for almost every V, there is a unique θ such that $h_\theta(u, u', y, v, H) = h_{\theta_0}(u, u', y, v, H)$ holds over every set of (u', y) for which $\bar{\mu}_{t-1}$ has positive measure. Therefore, the limiting likelihood is almost surely uniquely maximized by θ_0. It follows that the MLE $\theta_{T,N}$ converges in probability to the true parameter θ_0. □

Proof of Theorem 4.3. Recall that

$$\begin{aligned} \mathcal{L}_{T,N} &= \frac{1}{N} \sum_{t=1}^{T} \sum_{n=1}^{N} \log h_\theta(U^n_t, U^n_{t-1}, Y^n, V_{t'<t}, H^N_{t'<t}) \\ &= \sum_{t=1}^{T} \left\langle \log h_\theta(u_t, u_{t-1}, y, V_{t'<t}, H^N_{t'<t}), \mu^N_{t,t-1} \right\rangle. \end{aligned} \tag{34}$$

Define $\Xi^N_{t,t-1} = \sqrt{N}(\mu^N_{t,t-1} - \bar{\mu}_{t,t-1})$. $\Xi^N_{t,t-1} \xrightarrow{d} \bar{\Xi}_{t,t-1}$, where

$$\begin{aligned} \bar{\Xi}_{t,t-1}(u, u', dy) &= h(u, u', y, V_{t-1}, \bar{H}_{t'<t})\bar{\Xi}_{t-1}(u', dy) \\ &+ (\frac{\partial h}{\partial H}(u, u', y, V_{t-1}, \bar{H}_{t'<t}) \cdot \bar{E}_{t'<t})\bar{\mu}_{t-1}(u', dy) + \bar{\mathcal{M}}_t(u, u', dy), \\ Var[\bar{\mathcal{M}}_t(u, u', dy)] &= h_\theta(u, u', y, V_{t'<t}, \bar{H}_{t'<t})(1 - h_\theta(u, u', y, V_{t'<t}, \bar{H}_{t'<t}))\bar{\mu}_{t-1}(u', dy). \end{aligned} \tag{35}$$

This can be proven with the same approach which was used to show $\Xi^N_t \xrightarrow{d} \bar{\Xi}_t$.

By Taylor's theorem,

$$\sqrt{N}(\theta_{T,N} - \theta_0) = -(\frac{\partial^2 \mathcal{L}_{T,N}}{\partial\theta^2}(\theta^1_{T,N}))^{-1} \sqrt{N} \frac{\partial \mathcal{L}_{T,N}}{\partial\theta}(\theta_0), \tag{36}$$

where $\theta^1_{T,N} \in [\theta_{T,N}, \theta_0]$.

From the previous weak convergence result, convergence of the estimator $\theta_{T,N}$ to the true parameter θ_0 as $N \longrightarrow \infty$, Slutsky's theorem, and the continuous mapping theorem, we have that, for each V, $(\frac{\partial^2 \mathcal{L}_{T,N}}{\partial\theta^2}(\theta^1_{T,N}))^{-1}$ converges in probability to $(\frac{\partial^2 \mathcal{L}_{T,\infty}}{\partial\theta^2}(\theta_0))^{-1}$. Similarly, we have that, for each V,

$$\begin{aligned} \sqrt{N} \sum_{t=1}^{T} \left\langle \frac{\partial \log h_{\theta_0}}{\partial\theta}(u_t, u_{t-1}, y, V_{t'<t}, H^N_{t'<t}), \mu^N_{t,t-1} \right\rangle &\xrightarrow{d} \textstyle\sum_{t=1}^{T} [\Big\langle \frac{\partial \log h_{\theta_0}}{\partial\theta}(\cdot, \bar{H}_{t'<t}), \bar{\Xi}_{t,t-1} \Big\rangle \\ &+ \Big\langle \frac{\partial^2 \log h_{\theta_0}}{\partial\theta\partial H}(\cdot, \bar{H}_{t'<t}) \cdot \bar{E}_{t'<t}, \bar{\mu}_{t,t-1} \Big\rangle]. \end{aligned} \tag{37}$$

□

Appendix B. Additional Tables

Data Set	Loan-level Features	Monthly Updates	Termination Events
Subprime	zip code, FICO, LTV, initial interest rate, initial balance, mortgage type	Modification, REO, foreclosure, default, prepayment	Default or prepayment
Agency	MSA, FICO, LTV, initial interest rate, initial balance, mortgage type, first time home buyer indicator, number of units, occupancy status, combined loan-to-value, debt-to-income ratio, prepayment penalty indicator, property type, loan purpose, number of borrowers	current interest rate, modification indicator, repurchased indicator, current, days delinquent (30-59 days, 60-89 days, 90-119 days, 120-149 days, 150-179 days, or 180 or more days), and termination event if it occurs	Prepaid or matured, third party sale prior to 180 days delinquent, short sale or short payoff, deed-in-lieu of foreclosure, repurchase, REO acquisition, 180 days delinquent

Table 4. Comparison of the subprime and agency data sets.

Factor	Parameter
Constant	-4.476
National unemployment rate	-0.1502
National mortgage rate	-0.4464
FICO score	0.1787
First time homebuyer	-.0392
Number of units	-0.0282
Occupancy status: owned	0.0992
Occupancy status: investment property	-0.1272
Occupancy status: second home	-0.0063
Combined loan-to-value ratio	-0.0046
Loan-to-value ratio	-0.0296
Initial interest rate	0.7689
Prepayment penalty flag	0.0108
Property type: condo	0.0683
Property type: leasehold	.0422
Property type: PUD	0.1698
Property type: manufactured housing	-.0583
Property type: 1-4 fee simple	.1884
Property type: Co-op	.0464
Loan purpose: purchase	0.0531
Property type: cash-out refinance	-0.0541
Property type: no cash-out refinance	-0.0051
Number of borrowers: 1 borrower	-0.2330
Number of borrowers: more than 1 borrower	-0.0703
Debt-to-income	-0.0201

Table 5. Parameter fits for prepayment model for agency mortgage data set.

References

[1] K. Dunn and J. McConnell. Valuation of GNMA mortgage-backed securities. *The Journal of Finance*, 36(3):599–616, 1981.

[2] A. Hall. Valuing the mortgage borrower's prepayment option. *Real Estate Economics*, 13(3):229–247, 1985.

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

[4] R. Stanton. Rational prepayment and the valuation of mortgage-backed securities. *Review of Financial Studies*, 8(3):677–708, 1995.

[5] E. Schwartz and W. Torous. Prepayment and the valuation of mortgage-backed securities. *The Journal of Finance*, 44(2):375–392, June 1989.

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

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

[8] V. Gorovoy and V. Linetsky. Intensity-based valuation of residential mortgages: an analytically tractable model. *Mathematical Finance*, 17(4):541–573, 2007.

[9] H. Stein, A. Belikoff, K. Levin, and X. Tian. Analysis of mortgage-backed securities: before and after the credit crisis. *Credit Risk Frontiers: Subprime Crisis, Pricing and Hedging, CVA, MBS, Ratings, and Liquidity*, pages 345–394, 2007.

[10] T. Williams. Distributed calculations on fixed-income securities. In *Proceedings of the 2nd Workshop on High Performance Computational Finance*. Portland, OR, November 2009.

[11] S. Zenios. Parallel monte carlo simulation of mortgage-backed securities. *Financial Optimization*, page 325, 1996.

[12] J. Fermanian. A top-down approach for MBS, ABS and CDO of ABS: a consistent way to manage prepayment, default and interest rate risks. *Journal of Real Estate Finance and Economics*, 46(3), February 2013.

[13] S. Wu, L. Jiang, and J. Liang. Intensity-based models for pricing mortgage-backed securities with repayment risk under a cir process. *International Journal of Theoretical and Applied Finance*, 15(3), 2012.

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

[15] P. Kang and S. Zenios. Complete prepayment models for mortgage-backed securities. *Management Science*, 38(11):1665–1685, 1992.

[16] J. Mattey and N. Wallace. Housing-price cycles and prepayment rates of us mortgage pools. *The Journal of Real Estate Finance and Economics*, 23(2):161–184, 2001.

[17] S. Chinchalkar and R. Stein. Comparing loan-level and pool-level mortgage portfolio analysis. Technical report, Moody's Research Labs, 2010.

[18] N. Bush, B. M. Hambly, H. Haworth, L. Jin, and C. Reisinger. Stochastic evolution equations in portfolio credit modelling. *SIAM Journal of Financial Mathematics*, 2(1):627–664, 2011.

[19] J. Cvitanic, J. Ma, and J. Zhang. The law of large numbers for self-exciting correlated defaults. *Stochastic Processes and their Applications*, 122(8):2781–2810, 2012.

[20] P. Dai Pra, W.J. Runggaldier, E. Sartori, and M. Tolotti. Large portfolio losses: A dynamic contagion model. *The Annals of Applied Probability*, 19(1):347–394, 2009.

[21] P. Dai Pra and M. Tolotti. Heterogeneous credit portfolios and the dynamics of the aggregate losses. *Stochastic Processes and their Applications*, 119(9):2913–2944, 2009.

[22] K. Giesecke, K. Spiliopoulos, R.B. Sowers, and J.A. Sirignano. Large portfolio asymptotics for loss from default. *Mathematical Finance*, 2014, in press.

[23] K. Giesecke and S. Weber. Credit contagion and aggregate losses. *Journal of Economic Dynamics and Control*, 30(5):741–767, 2006.

[24] K. Spiliopoulos, J.A. Sirignano, and K. Giesecke. Fluctuation analysis for the loss from default. *Stochastic Processes and their Applications*, 124:2322–2362, 2014.

[25] R. Goodstein, P. Hanouna, C. Ramirez, and C. Stahel. Contagion effects in strategic mortgage defaults. GMU Working Paper in Economics No. 13-07, 2011.

[26] B. Ambrose and C. Capone. Modeling the conditional probability of foreclosure in the context of single-family mortgage default resolutions. *Real Estate Economics*, 26(3):391–429, 1998.

[27] Z. Lin, E. Rosenblatt, and V. Yao. Spillover effects of foreclosures on neighborhood property values. *Journal of Real Estate Finance and Economics*, 38(4):387–407, May 2009.

[28] J. Harding, Eric Rosenblatt, and V. Yao. The contagion effect of foreclosed properties. *Journal of Urban Economics*, 66:164–178, July 2009.

[29] C. Towe and C. Lawley. The contagion effect on neighboring foreclosures on own foreclosures. Working paper, University of Maryland and University of Manitoba, 2010.

[30] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[31] J. Sirignano and K. Giesecke. Geographic risk for mortgages. Working Paper, September 2014.

[32] E. Stewart and T. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley and Sons Inc., 1986.

[33] Jean-Pierre Fouque. La convergence en loi pour les processus a valeurs dans un espace nucleaire. *Annales de l'I.H.P.*, 20:225–245, 1984.

Department of Management Science and Engineering, Stanford University, Stanford, CA 94305
E-mail address: jasirign@stanford.edu

Department of Management Science and Engineering, Stanford University, Stanford, CA 94305
E-mail address: giesecke@stanford.edu