The Failure of Models That Predict Failure: Distance, Incentives and Defaults[*]

Uday Rajan[†]
Amit Seru[‡]
Vikrant Vig[§]

January 2009

[*] We thank Viral Acharya, Patrick Bolton, John Cochrane, Douglas Diamond, Anil Kashyap, Tobias Moskowitz, Raghuram Rajan, Jeremy Stein, Richard Thaler, Luigi Zingales and seminar participants at the Bank of England, FRB Chicago, Houston, IMF, LSE, Michigan, Michigan State, Naples, University of Chicago, Utah and Western Ontario for useful suggestions. Seru thanks the Initiative on Global Markets at the University of Chicago for financial support. Part of this work was undertaken when Seru was a visiting scholar at Sorin Capital Management. We are also indebted to Tanmoy Mukherjee for extensive discussions. All errors are our responsibility.

[†] Ross School of Business, University of Michigan, e-mail: urajan@umich.edu

[‡] Booth School of Business, University of Chicago, e-mail: amit.seru@chicagogsb.edu

[§] London Business School, e-mail: vvig@london.edu

The Failure of Models That Predict Failure: Distance, Incentives and Defaults

Abstract

Using data on securitized subprime mortgages issued in the period 1997–2006, we demonstrate that, as the degree of securitization increases, interest rates on new loans rely increasingly on hard information about borrowers. As a result, a statistical default model fitted in a low securitization period breaks down in the high securitization period in a systematic manner: it underpredicts defaults among borrowers for whom soft information is more valuable (i.e., borrowers with low documentation, low FICO scores and high loan-to-value ratios). We rationalize these findings in a theoretical model that highlights a reduction in lenders' incentives to collect soft information as securitization becomes common, resulting in worse loans being issued to borrowers with similar hard information characteristics. Our results partly explain why statistical default models severely underestimated defaults during the subprime mortgage crisis, and imply that these models are subject to a Lucas critique. Regulations that rely on such models to assess default risk may therefore be undermined by the actions of market participants.

I Introduction

There has been a genuine surprise among practitioners, regulators and investors as the valuation of subprime loans backed securities fell rapidly during the subprime crisis. The ABX index that tracks credit default swaps based on AAA subprime tranches fell by about 45% over the course of eight months starting in July 2007 (see Greenlaw, et al., 2008). Behind the valuation of these tranches is a statistical model such as the S&P LEVELS® 6.1 Model that estimates defaults on the underlying collateral. Similar models have been used widely across the financial markets, to enhance market liquidity and impose capital requirements on financial institutions. Why did statistical default models for subprime mortgages fare so poorly in this period?[1] We argue that a fundamental cause for this failure was that the models relied entirely on hard information variables such as borrower credit scores, and ignored changes in the incentives of lenders to collect soft information about borrowers.[2] Thus, echoing the classical Lucas critique (Lucas, 1976), these models failed to account for the change in the relationship between observable borrower characteristics and default likelihood caused by a fundamental change in lender behavior.

What changed the behavior of lenders in the subprime market? There was a tremendous growth in securitization (converting illiquid assets into liquid securities) in the subprime sector after 2000. Securitization increases the distance between the originator of the loan and the party that bears the default risk inherent in the loan. Since soft information about borrowers is unverifiable to a third party (as in Stein, 2002), the increase in distance results in lenders choosing to not collect soft information about borrowers. Consequently, among borrowers with similar hard information characteristics, the set that receives loans changes in a fundamental way as the securitization regime changes. This leads to a breakdown in the quality of predictions from default models that use parameters estimated using data from a period in which a low proportion of loans are securitized. Importantly, the breakdown is systematic, and therefore predictable: It occurs in the set of borrowers for whom soft information is especially valuable.

We formalize this intuition in a theoretical model of loan origination and derive a number of testable predictions. We test these predictions on a database that contains information on securitized subprime mortgage loans in the period 1997–2006 and find conclusive support. First, we demonstrate that the interest rate on new loans relies increasingly on hard information as securitization increases. Specifically, the R^2 of a regression of interest rates on borrower FICO scores and loan-to-value (LTV) ratios increases from 3% in 1997 to almost 50% in 2006. Further,

[1] For example, in November 2007, S&P adjusted its LEVELS® default model to increase predicted defaults on no documentation loans by approximately 60% (see Standard & Poor's, 2007).

[2] For example, risk calculators used by rating agencies such as S&P, Moody's and Fitch rely on estimating credit risk from hard information variables such as the borrower's credit (FICO) score and the geographic location of the property (see, for example, the FitchRatings report on the Fitch default model, October 2006).

conditioning on the FICO score, the variance of interest rates on loans shrinks over time. The latter effect occurs especially for borrowers with low FICO scores, on whom soft information is more important, implying a loss of soft information about borrowers. The effect survives when we control for standardization of other contractual terms over time.

Second, we estimate a statistical default model from loans issued in a period with a low degree of securitization (1997–2000), using hard information variables about borrowers. We show that the statistical model underpredicts defaults on loans issued in a regime with high securitization (2001 onwards). The degree of underprediction progressively worsens as the securitization increases, suggesting that at the same hard information characteristics, the set of borrowers receiving loans has worsened over time. Since lenders are no longer collecting soft information about borrowers, we expect the prediction errors to be particularly high when soft information is valuable; that is, for borrowers with low FICO scores and high LTV ratios. Indeed, we find a systematic variation in the prediction errors; they increase as the borrower's FICO score falls and the LTV ratio increases.

We perform two additional tests to confirm our results on the failure of the default model. First, we separately consider loans with full documentation and loans with low documentation. Full-documentation loans include hard information on a borrower's job, income and assets, making soft information less valuable for such borrowers. As expected, the prediction errors from the default model in the high securitization era are lower for full-documentation loans. Second, as a placebo test, we estimate a default model for low-documentation loans over a subset of the low securitization era, and examine its out-of-sample predictions on loans issued in 1999 and 2000 (also a low securitization period). The model performs significantly better than in our main test, and in particular yields prediction errors that are approximately zero on average.

We find that the default model underpredicts errors even for loans issued in the period 2001–2004, while house prices were increasing. Nevertheless, falling house prices (or, more broadly, an economic decline) likely contributed to the increase in defaults in the later part of the sample (i.e., loans issued in 2005 and 2006), and may have a disproportionately adverse impact on borrowers with low FICO scores. To account for this effect, we consider a stringent specification that both estimates the baseline model over a rolling window, and also explicitly accounts for the effects of changing house price. To do the latter, we determine the statewide change in house prices for two years *after* the loan has been issued, and explicitly include it as an explanatory variable in the default model. Approximately 50% of the prediction error survives the new specification, and the qualitative results remain: a default model estimated in a low securitization regime continues to systematically underpredict defaults in a high securitization regime.

Our empirical predictions follow directly from a theoretical model of loan origination, in which a lender may acquire both hard information (such as a FICO score) and soft information about a borrower. Following Stein (2002), by soft information we refer to any information (about the borrower or the property) that is not easily documentable or verifiable. Borrowers have types, and both hard and soft information play a valuable role in screening loan applicants. However, soft information is costly. A lender chooses to incur the cost of acquiring soft information if the hard information signal is imprecise and the lender plans to retain the loan on its balance sheet. Now consider a regime in which loans are securitized; i.e., sold to an investor rather than being retained on the books of the lender. Since soft information cannot be verified by an independent observer, it is perforce uncontractible, and the price investors offer for a loan (or pool of loans) must depend only on the associated hard information. This creates a moral hazard problem for the lender.

To see this, suppose there is a set of borrowers who all generate the same hard information signal, and the lender acquires soft information that perfectly distinguishes between good borrowers (those likely to repay the loan) and bad ones (those likely to default). Suppose further that investors price loans as if the lender is screening out bad borrowers, and only making loans to good borrowers. Since soft information cannot be credibly communicated to the investor, the lender has an incentive to deviate and issue loans to both types of borrowers. In equilibrium, investors compensate for the adverse selection and price the loans accordingly, recognizing that both types of borrowers are pooled in the loan sample. This further implies that the lender has no incentive to collect soft information. Thus, the model implies that the set of borrowers who receive loans changes in a fundamental way across securitization regimes.

Since our work directly provides a Lucas critique on the use of statistical default models, it implies that regulations based on such models (for instance, as recommended in the Basel II guidelines) can be undermined by the actions of market participants. More broadly, the paper provides evidence that supports theories on incentive effects related to hard and soft information about borrowers (Stein, 2002). A detailed discussion of both these issues and other implications of our paper is deferred to Section VI.

The rest of this paper is organized as follows. The model and theoretical results are contained in Section II, the data are described in Section III and Section IV details the main empirical findings. In Section V, we consider the robustness of our findings. Section VI elaborates on the connections of our work with the existing literature and discusses policy implications of our findings.

II Model

We develop a stylized theoretical model that captures the salient features of the loan market and highlights the effects of securitization on the tradeoffs faced by a lender in screening loan applicants. We focus in particular on the incentives of the lender to collect soft information about borrowers. Our goal is not to explain the features of the market, but rather to develop empirical predictions taking these features as given. These empirical predictions are then tested later in the paper.

There are three sets of agents in the model: borrowers, a single lender, and investors. At date 0, a borrower applies for a loan to be repaid at date 1. The loan size is homogeneous across types, and is normalized to 1. At date 0, the lender costlessly observes a hard information signal x about the borrower. Based on the hard information signal, the lender decides whether to incur a cost c to obtain a soft information signal y. Using all available information, the lender offers the borrower an interest rate r. The borrower accepts or rejects the loan offer. Finally, a fixed proportion of loans made by the lender, $\alpha \in [0, 1]$, are securitized.

There is a continuum of borrowers, with each borrower having a type $\theta \in \{\theta_h, \theta_\ell, \theta_b\}$. Types are independent and identically distributed across borrowers. Let p_j denote the prior probability a borrower has type θ_j. A borrower with type θ_j finds herself in a good state with respect to her personal finances at time 1 with probability θ_j. In this event, she repays her loan if the interest rate is sufficiently low (in a manner made precise below). With probability $1 - \theta_j$, she is in a bad state at time 1 and defaults, in which case the lender recovers zero. In equilibrium, the types will correspond to the likelihood of repayment on the loan. We assume that $\theta_h > \theta_\ell > \theta_b$.

For $j = h, \ell$, a borrower of type θ_j has a reservation interest rate $\tau(\theta_j)$ that depends on her (unmodeled) outside opportunities (which could include applying to and obtaining a loan from another lender). For convenience, let $\tau(\theta_h) = r_1$, $\tau(\theta_\ell) = r_2$. Less risky types have better outside opportunities, so the reservation interest rates satisfy $r_1 < r_2$. Both these types repay their loans in full in their respective good states if the interest rate is weakly less than r_2. Since neither type will accept a loan at an interest rate greater than r_2, the repayment probabilities at higher interest rates are irrelevant.

The lender has a cost of funds, or discount rate, normalized to zero. Define the net present value of a loan to type θ_j at interest rate $r \leq r_2$ as $v_j(r) = \theta_j(1 + r) - 1$ for $j = h, \ell, b$. We assume that $v_h(r_1) > v_\ell(r_2) > 0$, so that the high type offers a higher NPV than the low type at their respective reservation interest rates.

A borrower of type θ_b accepts any loan that is offered, so her reservation interest rate may be thought of as infinite. Thus, in what follows, we define $\tau(\theta_b) = \infty$. The θ_b type repays her loan in the good state if and only if the interest rate on the loan is no higher than $r_b > r_2$. The maximal interest rate r_b captures the idea that even in the good state a borrower will face a budget constraint. A loan made to this type has negative NPV regardless of the interest rate at which it is offered, so $v_b(r) < 0$ for all r. Therefore, in a full information world, types θ_h and θ_ℓ would obtain a loan, and type θ_b would not.

On each borrower, the lender obtains a hard information signal $x \in \{x_h, x_\ell, x_b\}$ at zero cost. The hard information incorporates verifiable data such as the borrower's FICO credit score and tax returns. Let $\delta(x_i \mid \theta_j)$ be the probability that the hard information signal is x_i, when the borrower's true type is θ_j. Hard information signals are conditionally independent across borrowers. The hard information signal is informative in the following sense: if $\theta_i > \theta_j$, then $\delta(x_h \mid \theta_i) \geq \delta(x_h \mid \theta_j)$ and $\delta(x_h \mid \theta_i) + \delta(x_\ell \mid \theta_i) \geq \delta(x_h \mid \theta_j) + \delta(x_\ell \mid \theta_j)$, with at least one of the relationships being strict.

Having seen the hard information signal, the lender may choose to obtain a soft information signal about the borrower, $y \in \{y_h, y_\ell, y_b\}$. The soft information signal is obtained at a cost c, which includes items such as the value of the time of a loan officer who has to interview the borrower or examine the borrower's file. Soft information here includes any information related to the likelihood of default that is not verifiable by a third party. The information could pertain to the borrower or to the property. It includes, for example, the likelihood that the borrower's job may be terminated or she will be credit-constrained in the future, and information on income or assets the borrower cannot document. It also includes information pertaining to the property, such as the quality of the appraisal. For example, if the lender finds out the reported home value has been inflated by the appraiser, it may infer that the borrower is less likely to be able to repay the loan.

Let $\gamma(y_i \mid \theta_j, x_k)$ be the probability that the soft information signal is y_i, given that the borrower's true type is θ_j and the hard information signal was x_k. Soft information signals are conditionally independent (given borrower type and hard information signal) across borrowers. The soft information signal is also informative: Suppose $\theta_i \geq \tilde{\theta}_i$ and $x_j \geq \tilde{x}_j$, with at least one strict inequality. Then, $\gamma(y_h \mid \theta_i, x_j) \geq \gamma(y_h \mid \tilde{\theta}_i, \tilde{x}_j)$ and $\gamma(y_h \mid \theta_i, x_j) + \gamma(y_\ell \mid \theta_i, x_j) \geq \gamma(y_h \mid \tilde{\theta}_i, x_j) + \gamma(y_\ell \mid \tilde{\theta}_i, \tilde{x}_j)$, with at least one strict inequality.

Given the signals it has observed, the lender chooses to either offer the borrower a loan at a specified interest rate or not offer a loan. A borrower with types θ_h or θ_ℓ accept a loan offer if the interest rate is weakly less than her reservation interest rate, and rejects otherwise. A type θ_b borrower always accepts a loan offer. For now, suppose there is no securitization, so that $\alpha = 0$. Then, it is straightforward to see that if a loan is offered, the optimal interest rate offered by the bank must be either r_1 (which is accepted by all three types of borrowers), r_2 (which is accepted only by types θ_ℓ and θ_b).

Given a hard information signal x, let $r^*(x)$ denote the interest rate that maximizes the lender's expected profit (if a loan is offered) when there is no securitization (i.e., $\alpha = 0$). We assume that if the hard information signal is x_h or x_ℓ, it is a strict best response for the lender to offer a loan, with the unique optimal interest rate being $r^*(x_h) = r_1$ if the hard information signal is x_h and $r^*(x_\ell) = r_2$ if the hard information signal is x_ℓ. That is, the posterior probability of type θ_h is sufficiently high when $x = x_h$ to enable r_1 to be the optimal interest rate in this case. Similarly, when $x = x_\ell$, the posterior probability of type θ_h is sufficiently low to ensure that r_2 is the optimal interest rate. We further assume that the posterior probability of type θ_b is strictly positive when $x = x_\ell$. Finally, if the hard information signal is x_b, posterior probabilities over types are such that it is strictly optimal to not offer a loan.

Similarly, given hard and soft information signals (x, y), let $r^*(x, y)$ denote the interest rate that maximizes the lender's profit when $\alpha = 0$. We assume that when the hard information signal is x_ℓ and the soft information signal is y_h or y_ℓ, the lender offers a loan, with the unique optimal interest rate being $r^*(x_\ell, y_h) = r_1$ and $r^*(x_\ell, y_\ell) = r_2$. If the hard information signal is x_ℓ and the soft information signal is y_b, again it is optimal to not offer a loan.

Finally, we assume that the cost of acquiring the soft information signal satisfies the following restrictions. For $i = h, \ell, b$, let $\mu_i(x)$ denote the posterior probability that the borrower's type is θ_i, given that the hard information signal was x. Then,

(i) $c \geq \max\{\mu_\ell(x_h)\theta_\ell(r_2 - r_1) - \mu_b(x_h)v_b(r_1),\ \mu_h(x_b)v_h(r_1) + \mu_\ell(x_b)v_\ell(r_2)\}$.

(ii) $c \leq \sum_{i=h,\ell,b} \mu_i(x_\ell)\gamma(y_h \mid \theta_i, x_\ell)v_i(r_1) - \sum_{i=\ell,b} \mu_i(x_\ell)(1 - \gamma(y_\ell \mid \theta_i, x_\ell))v_i(r_2)$.

Part (i) above is satisfied if the hard information signal is precise enough when the signal received is x_h or x_b. For example, suppose the high type always generates signal x_h, and the bad type always generates signal x_b, with the low type generating all three signals with positive probability. Then, (i) reduces to the requirement that $c \geq 0$.

Part (ii) essentially requires that the signal y should be precise enough when the hard information signal is x_ℓ to make it worthwhile to collect soft information. The optimal interest rate offer given a hard information signal x_ℓ is r_2. At this interest rate, type θ_h rejects the offer, and types θ_ℓ, θ_b accept. Thus, further screening (via the soft information signal) can potentially add value in two ways: by identifying θ_h borrowers who can be offered r_1 and θ_b borrowers who can be shut out altogether. Suppose the soft information signal is fully-revealing, with $\gamma(y_j \mid \theta_j, x_\ell) = 1$ for each j. Then, part (ii) reduces to $c \leq \mu_h(x_\ell)v_h(r_1) - \mu_b(x_\ell)v_b(r_2)$. The first term on the right-hand side is the additional payoff from θ_h borrowers who now accept a loan at r_1, and the last term the additional payoff from θ_b borrowers who are screened out (recall that $v_b(r_2) < 0$).

Given our assumptions about c, the second-best outcome is as follows. If the hard information signal is x_h or x_b, the lender does not acquire soft information. However, when the hard information signal is x_ℓ, soft information is valuable. In this outcome, the interest rate offered to the borrower is r_1 if the signals received are x_h or (x_ℓ, y_h) and r_2 if the signals received are x_ℓ or (x_ℓ, y_ℓ), with no loan being offered if the signals received are x_b or (x_ℓ, y_b).

A loan is said to be securitized if it is sold to investors. For any loan made by the lender, investors observe the interest rate on the loan, r, and the hard information associated with the borrower, x. Even if it is acquired, the soft information, y, is not verifiable and therefore not contractible between the investors and the lender. We assume that investors do not observe (or, equivalently, cannot credibly condition on) whether the lender acquires soft information. Financial markets are perfectly competitive, so the price of a loan equals its expected payoff, and investors earn zero profit. Let $P(x, r)$ denote the price of a loan with hard information signal x and interest rate r.

Any particular loan made by a lender is securitized with an exogenous probability α. With probability $(1 - \alpha)$, the lender must retain the loan. It is common in the residential mortgage market for a lender to offer a basket of loans to investors, who randomly select loans in every category. Thus, on any given loan, there is a positive probability the lender will have to retain it. For simplicity, we do not allow the lender to choose whether to retain a loan or offer it to investors. Finally, in our model, securitization corresponds to an outright sale of the loan, without recourse. This again corresponds to practices in the mortgage market. Typically, loans are sold with only a three-month recourse (that is, the investor may return the loan if it defaults within three months), and entire loans are sold, with tranching only occurring at a subsequent stage of the process. In Section II.B, we outline our reasons for treating securitization as exogenous and also comment on the implications of allowing the lender to choose which loans to offer for securitization.

II.A Equilibrium

We consider a perfect Bayesian equilibrium of the game. Given the securitization probability α and the pricing schedule the lender believes investors will offer, the lender chooses whether to acquire the costly soft information signal on each borrower and the interest rate to offer to each borrower. Let $\rho(r \mid x, y)$ denote the probability the lender's offers interest rate r, given signal pair (x, y). If $\rho(r \mid x, y) = 0$ for all r, no loan is offered. With a slight abuse of notation, we let $\rho(\cdot \mid x)$ denote the lender's strategy when soft information is not acquired. Let $\lambda_i(x, \rho)$ denote the posterior probability of type θ_i, given a hard information signal x and lender strategy ρ. In equilibrium, it must be that if $\rho(r \mid x, y) > 0$, then $P(x, r) = (1 + r)\,\frac{\sum_{\{i: r \le \tau(\theta_i)\}} \lambda_i(x,\rho)\theta_i}{\sum_{\{i: r \le \tau(\theta_i)\}} \lambda_i(x,\rho)}$, the expected payoff of the loan at time 1. As usual, perfect Bayesian equilibrium does not restrict investors' beliefs off the equilibrium path (i.e., if $\rho(r \mid x, y) = 0$ or $\rho(r \mid x) = 0$). In such cases, we impose the following beliefs on investors: the posterior probability of type θ_i is $\frac{\mu_i(x) 1_{\{r \le \tau(\theta_i)\}}}{\sum_{\{j: r \le \tau(\theta_j)\}} \mu_j(x)}$ if $r \le r_2$, where $1_{\{r \le \tau(\theta_i)\}}$ is an indicator variable that has value 1 if $r \le \tau(\theta_i)$ and 0 otherwise.

If $r > r_2$, investors believe the borrower is of type θ_b, the only type that will accept a loan at that interest rate.

Let $\psi_i(x, y)$ denote the posterior probability that the borrower's type is θ_i, given that the hard information signal is x and the soft information signal is y. If soft information is not collected, let $\psi_i(x, y) = \mu_i(x)$ for each y. Suppose the lender offers an interest rate r to a borrower with signal pair (x, y). The overall expected payoff of the lender is $u(r, \rho) = (1 - \alpha) \sum_{\{i:r\leq\tau(\theta_i)\}} \psi_i(x, y)v_i(r) + \alpha[P(x, r) - 1] \sum_{\{i:r\leq\tau(\theta_i)\}} \psi_i(x, y)$ as long as $r \leq r_2$, with interest rates strictly greater than r_2 being trivially suboptimal (since only the θ_b type accepts such rates).

The degree of securitization critically affects the lender's incentive to collect soft information. For example, suppose $\alpha = 1$, so that all loans are securitized. Consider borrowers with a hard information signal x_ℓ. Even if the lender can perfectly identify θ_b types via acquiring soft information, it will issue loans to these types at an interest r_1 or r_2 and sell these loans to investors. Since investors cannot be fooled in equilibrium, they will price loans assuming that the borrower pool includes θ_b types. Therefore, the lender will treat all borrowers with the same hard information equally, regardless of their soft information. But then, of course, it is sub-optimal to incur any cost to acquire soft information.

We explore two kinds of equilibria. An efficient soft information equilibrium delivers the second-best outcome, whereas a hard information equilibrium is one in which the lender relies exclusively on hard information.

Definition 1 (i) In an efficient soft information equilibrium, the lender collects soft information if and only if $x = x_\ell$, and adopts the following interest rate strategy: $\rho(r_1 \mid x, y) = 1$ if $x = x_h$ or $(x, y) = (x_\ell, y_h)$, and $\rho(r_2 \mid x, y) = 1$ if $(x, y) = (x_\ell, y_\ell)$, with no loan being offered if $x = x_b$ or $(x, y) = (x_\ell, y_b)$.
(ii) In a hard information equilibrium, the lender does not collect soft information, and adopts the following interest rate strategy: $\rho(r_1 \mid x_h) = \rho(r_2 \mid x_\ell) = 1$, with no loan being offered if $x = x_b$.

We show that, for the lender to collect soft information, the degree of securitization must be sufficiently low. In the proof of Proposition 1, we show that the soft information equilibrium is the unique equilibrium when there is no securitization, with $\alpha = 0$. Proofs of both propositions are in the Appendix.

Proposition 1 *There exists a level of securitization $\underline{\alpha} \in (0, 1)$ such that an efficient soft information equilibrium exists if and only $\alpha \leq \underline{\alpha}$.*

Next, we show that the hard information equilibrium obtains in a regime with a high degree of securitization. Essentially, collecting soft information represents a moral hazard problem for the lender. Unless there is a sufficiently high probability the lender will have to retain the loan, the moral hazard cannot be overcome. We show in the proof of Proposition 2 that the hard information equilibrium is the unique equilibrium when there is complete securitization, with $\alpha = 1$.

Proposition 2 *There exists a level of securitization $\bar{\alpha} \in (0, 1)$ such that a hard information equilibrium exists if and only if $\alpha \geq \bar{\alpha}$.*

Thus, if the degree of securitization is low, the lender collects soft information when the hard information signal is x_ℓ, and prices efficiently conditional on both hard and soft information. In this case, since there is a substantial probability the lender must retain the loan, we obtain the second-best outcome. However, when the degree of securitization is high, the moral hazard problem with respect to collecting soft information is too severe, and only hard information is obtained by the lender.

II.B Remarks on Model Features

We now comment on two features of our model: the exogenous degree of securitization, and the notion that the lender cannot select which loans to securitize. First, consider the degree of securitization, α. In practice, securitization offers several benefits to both lenders and investors that we do not model here. For lenders, as Bolton and Freixas (2000) point out, securitization frees up capital that can be used to make additional investments. If a bank holds a loan on its balance sheet, it is subject to minimum capital requirements, which must be met before it can expand lending. A lender wishing to increase its market share or its sales will thus find it attractive to securitize loans.[3] On the investor side, securitization increases opportunities for risk-sharing.

Although we focus only on one cost of securitization, the degree of securitization in the data was surely determined in equilibrium based on benefits and costs to lenders and investors. Incorporating the benefits of securitization to endogenize α in our model will not change the empirical predictions. Hence, for simplicity, we leave α as exogenous, and instead focus on the effects of changing α on incentives to collect soft information.

Next, suppose the lender in our model can choose which loans to offer for securitization. In each case, the hard information signal must still be communicated to the investors. However,

[3] Baumol (1958) argues that, under separation of ownership and control, firms have an incentive to maximize sales rather than profit.

the lender can condition its retention strategy based on the soft information signal.

For now, suppose all offered loans are securitized. If it is optimal for the lender to acquire soft information when $x = x_\ell$, it must adopt the following interest rate strategy. If the soft information signal is y_h, it offers interest rate r_1 and retains the loan. If the soft information signal is y_ℓ or y_b, it offers interest rate r_2 and sells the loan. The key is that the lender cannot commit to screen out borrowers with soft information signal y_b. Fixing the pricing strategy of the investors, it is always optimal for the lender to issue loans to borrowers with signals (x_ℓ, y_b), and sell these loans to investors.

The intuition of our model therefore goes through if the lender can selectively retain loans. If the securitization probability is sufficiently high, the lender will inevitably make loans to borrowers with signals (x_ℓ, y_b). This necessarily implies that the average quality of loans issued in a high securitization regime, given interest rate r_2, is worse than the average quality in a low securitization regime, even if the lender collects soft information in the former regime. Further, there exists some cost of acquiring soft information $\hat{c}$ at which the lender will not collect soft information, as long as the posterior probability of type θ_b is sufficiently high when $x = x_\ell$.

III Data

Our primary data set is obtained from a data vendor which provides a detailed perspective on the non-agency mortgage-backed securities market, and contains information on individual securitized loans. The data include information on issuers, broker dealers, deal underwriters, servicers, master servicers, bond and trust administrators, trustees, and other third parties. As of December 2006, more than 8,000 home equity and nonprime loan pools (over 7,000 active) that include 16.5 million loans (more than 7 million active) with over \$1.6 trillion in outstanding balances were included. Estimates from the data vendor suggest that as of 2006, the data covers over 90% of the subprime loans that are securitized.

We focus our analysis on subprime loans. As Mayer and Pence (2008) and Gerardi, et al. (2008) point out, there is no universally accepted definition of “subprime.”[4] Broadly, a borrower is classified as subprime if she has had a recent negative credit event. Occasionally, a lender signals a borrower with a good credit score is subprime, by charging higher than usual fees on a loan. In our data, the vendor identifies loans as subprime or Alt-A (thought to be less risky than subprime, but riskier than agency loans).

The subprime sector of the mortgage market provides an excellent test-bed for our predictions. Soft information about borrowers is likely to be especially valuable in this sector,

[4] Chomsisengphet and Pennington-Cross (2006) provide a history of the subprime market.

compared to either prime or Alt-A loans.[5] Gramlich (2007) shows that securitization in the subprime market grew rapidly after 2000. To provide a reasonable length of time with both low and high securitization, we consider loans issued in the period January 1997 to December 2006.

The data set includes all standard loan application variables such as the loan amount, term, loan-to-value (LTV) ratio, credit score, interest rate, and type of loan. We use both the credit score and the LTV ratio as hard information signals about the creditworthiness of the borrower.

A FICO score is a summary measure of the borrower's credit quality. These scores are calculated using various measures of credit history, such as types of credit in use and amount of outstanding debt, but do *not* include any information about a borrower's income or assets (Fishelson-Holstein, 2004). The software used to generate the score from individual credit reports is licensed by the Fair Isaac Corporation to the three major credit repositories – TransUnion, Experian, and Equifax. These repositories, in turn, sell FICO scores and credit reports to lenders and consumers. FICO scores provide a ranking of potential borrowers by the probability of having some negative credit event in the next two years. Probabilities are rescaled as whole numbers in a rang of 400–900, though nearly all scores are between 500 and 800, with a higher score implying a lower probability of a negative event. The negative credit events foreshadowed by the FICO score can be as small as one missed payment or as large as bankruptcy. Borrowers with lower scores are proportionally more likely to have all types of negative credit events than are borrowers with higher scores.

By design, therefore, a FICO score measures the probability of a negative credit event over a two-year horizon.[6] Holloway, MacDonald and Straka (1993) show that the ability FICO scores observed at loan origination to predict mortgage defaults falls by about 25 percent once one moves to a three-to-five year performance window. Thus, when we consider default models, we restrict attention to defaults that occur within 24 months of loan origination.

The loan-to-value ratio (LTV) of the loan, which measures the amount of the loan expressed as a percentage of the value of the home, also serves as a signal of borrower quality. Since the FICO score does not include information about the borrower's assets or income, the LTV ratio provides a proxy for the wealth of the borrower. Those who choose low LTV loans are likely to have greater wealth and hence are less likely to default.

Borrower quality can also be gauged by the level of documentation collected by the lender when taking the loan. The documents collected provide historical and current information about the income and assets of the borrower. Documentation in the market (and reported in

[5] Further, the coverage of prime and Alt-A loans in the data set is limited.

[6] Mortgage lenders should be interested in credit risk over a much longer period of time. The increasing usage of FICO scores in automated underwriting systems thus indicates that lenders have attained a level of comfort with their value in determining lifetime default probabilities.

the database) is categorized as full, limited or no documentation. Borrowers with full documentation provide verification of income as well as assets. Borrowers with limited documentation provide no information about their income but do provide some information about their assets. “No-documentation” borrowers provide no information about income or assets, which is a rare degree of screening leniency on the part of lenders. In our analysis, we combine limited- and no-documentation borrowers and call them low-documentation borrowers. Our results are unchanged if we remove the small proportion of loans which are no documentation.

Other variables include the type of the mortgage loan (fixed rate, adjustable rate, balloon or hybrid), and whether the loan is provided for the purchase of a principal residence, to refinance an existing loan, or to buy an additional property. We present results exclusively on loans for first-time home purchases. We do not report the results of our analysis on loans for refinancing since the nature of the results is qualitatively similar. We ignore loans on investment properties, which are more speculative in nature, and likely to come from wealthier borrowers. The zip code of the property associated with each loan is included in the data set. Finally, there is also information about the property being financed by the borrower, and the purpose of the loan. Most of the loans in our sample are for owner-occupied single-family residences, townhouses, or condominiums. Therefore, to ensure reasonable comparisons we restrict the loans in our sample to these groups. We also exclude non-conventional properties, such as those that are FHA or VA insured, pledged properties, and buy down mortgages.

IV Empirical Results

IV.A Descriptive Statistics

In Figure 1, we plot the percentage of new loans that have been securitized in the subprime sector (also known as the B&C loan market) since 1997. As can be observed, the percentage of loans securitized in this market grew steadily from about 30% in 1997 to almost 85% in 2006. As mentioned by Greenspan (2008), there was a surge in investor demand for securitized loans over this period. Due to an unprecedented budget surplus, the US Treasury had engaged in a buyback program for 30-year bonds in 2000–01, and had ceased to issue new 30-year bonds between August 2001 and February 2006.[7] Coincidentally, there was a rapid increase in CDO volume over this period, with a significant proportion containing subprime assets.[8]

[7] See, for example, “30-Year Treasury Bond Returns and Demand Is Strong,” the *New York Times*, Feb 9, 2006.

[8] The volume of CDOs issued in 2006 reached $386 billion, with home equity loans (largely from the subprime sector) providing for 26% of the underlying assets (from “Factbox - CDOs: ABS and other sundry collateral,” reuters.com, June 28, 2007).

We do not have information on the degree of securitization of individual lenders in the data. Therefore, we conduct our empirical analysis on aggregate data from the entire market. In going from our theoretical model with a single lender to implications for aggregate data, it is important to remember that lenders in this market were heterogeneous, and included commercial banks, thrifts, independent mortgage companies, and bank subsidiaries (see, for example, Gramlich, 2007). We expect that different lenders would cross over from a low to a high degree of securitization at different points of time. As a result, when aggregated across lenders, the data exhibit a steady increase in the degree of securitization over time. We therefore conduct many of our tests on a year-by-year basis, to examine whether incremental effects of increased securitization can be observed in the aggregate data. Broadly, we consider 1997–2000 to be a low securitization regime (with about 35% of subprime loans being securitized on average), and the period 2001 and later to involve high securitization (with about 70% of loans being securitized on average).

We report year-by-year summary statistics on our sample in Table I. The number of securitized subprime loans increases more than fourfold from 2001 to 2006. This pattern is similar to what is described by Demyanyk and Van Hemert (2007) and Gramlich (2007). The market has also witnessed an increase in the proportion of loans low (i.e., limited or no) documentation, from about 25% in 1997 to about 45% in 2006, which is consistent with a worsening quality of loans over time.

LTV ratios have gone up over time, as borrowers have put in less equity into their homes when financing loans. This increase is consistent with a better appetite of market participants to absorb risk. In fact, this is often considered the bright side of securitization – borrowers are able to obtain loans at better credit terms since the default risk is being borne by investors who can bear more risk than individual banks. The average FICO score of individuals who access the subprime market has been increasing over time, from 611 in 1997 to 636 in 2006. This increase in the average FICO score is consistent with a rule-of-thumb leading to a larger expansion of the market above the 620 threshold as documented in Keys et al. (2008). Though not reported in the table, average LTV ratios are lower and FICO scores higher for low-documentation loans, as compared to the full-documentation sample. This possibly reflects the additional uncertainty lenders have about the quality of low-documentation borrowers. The trends are similar for loan-to-value ratios and FICO scores in the two documentation groups.

IV.B Increased Reliance on Hard Information

Our first prediction is that, as the regime moves from the soft information to hard information equilibrium, there is increased reliance on hard information variables. We test this prediction in two ways.

IV.B.1 Regression of Interest Rate on FICO Score and LTV Ratio

In our model, borrowers with a high hard information signal ($x = x_h$ in the model) obtain the interest rate r_1 in both a low and a high securitization regime. Consider, however, borrowers with a low hard information signal ($x = x_\ell$ in the model). In a low securitization regime, the lender collects soft information which is unobserved by both investors and the econometrician. Thus, conditional on a low hard information signal, there will be a range of interest rates offered to borrowers, with a relatively weak link between the hard information and the interest rate. However, in a high securitization regime, soft information is not collected, and there is no diversity in interest rates. The FICO score and LTV ratio of a borrower represent the hard information signals in the data. As mentioned earlier, the FICO score is a direct and independent measure of a borrower's default probability. The LTV ratio also serves as a hard information signal about the creditworthiness of the borrower since borrowers with greater wealth (who are less likely to default) are more likely to choose loans with low LTV. Thus, when we consider a sample of mortgage loans that have been issued, the x_h signal naturally corresponds to high FICO scores and low LTV ratios, and the x_ℓ signal to low FICO scores and high LTV ratios.

To capture the reliance on hard information, we track the R^2 of a regression of interest rates on FICO scores and LTV ratios. The idea is that this measure should tell us how much of the variation in interest rates in a given year can be explained just by examining the variation in these variables. We can assess whether there has been increased reliance on FICO scores and LTV ratios by assessing how this measure changes over time. If lenders also use soft information to choose interest rates, the coefficients on FICO and LTV in this regression suffer from an omitted variable bias and the R^2 of the regression will be biased downward.

More concretely, we estimate the following regression for each loan i separately for every year in the sample:

$$r_i = \alpha + \beta_{FICO} \times FICO_i + \beta_{LTV} \times LTV_i + \epsilon_i, \tag{1}$$

where r_i is the interest rate on loan i, $FICO_i$ the FICO score of the borrower, LTV_i the LTV ratio on loan i, and ϵ_i an error term. This regression is estimated using both low-documentation and full-documentation loans.

We report β_{FICO}, β_{LTV} and R^2 in Panel B of Table II. Consistent with our first prediction, column 3 of the table shows that there is a drastic increase in the R^2 of this regression over the years. Starting from about 3% in 1997, the R^2 increases to almost 50% by the end of the sample, the increase being especially rapid after 2000. This is even more stark if one considers the sharp increase in the number of observations over the years. As expected, β_{FICO} is consistently negative (higher FICO scores obtain lower interest rates), and β_{LTV} is consistently positive (higher LTV ratios result in higher interest rates). In the low securitization regime, the hard information variables explain very little variation in interest rates suggesting that the omitted variables are particularly important in these years. As the securitization regime shifts, the same hard information variables explain a large amount of variation in interest rates indicating the declining importance of the omitted variables. Since soft information, by its very nature, is not observed by anyone but the lender, it is one of the omitted variables. Consequently, our results are consistent with the importance of soft information in determining interest rates on new loans declining with securitization.[9]

Since the LTV ratios observed during the sample period are lumpy (with masses at 80%, 90%, 95% and 100%), we also re-estimate the regression in different LTV buckets using only the FICO score as the dependent variable. The R^2 of such a regression using loans with LTV between 50% and 95% (about 80% of the sample) improves from 1% in 1997 to about 20% in 2006, suggesting that the FICO score by itself is an important hard information variable. For robustness, we also add dummy variables that capture the documentation level of the loan (low or full) and loan type (ARM/FRM) to equation (1), which yields qualitatively similar results. In this specification (unreported), the R^2 improves from about 5% in 1997 to about 57% in 2006. Since interest rates will also depend on macro-economic factors, we re-estimated equation (1) adding the average monthly yield on the nominal 10-year Treasury note as an independent variable. There is no quantitative difference in the R^2 results.[10]

IV.B.2 Shrinkage of the Distribution of Interest Rates

Another way to test the relationship between hard information and interest rates is to consider the dispersion of interest rates given the hard information signal. In the model, in the low securitization regime, borrowers who generate hard information signal x_ℓ eventually obtain rates of interest r_1 or r_2, depending on the soft information signal they generate. Under high securitization, since soft information is not acquired, all borrowers with signal x_ℓ obtain the interest rate r_2. Thus, the dispersion of interest rates with a low hard information signal is reduced under high securitization. Note that borrowers who generate a high hard information signal obtain interest rate r_1 under both securitization regimes. As mentioned above, LTV ra-

[9] An increased dependence of interest rates on FICO scores over time may also be due to credit scores becoming more accurate at evaluating creditworthiness as more data about subprime borrowers become available. However, an improvement in the informational quality of the FICO score does not imply an increase in prediction errors from the statistical default model. Moreover, as we show in Section IV.C.1, the prediction errors are positive even when the baseline default model is estimated over a period in which the subprime market had matured and a longer credit history about borrowers had become available.

[10] During the sample period, there were some bank mergers. As banks become large, interest rates will depend more on hard information, due to the effects identified by Stein (2002). To rule out this explanation, we re-estimated equation (1) only for banks that did not engage in mergers over the sample period. Our results remain.

tios in our sample are lumpy so our tests in this section focus on the FICO score as the hard information variable. Our prediction, therefore, is that the variance of interest rates offered on newly originated mortgages should fall for low FICO scores, while it should remain relatively unaffected for higher FICO scores.

To test this, we calculate the standard deviation of interest rates at each FICO score and track it over time. Specifically, we first compute $\sigma_{it} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(r_{ijt} - \bar{r}_{it})^2}$, where r_{ijt} is the interest rate on the j^{th} loan with FICO score i in year t, and $\bar{r}_{ijt} = \frac{1}{N}\sum_{j=1}^{N} r_{ijt}$ is the mean interest rate. Next, we pool observations into FICO score buckets of 10 points starting from a score of 500 and ending at 800 (i.e., the buckets are FICO scores 500-509, 510-519,...). We then estimate the following regression separately for each bucket b:

$$\sigma_{bt} = \alpha_b + \beta_b \times t + \epsilon_{bt}, \qquad (2)$$

where t indexes year and ϵ_{bt} is an error term. The coefficient β_b provides a sense of how the dispersion of interest rates within each FICO score bucket changes over time, and thus allows us to examine how interest rates are dispersed across the FICO score spectrum as the securitization moves from a low regime to a high one. We expect β_b to be large and negative for low FICO scores, i.e., we expect a shrinkage of dispersion in contracts at low FICO scores. Again, we use both low-documentation and full-documentation loans in our test.

The results of this estimation are displayed in Table III. As can be observed, loans at lower FICO scores – from 500 to 599 – see a reduction of about 0.15 per year in the dispersion of interest rates relative to the mean interest rate at that score. In contrast, loans at FICO scores in the higher range (600 and above) see a corresponding reduction of about 0.05 per year. Relative to the average standard deviation of interest rates across years, this translates to about 6.8% shrinkage at lower FICO scores (average standard deviation = 2.2) and about 2.5% shrinkage per year at higher FICO scores (average standard deviation 2). The magnitude of shrinkage can also be interpreted relative to the mean interest rate. Across sample years, the mean interest rate is 9.2% at FICO scores 500–599 and 8.1% at FICO scores 600 and higher. Thus, scaling the degree of shrinkage by the mean interest rate yields the same results.

While the loan-to-value ratio is also a hard information signal, it is cumbersome to conduct and report our results for each (FICO, LTV) pair. To condition for LTV ratios, we conduct two additional tests. First, we re-estimate equation (2) separately at LTV ratios of 80%, 90% and 95%, which together represent about 70% of the observations. The results for high LTV ratios (i.e., LTV of 90% and 95%) are qualitatively similar to those reported in Table III. This is consistent with more shrinkage occurring where soft information about the borrowers is important, i.e., low FICO scores and high LTVs. Second, we also include the dispersion of loan-to-value ratio (calculated at each FICO score in a similar manner to the dispersion for the interest rate variable) and again find similar results.

During our sample period, there was some standardization of terms of mortgage loans for transparency reasons (see Kroszner, 2007). Standardization itself may imply that there is a shrinkage in the distribution of observed interest rates over time. However, note that we find not just a shrinkage in the interest rate distribution, but also that the shrinkage occurs in significantly greater amounts for borrowers at low FICO scores and high LTV ratios. This cannot be explained by the standardization of contractual terms unless the loan terms were already standardized at high FICO scores by 1997.

Nevertheless, we conduct an additional test to alleviate this concern. The idea is that transparency considerations should shrink other dimensions of contractual terms simultaneously. We extend equation (2) to condition for shrinkage in the dispersion of not just the loan-to-value ratio, but also other contractual terms including ARMs, prepayment penalty etc. at each FICO score in each year. The shrinkage for each of these variables is constructed in a similar manner to the shrinkage for the interest rate variable. The results of this estimation are visually presented in Figure 2 where we report β_b and its 95% confidence interval. As can be observed, our results are robust even after conditioning for simultaneous shrinkage in the dispersion of other contractual features.[11] Overall, the evidence that there is greater shrinkage in the distribution of interest rates at lower FICO scores in the high securitization regime is consistent with our second prediction.

IV.C Failing to Predict Failure

Finally, consider the effect of securitization on mortgage defaults. In the model, in the low securitization regime, the lender acquires soft information when the hard information signal is x_ℓ. As a result, borrowers who generate the bad soft information signal y_b are screened out. Given the signal structure, the pool of borrowers that are screened out contains a disproportionate number of bad (θ_b) types. Thus, the pool of borrowers who are offered loans contains a disproportionately small number of bad types. Since type directly corresponds to default probability in the model, defaults are expected to be low under low securitization.

Under high securitization, the lender does not collect any soft information. Thus, no further screening occurs among borrowers who generate hard information signal x_ℓ. That is, the pool of borrowers who obtain loans given x_ℓ contains a greater proportion of bad types (who have the highest default rate) under high securitization. Therefore, any default model estimated from the low securitization regime will underpredict defaults under high securitization for borrowers with low FICO scores and high LTV ratios. However, at high FICO scores and low LTV ratios (i.e.,

[11] We also estimate the regressions separately for low-documentation and full-documentation loans and find similar results for both sets of loans.

when the hard information signal is x_h), the set of borrowers obtaining loans is the same under both securitization regimes. Thus, at high FICO scores and low LTV ratios, a default model estimated under low securitization should retain its predictability under high securitization.

IV.C.1 Main Test

We first estimate a simple default model in a base or test period with a low degree of securitization. The model we estimate is similar in philosophy to the S&P LEVELS® model (Standard & Poor's, 2007). We consider the period 1997 to 2000 to be a low securitization era. We then fix the model coefficients and examine the prediction errors from the model during the high securitization regime.

For any loan i issued in the period 1997 to 2000, let X_i denote the vector of explanatory variables. This vector includes a constant, the FICO score of the borrower ($FICO_i$), the LTV ratio (LTV_i), the interest rate on the loan r_i, a dummy variable ARM_i that takes value 1 if the loan is an adjustable rate mortgage, a dummy variable FRM_i that takes value 1 if the loan is a fixed rate mortgage (the third type of mortgage, hybrid mortgage, represents the basic specification when $ARM_i = FRM_i = 0$), and a dummy variable $Prepay_i$ if the loan has a prepayment penalty. Finally, let I_i^{Low} be a dummy variable that takes value 1 if loan i has low documentation (i.e., no documentation or limited documentation) and 0 if the loan has full documentation.

We estimate the following logit model on loans issued in the period 1997 to 2000:

$$\text{Prob}(\text{Default}_i = 1) \ = \ \Phi(\beta \cdot X_i + \beta^{Low} \cdot I_i^{Low} X_i), \qquad (3)$$

where $\Phi(\cdot)$ is the logistic distribution function. Here, β and β^{Low} are vectors of coefficients, one for each explanatory variable included in X_i. Notice that we effectively estimate separate models for full- and low-documentation loans. The extent of hard information available to a lender differs fundamentally across full- and low-documentation loans, since on the former the lender has documentation about the borrower's job, income and assets. We choose a flexible specification to allow the effect of all explanatory variables to vary across the two kinds of loans.

Panel A of Table IV shows the estimates from the baseline model. A high credit score and low interest rate are both associated with lowering the probability that the borrower will default in the subsequent two years, for both full-documentation and low-documentation loans. Holding all else constant, the marginal effect of LTV ratios on defaults is small for both kinds of loans. The marginal effect is positive for full-documentation loans, as expected. However, for low-documentation loans, the marginal effect is negative. There are at least two possible explanations for this. First, some of the effect of changing LTV are captured by the interest rate r (as shown in Section IV.B.1, the interest rate increases with LTV). Second, since the baseline model considers only the low securitization period, we expect low-documentation borrowers with high LTV ratios to be screened more intensively on soft information, compared to full-documentation borrowers.

Next, we use the coefficients of the baseline model to predict the probability of default on a loan in the two subsequent years following loan origination, for loans issued from 2001 to 2006. Concretely, let $\hat{\beta}_{1,t}$ and $\hat{\beta}_{1,t}^{Low}$ be the coefficients estimated from equation (3) for the baseline model over the period 1 to t (where year 1 is 1997 and year t is 2000). Then, for $k = 1, 2, \cdots, 6$, we estimate the predicted probability that a loan i issued at $t + k$ will default in the next 24 months (keeping the baseline coefficients fixed) as $\textit{Predicted Default}_{i,t+k} \equiv \text{Prob}(\widehat{\text{Default}}_{i,t+k} = 1)$, where:

$$\text{Prob}(\widehat{\text{Default}}_{i,t+k} = 1) = \Phi(\hat{\beta}_{1,t} \cdot X_{i,t+k} + \hat{\beta}_{1,t}^{Low} \cdot I_{i,t+k}^{Low} X_{i,t+k}).$$

We then examine the actual default experience of loans issued in each of years 2001 to 2006, assigning $\textit{Actual Default}_{i,t+k} = 1$ if loan i issued in year $t+k$ defaults within 24 months of issue, and zero otherwise.[12] The prediction error is computed as $\textit{Error}_{i,t+k} = \textit{Actual Default}_{i,t+k} - \textit{Predicted Default}_{i,t+k}$. If the model indeed underpredicts defaults in the high securitization era, the prediction error should be positive on average. Further, if there is systematic underprediction at low FICO scores and high LTV ratios, the prediction error should decline in magnitude as the FICO score increases and LTV ratio falls.

We estimate yearly the regression for borrower i in year $t + k$ (where $t = 2000$ and $k = 1, 2, \cdots, 6$) as follows:

$$\textit{Error}_{i,t+k} = \alpha + \beta_{FICO} \times \textit{FICO}_{i,t+k} + \beta_{LTV} \times \textit{LTV}_{i,t+k}.$$

Panel B of Table IV reports the coefficients on the FICO scores and LTV ratio for loans issued in each of the years 2001 to 2006. As can be observed from columns 1 and 2, the β_{FICO} is negative while β_{LTV} is positive and significant across 2001 to 2006. The magnitudes seem large. For instance, a 1 standard deviation increase in the FICO score (about 70 points) leads to a reduction in the prediction error of about 33.5% for 2006 loans. Similarly, a 1 standard deviation increase in LTV ratio (about 10%) leads to a reduction in prediction error of about 9.4% for 2006 loans.

To gauge whether there is indeed underprediction by the baseline model, we need to examine whether the prediction errors are positive on average. As shown in column 5 of Panel B, this is indeed the case in each year. Further, the average prediction error increases over time as securitization increases, implying that the fit of the baseline model worsens over time.[13]

[12] We have data through May 2008, so for loans issued after May 2006, the *Actual Default* variable is based on a window less than 24 months.

[13] As another indicator of a worsening fit, in unreported tests, we also examined how well predicted defaults

Moreover, the magnitudes of the prediction errors are large relative to actual defaults (reported in the last column). For instance, among loans of 2004 vintage, the mean prediction error of 7.8% reflects an underprediction of about 55% on actual defaults of 13.9%. Together, our findings in Panel B suggest that the prediction errors are positive and are higher for low FICO scores and high LTV ratios.

As a confirmation that prediction errors are positive, we plot the Epanechnikov kernel density of mean prediction errors over time.[14] As is clear from Figure 3, the distributions show that, on average, the mean prediction error has been positive across years. If the predictions of the default model are correct on average, we expect the distributions of the prediction errors to be centered around zero. However, as seen from the figure, there are very few observations with negative mean prediction errors. Further, the distribution of the mean prediction error progressively shifts to the right over time, as securitization becomes more prevalent in the subprime market.

Our test above estimates the coefficients of the model in the window 1997 to 2000, and considers the prediction errors in the period 2001 to 2006. As seen from Figure 1, there was a steady increase in securitization over the latter period. Hence, an alternative way to conduct this test is to use as much data as available for each year to tease out the incremental effect of additional securitization on the prediction errors of a default model. Using a rolling window, we predict defaults for loans issued in years 2005 and 2006, which allows the baseline model to include a few years of data from the high securitization regime. Thus, we expect the prediction errors to be smaller. For 2005 loans, the baseline model is estimated over the period 1997 to 2004, and for 2006 loans the base period is 1997 to 2005. The results, shown in Figure 4, are qualitatively similar. The average prediction error in this specification is 8.3% for 2005 loans (compared to 14.7% in the baseline specification) and 15.1% for 2006 loans (compared to 25.5% in the baseline specification). Thus, while the magnitude of the prediction errors falls, the default model continues to underperform in the high securitization regime even with a rolling window adjustment.

on 2001–2006 loans explained actual defaults in a logistic regression. As compared to the pseudo R^2 of 7% for the baseline model (over 1997–2000) the pseudo R^2 of the regression of actual defaults on predicted defaults falls steadily from about 3% in 2001 and 1% in 2006.

[14] Plotting each of the error data points results in a dense figure with a large file size. To ensure manageable file sizes, all the kernel density figures in the paper are constructed as follows. For each year, at each FICO score, we determine the mean prediction error. We then plot the kernel density using the mean errors at each FICO score. We also plotted the densities weighing the errors by the actual number of loans at each FICO score. The plots look similar.

IV.C.2 Confirmatory Test: Low- and Full-Documentation Loans

Fixing a FICO score and LTV ratio, soft information should be more important for low-documentation loans. Thus, all else equal, a default model fitted during a low securitization era should perform better (in terms of default predictions in the high securitization period) on full-documentation loans compared to low-documentation loans. Importantly, the distribution of full- and low-documentation loans across zip codes is similar. To check this, we sorted the volume of each kind of loan by zip code over 2001–2006, and considered the top 25% of zip codes in each case (which contribute over 60% of the volume of each kind of loan). A large proportion of zip codes (about 82%) are common across the two lists. In Figure 5, we plot the top 25% of zip codes for each kind of loan. As can be seen, there is substantial overlap across the two kinds of loans. Thus, under the assumption that low- and full-documentation borrowers are equally sensitive to changes in the economy, any differential effects across the two kinds of loans are insulated from macroeconomic and zip-code level shocks such as unemployment and changes in house prices.[15]

To evaluate how prediction errors vary across the two kinds of loans, we use a rolling window specification and fit separate baseline models for full- and low-documentation loans. That is, for predicting default probabilities on loans issued in year $t+1$, the baseline model is estimated over years 1 through t, where year 1 is 1997. For each kind of loan $s = Low, Full$, the baseline specification is a logit model of the form

$$\text{Prob}(\text{Default}_i^s = 1) \; = \; \Phi(\beta_{1,t}^s \cdot X_i^s),$$

where the vector X_i is the same as described in Section IV.C.1. Let $\hat{\beta}_{1,t}^s$ be the estimated coefficients from this regression. The predicted default probability for loans issued in year $t+1$ is then estimated as

$$\text{Prob}(\widehat{\text{Default}}_{i,t+1}^s = 1) = \Phi(\hat{\beta}_{1,t}^s \cdot X_{i,t+1}^s),$$

We report the mean prediction errors for full- and low-documentation loans in Table V. As seen from the table, the mean errors are substantially higher among low-documentation loans for loans issued in 2003 and later. For 2001 and 2002 loans, the mean prediction errors are not significantly different across the two kinds of loans. Figures 6 (a) and (b) plot the Epanechnikov kernel density of mean prediction errors at each FICO score over time separately for full and low-documentation loans, and confirm the same observation. The plots also suggest that, for full-documentation loans, the relationship between model errors across time is weaker than for low-documentation loans.

[15] In Section V, we explicitly consider the role of changing house prices on default predictions more generally, and on predictions for low- and full-documentation loans in particular.

IV.C.3 Control Test: Low Securitization Regime

Across different years in the low securitization regime, there should be no substantive change in a lender's incentives to collect soft information. Thus, our hypothesis is that the quality of loans issued during the low securitization years will be approximately similar from year to year. Therefore, as a placebo test, we assess whether a default model estimated during low securitization regime predicts defaults reasonably in a period with relatively *low* securitization.

To conduct the test, we predict defaults on low-documentation loans issued in 1999 and 2000, using a baseline model estimated from 1997 and 1998 for 1999 loans, and 1997 through 1999 for 2000 loans (i.e., employing a rolling window). We then regress the prediction errors on FICO score and LTV ratio for each year 1999 and 2000. The results are reported in Table VI. As can be observed, in contrast to the results in Table IV, the β_{FICO} and β_{LTV} coefficients are insignificant suggesting that there is no systematic underprediction by the baseline model. The mean prediction error is not significantly different from zero, and is also substantially smaller in magnitude than the mean errors reported in Table V for years 2001 and beyond. The same result is confirmed in Figure 7, where we plot the kernel distribution of the mean prediction error at each FICO score. As can be observed, in contrast to Figure 3, the mean errors are centered around 0. Thus, the control test is consistent with our predictions.

As mentioned earlier, the market for full-documentation loans evolved more quickly than the market for low-documentation loans. By 2000, for example, there are about 90,000 full-documentation loans and only 25,000 low documentation loans in our sample. Thus, we are confident that the years 1997 through 2000 represent a period of low securitization for low-documentation loans, even though the market for full-documentation loans may have been more advanced. Nevertheless, we repeated the control test on full-documentation loans. The mean prediction error is approximately zero for 1999 loans, and 1.8% for 2000 loans. Although the latter is significantly different from zero, the mean error is nevertheless substantially smaller than the means in later years.

V Robustness

We now consider the robustness of our findings by evaluating the role of a few other explanations for the increase in defaults.

Falling house prices

There is no doubt that falling house prices are partly responsible for the surge in defaults for loans issued in 2005 and 2006. However, only in August 2007 did the composite (i.e., national level) Case-Shiller index indicate a fall from its value 24 months earlier. As a result, loans issued in 2004 and before did not suffer from a fall in house prices over the next 24 months, yet as shown in Table IV and Figure 3, the prediction errors from a default model remain high.[16] Further, in our comparison between full- and low-documentation loans, both are subject to the same effects of changing house price, since the distribution of both kinds of loans across zip codes is similar. Finally, in this section, later in this section, we explicitly include the future change in house prices at the state level as an explanatory variable.

As a more direct test to soak up the effects of falling house prices on defaults, we explicitly include a house price appreciation variable in the statistical default model. For each loan, we construct the house price appreciation (HPA) variable as follows. We begin with the state-level quarterly house price index constructed by the Office of Federal Housing Enterprise Oversight. For each state s, a house price index for each year t, $h_{s,t}$, is constructed as a simple average of the indices over four quarters (for 2008, only three quarters are used). Consider loan i issued in state s in year t. The house price appreciation variable for loan i is set to the growth rate of house prices over the next two years, $HPA_i = \frac{h_{s,t+2} - h_{s,t}}{h_{s,t}}$. We include HPA_i in the vector of loan characteristics X_i in both the baseline and predictive regressions.

Our specification is stringent: This is clearly more information than available to an econometrician at the time the forecast is made and will soak up more variation in defaults than a prediction made in real time. Gerardi, et al. suggest that market participants could have anticipated how sensitive foreclosures were to market prices, but not the change in home prices. Since we directly assume knowledge of the future path of house prices, we circumvent the issue of participants' beliefs.

We predict default probabilities for loans issued in each of the years 2001 through 2006 using a rolling window specification after including the HPA variable (both by itself and interacted with I^{Low}, the low-documentation dummy) on the right-hand side. The predicted default probabilities for loans issued in year $t + 1$ are based on coefficients estimated over years 1 through t, where year 1 is 1997. In Figure 8, we plot the Epanechnikov kernel density of mean prediction errors (computed at each FICO score) in each year 2001 through 2006. For ease of comparison, the figure has six panels, each panel showing the kernel density of mean out-of-sample prediction errors in a given year with and without including house price appreciation as an explanatory variable, using a rolling estimation window in each case.

[16] There are two possible explanations for borrowers defaulting when house prices increase. First, over 70% of the loans in our sample have a prepayment penalty, increasing the transaction cost to a borrower of selling the house. Second, some borrowers who experience an increase in home prices may be taking out additional home equity loans, effectively maintaining a higher LTV ratio than reported in the sample. The latter effect is consistent with our channel of hard versus soft information, since soft information includes the likelihood that a borrower will be credit-constrained in the future and will take out additional home loans.

Two observations emerge from the figure. First, for 2001–2004 loans, there is not much difference in the two kernel densities. In fact, for 2002–2003 loans, including the house price effect slightly magnifies the prediction errors. Second, the prediction errors for loans issued in 2005 and 2006 are indeed reduced in magnitude when the effect of house prices is included. In particular, using a rolling window for estimating the baseline model, the mean prediction error for 2005 loans falls from 8.3% to 4.9% when HPA is included as an explanatory variable, and for 2006 loans falls from 15.1% to 6.1%. Thus, for these two years, approximately 50% of the mean prediction error survives over and above the effect of falling house prices. Therefore, falling house prices account for a significant proportion of defaults on loans issued in 2005 and 2006. However, even after accounting fully for the effect of falling house prices on defaults, the prediction errors exhibit the patterns predicted by our theoretical model.

In unreported tests, we repeat the analysis separately on low- and full-documentation loans, to account for the possibility that low-documentation borrowers are more sensitive to economic downturns. We confirm that for loans issued in 2001–2004, the results are similar to those reported in Section IV.C.2. For loans in 2005 and 2006, the magnitudes of the prediction errors are reduced for both low- and full-documentation loans when house price changes are taken into account, but the errors continue to be larger for low-documentation loans.

Other alternatives

One benefit of securitization is a lower cost of capital for the lender. As the cost of capital falls, some risky borrowers who represent negative NPV projects at the higher cost of capital now become positive NPV projects. Thus, given a set borrowers with the same FICO score, the lender will naturally make loans to more risky borrowers at the lower cost of capital. Therefore, as securitization increases, the quality of loans issued will worsen, leading to positive prediction errors from a statistical default model.

However, the lower cost of capital channel has a very different prediction on the interest rate distribution, compared to our channel of loss of soft information. Even at a lower cost of capital, there should be a difference in the interest rates charged to a more risky and a less risky borrower. Thus, over time, if the pool of issued loans includes borrowers with greater risk, the dispersion of interest rates at a given FICO score should increase. However, as shown in Section IV.B.2, the dispersion of interest rates falls as securitization increases, especially at low FICO scores. This pattern is consistent with a loss of soft information for low hard information signals, but not the riskier borrowers channel.[17]

[17] Additional evidence against the cost of capital channel is provided by Keys, et al. (2008), who conduct a cross-sectional test using similar data, and show that defaults on a portfolio that is more likely to be securitized exceed defaults on a portfolio that has similar risk characteristics but is less likely to be securitized. Their test

Another possible explanation for our results is that borrowers may be able to manipulate FICO scores. As Mayer and Pence (2008) point out, there is no evidence to this effect. Nevertheless, suppose such manipulation were possible. Then, since borrowers will only manipulate their FICO scores upward, it is conceivable that in later years the quality of borrowers at a given FICO score is lower than the corresponding quality at the same FICO score in the earlier years. In such a situation, again a statistical default model will underpredict defaults in later years.

For the manipulation channel to explain our results on defaults, it must be the case that, first, manipulation of FICO scores increases with securitization. Second, it must be that, in the high securitization years, a greater proportion of low FICO scores have been manipulated upwards, compared to high FICO scores. Third, a greater degree of manipulation must have occurred among borrowers who accept low-documentation loans, compared to those who accept full documentation loans.

To investigate these conjectures, we rely on another dataset of subprime loans that continues to track the FICO scores of borrowers after loan origination. Borrowers who manipulate their FICO scores before loan issuance should experience a fall in FICO score shortly after receiving a loan (since a permanent change in the credit score cannot be manipulation). We find that borrowers at high FICO scores are more likely to experience such a reduction within six months and within one year of obtaining a loan, an opposite effect to what is predicted by the manipulation channel. In addition, there are no differences on this dimension between borrowers who obtain low- and full-documentation loans.[18]

Even accepting these caveats, if manipulation of FICO scores was indeed commonplace, it is hard to imagine that lenders were unaware of such manipulation. At any point of time, a borrower's credit report provides a credit history for the borrower over the past few months. Thus, a careful perusal of the credit report may uncover the likelihood of FICO score manipulation. Whether a borrower is likely to have manipulated the FICO score is therefore soft information, and can be uncovered by a lender by incurring costly effort. A failure to incur that effort is therefore consistent with our channel based on loss of soft information at loan issuance.[19]

to rule out the cost of capital channel also involves the dispersion of the interest rate distribution.

[18] For brevity, we do not report the details of these tests in the paper. The dataset covers loans serviced by the top ten subprime mortgage servicers in the U.S., who account for over 60% of loans in this market.

[19] A similar argument applies if the reported value of a house is manipulated upwards by a borrower or appraiser, resulting in a reported LTV ratio that underestimates the true ratio of loan to value.

VI Discussion and Conclusion

VI.A Connections with Literature

The notion in our model that distance increases the reliance on hard information since soft information cannot be contracted on outside the firm may be viewed as an extension of the work of Stein (2002). As Stein demonstrates, the inability to communicate soft information in a hierarchical firm results in line managers losing the incentive to acquire soft information about projects. Thus, even within the firm, distance leads to a greater reliance on hard information. Our result can also be broadly viewed in a multi-tasking framework. Consider the lender to be an agent with two tasks: acquiring hard information, and acquiring soft information. As the reward to acquiring soft information decreases, the agent will naturally spend less time or effort on that task.[20]

Since soft information, by definition, is unobservable to econometricians, empirical tests of Stein's model have been indirect. For instance, Berger, et al. (2005), analyze a data set on small business lending, and find that large banks lend at a greater geographic distance than small banks, and interact with borrowers in more impersonal ways. Similarly, Petersen and Rajan (2002) find that, over time, the distance between banks and small business borrowers has been increasing, in part because hard information about borrowers is more readily available. Relatedly, Cole, Goldberg and White (1998) and Liberti and Mian (2008) find that loan approvals by large banks and at higher levels within a bank are more sensitive to financial statement variables. In our work, we are able to directly determine the extent to which interest rates on loans rely on hard information about borrowers. By inference, an increasing reliance on hard information implies a decreasing reliance on the unobserved soft information. In our tests on default models, we control for all hard information about the borrower that is available to investors. Thus, the errors in predictions from these models may be traced to unobserved soft information.

In related work, Einav, Jenkins and Levin (2008) consider subprime auto loans, and show that the profitability of dealerships at a lender increases when they improve the use of hard information about borrowers. They find that the increased profits come both from eliminating loans to the riskiest borrowers and from superior credit terms provided to the safest borrowers, which is similar to the benefits of screening that we consider in our model.

We show that increased reliance on hard information in the subprime mortgage sector has led to a shrinkage of the dispersion of interest rates on new loans. Such a shrinkage is equivalent

[20] Inderst (2008) considers a multi-tasking model in which a loan officer is responsible for both generating new loans and for acquiring soft information, and shows that competition between banks leads to a regime in which the reliance on hard information increases.

to standardization of prices. Some shrinkage in dispersion of prices, of course, will occur simply due to standardization in the features of the contract, driven by transparency and liquidity considerations (see Kroszner, 2007). Our results control for these considerations, and are thus able to identify an additional channel that reduces dispersion in interest rates. This additional channel directly relates to the effects of a regime change in the determination of interest rates on new loans.

Our theoretical model builds on the literature on loan sales. We combine the insights of Gorton and Pennacchi (1995), who demonstrate that the level of credit screening falls as a bank's share of the loan falls, and Parlour and Plantin (2008), who consider the effects of adverse selection at the stage of the sale.[21] A key distinction is our emphasis on hard versus soft information, with the moral hazard (and possible private information about loan quality) being associated with soft information. We do not formally model the lender's incentive to securitize loans. On this question, Pennacchi (1988) shows that loan sales increase when a bank's internal funding costs rise and the bank attempts to free up capital.

More broadly, the paper argues that any attempt by banks to remove loans from their balance sheets (whether by selling them outright, creating a Special Investment Vehicle, or via creative off-balance sheet financing) will result in a similar incentive problem in terms of collecting soft information, and give rise to a similar regime shift in the quality of loans issued. Since capital is scarce, we expect continual financial innovation that strives to free up bank capital. Thus, in any market in which soft information is valuable, the past will not be a reliable indicator of the future. Consider, for example, the growing secondary market for bank-issued corporate loans. If the borrower is a publicly traded firm with a good credit rating, soft information is likely of relatively little value, since the rest of the financial market is constantly producing information about the firm. On the other hand, if the firm is private and has a low credit rating, soft information is more valuable. In such a situation, securitization of the bank loan inevitably leads to an incentive problem with respect to acquiring soft information.[22]

There is, of course, a growing literature on the subprime crisis.[23] As Mayer, Pence and Sherlund (2008) point out, falling house prices have also played a role in the increase in subprime mortgage defaults. Gerardi, et al. (2008) find that the sensitivity of foreclosures to home prices

[21] Gorton and Souleles (2008) point out that the adverse selection problem faced by an investor in a Special Purpose Vehicle is mitigated in a repeated game. Duffee and Zhou (2001) show that if a bank uses credit default swaps the adverse selection problem in loan sales is exacerbated.

[22] Drucker and Puri (2008) demonstrate that sold loans in the secondary loan market contain more restrictive covenants than unsold ones, which is consistent with lenders facing an incentive problem with respect to sold loans.

[23] See, for example, Dell'Ariccia, Igan and Laeven (2008), Demyanyk and Van Hemert, (2008), Doms, Furlong and Krainer (2007), Gerardi, et al. (2008), Keys, et al. (2008), Mayer, Pence and Sherlund (2008), Mian and Sufi (2008) and Purnanandam (2008).

was predictable, but analyst reports suggest that participants believed that a fall in house prices was a low probability event. In contrast, our findings suggest that the failure of default models occurs vintage by vintage and especially at low FICO scores and for low-documentation loans, whereas house prices presumably fell across the creditworthiness spectrum and for full-documentation loans as well.

VI.B Implications of Our Results

Establishing a liquid market for a complicated security requires standardization of not just the terms of the security, but also of the fundamental valuation model for the security, both of which help investors to better understand the security. Inevitably, the process of constructing and validating a model will include testing it against previous data. We argue in this paper that the growth of the secondary market for a security can have an important incentive effect that affects the quality of the collateral behind the security itself. The associated regime change will imply that even a model that fits historical data well will necessarily fail to predict cash flows, and hence values, going forward.

While we focus on a particular statistical default model, similar models are widely used by market participants for diverse purposes such as making loans to consumers (for example, using the FICO score), assessing capital requirements on lenders and determining the ratings of CDO tranches. Our critique applies to all such models, since they all use historical data in some manner to predict future defaults. Importantly, the effects we document are systematic and stronger for borrowers with low FICO scores and low-documentation. The magnitude of the prediction errors is large even after controlling for falling house prices, especially for loans issued in 2005 and 2006. The effects are reasonable since our notion of soft information is broad, and includes any information related to default that is not easily documentable or verifiable by a third party. It is plausible that the magnitudes of the prediction errors owe in part to interaction between securitization and house prices, an effect we do not explicitly consider in our work.

It is worth emphasizing that in our theoretical model, investors are fully rational, and price loans fully taking into account that the loan quality depends on the securitization regime. Nevertheless, we show in the data that a mechanical prediction that uses a regression model from a low securitization regime will systematically underpredict loan defaults in a high securitization regime, especially in the set of loans where soft information is important. Thus, regulators and investors using such a statistical model to price loans would be caught by surprise. For example, in November 2007, Standard and Poor's adjusted their default model to reduce the reliance on the FICO score as a predictor of default (Standard & Poor's, 2007). This is consistent with default models used in industry failing to compensate for loss of soft information in issuing loans.

Of course, it is difficult to establish whether market participants rationally anticipated an increase in defaults. As an indirect test, we consider the subordination levels of AAA tranches for new non-agency pools consisting of loans originated in 2005 and 2006. We have already shown (Figures 3 and 8) that a statistical default model most severely underestimates actual defaults in 2005 and 2006. The subordination level measures the magnitude of losses an equity tranche can absorb, before the principal of the AAA tranches is at risk. Thus, if rating agencies were correctly forecasting future defaults, the subordination levels in the pools must have a positive correlation with the prediction errors of the default model (otherwise the tranches should not have been rated AAA). Figures 9 (a) and (b) show the subordination level plotted against the mean prediction error of the pool. As is evident, the relationship is weak at best, suggesting that rating agencies were unaware of or chose to overlook the underlying regime change in the quality of loans issued as securitization increased. These results are consistent with the suggestions of Benmelech and Dlugosz (2008) and Griffin and Tang (2008), who argue that ratings of CDO tranches were aggressive relative to realistic forward-looking scenarios.

We highlight a dimension of model risk (i.e., the risk of having an incorrect model) that cannot be corrected by mere application of statistical technique. While model risk is often recognized as an important phenomenon, the term is often understood to mean an incomplete set of data, (conceptual) errors in a statistical model, or both and as a result the focus in the literature has been on testing the consistency and robustness of inputs that go into these models. Collecting more historical data, possibly on extreme (and possibly rare) values of inputs, is one of the key corrections that is frequently suggested. However, when incentive effects lead to a change in the underlying regime, the coefficients from a statistical model estimated on past data have no validity going forward. This holds regardless of how sophisticated the model is, or how well it fits the prior data. Importantly, collecting historical data over a longer time period is likely to exacerbate the problem by aggregating data from different regimes.

The inescapable conclusion of a Lucas critique is that actions of market participants will undermine any rigid regulation. This has a direct implication for the Basel II guidelines which assign risk to asset classes relying in part on probability of default models (either models by "external credit assessment institutions," i.e., rating agencies, or internal bank models; see, for example, Basel Committee on Banking Supervision, 2006). Recent policy discussions have focused on the role of capital requirements in the subprime crisis.[24] We contribute to this debate by highlighting the role of incentives in determining the riskiness of loans, and in turn affecting the performance of models used to determine capital requirements. Our findings suggest that a blind reliance on statistical default models will result in a failure to assess (and thus regulate) risks taken by financial institutions. Moreover, a reliance on a handful of external

[24] For a detailed perspective, see Kashyap, Rajan and Stein (2008).

credit assessment institutions to determine the riskiness of loans will amplify the effects of errors in the basic model, including those caused by a change in the underlying regime.

What can market participants do to better predict the future? Even sophisticated agents such as regulators setting capital requirements or rating agencies will take some time to learn about the exact magnitudes of relevant variables following a regime change. Nevertheless, we certainly expect them to be aware that incentive effects may lead to such a regime change, which can systematically bias default predictions downward. Once sufficient data has accumulated in the new regime, a statistical model can be reliably estimated (until the regime changes yet again). During the learning phase, however, participants need to be particularly aware that predictions from the default model are probabilistic and the set of possible future scenarios has expanded in an adverse way. Thus, the assessment of default risk must be extra conservative during this period. We expect that agents in the market will eventually learn that the regime has changed. The challenge for market participants is to recognize such shifts in real time.

VII Appendix: Proofs

Proof of Proposition 1

First, suppose $\alpha = 0$. We prove that the unique equilibrium is the efficient soft information equilibrium.

Suppose the lender observes hard information signal x_h. If the lender chooses to not acquire the soft information signal, the optimal interest rate to offer (by assumption) is $r^*(x_h) = r_1$. This interest rate will be accepted by all three types of borrower, since $r_1 \leq \tau(\theta_i)$ for each i. Thus, the lender's expected profit on the loan is $\tilde{\pi} = \sum_{i=h,\ell,b} \mu_i(x_h)v_i(r_1) - 1$. Suppose the soft information signal fully revealed the borrower's type. The lender's expected profit from acquiring the soft information signal is then $\hat{\pi} = \mu_h(x_h)v_h(r_1) + \mu_l(x_h)v_\ell(r_2) - c$ (since type θ_b is not offered a loan).

Thus, the lender will not acquire the soft information signal when $x = x_h$ if $\tilde{\pi} \geq \hat{\pi}$, or

$$\sum_{i=h,\ell,b} \mu_i(x_h)[\theta_i(1+r_1) - 1] \quad \geq \quad \mu_h(x_h)[\theta_h(1+r_1)-1] + \mu_l(x_h)[\theta_\ell(1+r_2)-1] - c, \qquad (4)$$

which reduces to $c \geq \mu_l(x_h)\theta_\ell(r_2 - r_1) - \mu_b(x_h)v_b(r_1)$.

Next, suppose the hard information signal is x_b. If the lender does not acquire the soft information signal, it does not offer a loan, and obtains a zero profit. Suppose the soft information signal is fully revealing, and the lender chooses to acquire it. Then, its expected profit is $\hat{\pi} = \sum_{i=h,\ell} \mu_i(x_b)v_i(r_i) - c$, since again the type θ_b is not offered a loan. Thus, the lender will not acquire the soft information signal when $x = x_b$ if $c \geq \sum_{i=h,\ell} \mu_i(x_b)v_i(r_i)$.

Thus, if

$$c \quad \geq \quad \max\{\ \mu_l(x_h)\theta_\ell(r_2 - r_1) - \mu_b(x_h)v_b(r_1),\ \sum_{i=h,\ell} \mu_i(x_b)v_i(r_i)\ \}, \qquad (5)$$

the lender will not acquire the soft information signal if $x \in \{x_h, x_b\}$ even when $\alpha = 0$ and the signal is fully revealing.

Finally, suppose the hard information signal is x_ℓ and the soft information signal is noisy. If the lender does not acquire the soft information signal, it offers the interest rate r_2 to the borrower. Only types θ_ℓ and θ_b accept, so the lender's expected profit is $\tilde{\pi} = \sum_{i=\ell,b} \mu_i(x_\ell)v_i(r_2)$. Suppose the lender does acquire the soft information signal. By assumption, the optimal interest rate given hard information signal x_ℓ and soft information signal y_j is $\tau(\theta_i)$ for $j = h, \ell$ with no loan being offered if the signal y_b is obtained. Thus, the lender's expected profit if it acquires the soft information signal is $\hat{\pi} = \sum_{i=h,\ell,b} \mu_i(x_\ell)\gamma(y_h \mid \theta_h, x_\ell)v_i(r_1) + \sum_{i=\ell,b} \mu_i(x_\ell)\gamma(y_\ell \mid \theta_i, x_\ell)v_i(r_2) - c$. The lender will acquire the soft information signal if $\hat{\pi} \geq \tilde{\pi}$, which reduces to

$$c \quad \leq \quad \sum_{i=h,\ell,b} \mu_i(x_\ell)\gamma(y_h \mid \theta_i, x_\ell)v_i(r_1) - \sum_{i=\ell,b} \mu_i(x_\ell)(1 - \gamma(y_\ell \mid \theta_i, x_\ell))v_i(r_2). \qquad (6)$$

By assumption, (5) and (6) are both satisfied. Thus, when $\alpha = 0$, the lender acquires the soft information signal if and only if $x = x_\ell$. Since the lender is strictly worse off following any deviation either in acquiring the soft information signal or in offered interest rates, this is a unique equilibrium.

Now consider the case of $\alpha > 0$. In the conjectured equilibrium, it must be that $P(x_h, r_1) - 1 = \sum_{i=h,\ell,b} \mu_i(x_h)v_i(r_1)$. Further, investors must correctly infer that, if the hard information signal is x_ℓ and the interest rate r_1, the soft information signal was y_h. Thus, $P(x_\ell, r_1) - 1 = \frac{\sum_{i=h,\ell,b} \psi_i(x_\ell,y_h)v_i(r_1)}{\sum_{i=h,\ell,b} \psi_i(x_\ell,y_h)}$, and similarly $P(x_\ell, r_2) - 1 = \frac{\sum_{i=\ell,b} \psi_i(x_\ell,y_\ell)v_i(r_2)}{\sum_{i=\ell,b} \psi_i(x_\ell,y_h)}$. The lender's expected payoff in such an equilibrium therefore equals the second-best payoff. It is immediate that the lender earns a higher payoff than it could by not acquiring soft information when the hard information signal is x_ℓ, or by collecting soft information when the hard information signal is x_h or x_b. Further, the lender's strategy when $x = x_h$ or $x = x_b$ is clearly optimal: no other strategy can increase payoff.

Thus, we only need to show that the lender is following an optimal interest rate strategy when $x = x_\ell$ and it collects soft information. Note that, if a loan is offered, the optimal interest rate for each realization of signals must be one of r_1 or r_2. Any interest rate less than r_1 has a lower payoff than an offer of r_1, and any interest rate between r_1 and r_2 has a lower payoff than an offer of r_2. Finally, any interest rate strictly greater than r_2 has a negative payoff.

Consider the lender's interest rate strategy for each realization of the soft information signal when $x = x_\ell$. Suppose first that $y = y_h$. In equilibrium, the lender offers interest rate r_1 to the borrower and obtains the payoff $u(r_1, \rho \mid y_h) = \sum_{i=h,\ell,b} \psi_i(x_\ell, y_h)v_i(r_1)$. Consider possible deviations by the lender. First, suppose the lender deviates and offers r_2. Since the pair (x_ℓ, r_2) is observed in equilibrium, the price for such a loan satisfies $P(x_\ell, r_2) - 1 = \frac{\sum_{i=\ell,b} \psi_i(x_\ell,y_\ell)v_i(r_2)}{\sum_{i=\ell,b} \psi_i(x_\ell,y_\ell)}$, so that the lender's expected payoff is

$$u(r_2, \rho \mid y_h) \quad = \quad (1-\alpha)\sum_{i=\ell,b} \psi_i(x_\ell, y_h)v_i(r_2) + \alpha \sum_{i=\ell,b} \psi_i(x_\ell, y_h)\frac{\sum_{i=\ell,b} \psi_i(x_\ell, y_\ell)v_i(r_2)}{\sum_{i=\ell,b} \psi_i(x_\ell, y_\ell)}. \quad (7)$$

Since it is a strict best response to offer a loan at r_1 when $\alpha = 0$ and $(x, y) = (x_\ell, y_h)$, it must be that $\sum_{i=h,\ell,b} \psi_i(x_\ell, y_h)v_i(r_1) > \max\{0, \sum_{i=\ell,b} \psi_i(x_\ell, y_h)v_i(r_2)\}$. Hence, there exists an $\alpha_1 \in (0, 1]$ (possibly equal to 1) such that the deviation is suboptimal if $\alpha \leq \alpha_1$. Further, a deviation to not offering a loan is suboptimal for all α, since such a deviation entails zero profit.

Next, suppose $y = y_\ell$. A deviation to offering no loan is sub-optimal, since it reduces the lender's payoff to zero. Consider, however, a deviation to $r = r_1$. Given the lender's equilibrium interest rate strategy ρ, investors price the loan as if $y = y_h$. Thus, the lender's expected payoff from the deviation is

$$u(r_1, \rho \mid y_\ell) = (1-\alpha)\sum_{i=h,\ell,b} \psi_i(x_\ell, y_\ell)v_i(r_1) + \alpha \sum_{i=\ell,b} \psi_i(x_\ell, y_\ell) \sum_{i=h,\ell,b} \psi_i(x_\ell, y_h)v_i(r_1). \qquad (8)$$

In equilibrium, the lender's payoff given (x_ℓ, y_ℓ) is

$$u(r_2, \rho \mid y_\ell) = (1-\alpha) \sum_{i=\ell,b} \psi_i(x_\ell, y_\ell) v_i(r_2) + \alpha \sum_{i=\ell,b} \psi_i(x_\ell, y_\ell) v_i(r_2). \tag{9}$$

Since r_2 is the optimal interest rate when $(x, y) = (x_\ell, y_\ell)$, it follows that $\sum_{i=h,\ell,b} \psi_i(x_\ell, y_\ell) v_i(r_1) < \sum_{i=\ell,b} \psi_i(x_\ell, y_\ell) v_i(r_2)$. Hence, there exists an $\alpha_2 \in (0, 1]$ (possibly equal to 1) such that $u(r_2, \rho \mid y_\ell) \geq u(r_1, \rho \mid y_\ell)$ if and only if $\alpha \leq \alpha_2$.

Finally, suppose $y = y_b$. In equilibrium, the lender does not offer a loan and obtains a zero payoff. However, the lender can deviate to r_1 or r_2. Suppose the optimal deviation is to r_1. Investors assume the hard information signal was y_h, and price the loan accordingly. Therefore, the payoff from such a deviation is

$$u(r_1, \rho \mid y_b) = (1-\alpha) \sum_{i=h,\ell,b} \psi_i(x_\ell, y_b) v_i(r_1) + \alpha \sum_{i=h,\ell,b} \psi_i(x_\ell, y_h) v_i(r_1). \tag{10}$$

Since $\sum_{i=h,\ell,b} \psi_i(x_\ell, y_h) v_i(r_1) > 0 > \sum_{i=h,\ell,b} \psi_i(x_\ell, y_b) v_i(r_1)$, it follows that there exists an $\tilde{\alpha}_h \in (0, 1)$ such that the lender will not deviate to r_1 if and only if $\alpha \leq \tilde{\alpha}_h$. A similar result follows if r_2 is the optimal deviation instead: There exists an $\tilde{\alpha}_\ell \in (0, 1]$ such that the lender will not deviate to r_2 if and only if $\alpha \leq \tilde{\alpha}_\ell$.

Finally, define $\underline{\alpha} = \min\{\alpha_1, \alpha_2, \tilde{\alpha}_h, \tilde{\alpha}_\ell\}$. Then, an efficient soft information equilibrium exists if and only if $\alpha \leq \underline{\alpha}$. ■

Proof of Proposition 2

In a hard information equilibrium, $P(x, r) - 1 = \frac{\sum_{\{i: r \leq \tau(\theta_i)\}} \mu_i(x) v_i(r)}{\sum_{\{i: r \leq \tau(\theta_i)\}} \mu_i(x)}$ for each $(x, r) = (x_h, r_1)$ or (x_ℓ, r_2). For any other combination of (x, r), we impose the belief that the posterior probability of type θ_i is $\frac{\mu_i(x) 1_{\{r \leq \tau(\theta_i)\}}}{\sum_{\{j: r \leq \tau(\theta_j)\}} \mu_j(x)}$ if $r \leq r_2$, and the borrower has type θ_b if $r > r_2$. Hence, for all $r \leq r_2$, the pricing function for loans satisfies $P(x, r) - 1 = \frac{\sum_{\{i: r \leq \tau(\theta_i)\}} \mu_i(x) v_i(r)}{\sum_{\{i: r \leq \tau(\theta_i)\}} \mu_i(x)}$.

We now show that, for α sufficiently high, the lender is playing a best response in a hard information equilibrium. First, suppose the hard information signal is x_h, and the lender deviates and acquires soft information. From the proof of Proposition 1, it is sub-optimal for the lender to acquire the soft information signal if the loan will be retained. The payoff if the loan is sold is at most $P(x_h, r_1) - c$, since $P(x_h, r_1) \geq P(x_h, r)$ for all $r \leq r_2$. In equilibrium, the lender's payoff is $u(r_1 \mid x_h) = P(x_h, r_1)$. Thus, whether the loan is retained or sold, the lender's profit is reduced, so the deviation is sub-optimal.

Next, suppose the hard information signal is x_b, and the lender deviates and acquires soft information. From the proof of Proposition 1, it is sub-optimal for the lender to acquire the soft information signal if the loan will be retained. If the loan is sold, the lender's payoff is negative, since $P(x_b, r) < 0$ for all $r \leq r_2$. Thus, whether the loan is retained or sold, the lender's profit is reduced, and the deviation is sub-optimal.

Finally, suppose the hard information signal is x_ℓ and the lender deviates and acquires soft information. To improve payoff, the deviation must be accompanied by an interest rate strategy that offers different interest rates to at least two types of borrower. Suppose the optimal deviation involves offering a loan at r_1 to borrowers with signals (x_ℓ, y_h) and at r_2 to borrowers with signals (x_ℓ, y_ℓ), and not offering a loan to borrowers with signals (x_ℓ, y_b). If the loan is retained, the expected payoff from this action is (as in the proof of Proposition 1 above) $\hat{\pi}^d = \sum_{i=h,\ell,b} \mu_i(x_\ell)\gamma(y_h \mid \theta_h, x_\ell)v_i(r_1) + \sum_{i=\ell,b} \mu_i(x_\ell)\gamma(y_\ell \mid \theta_i, x_\ell)v_i(r_2) - c$. If the loan is sold, the expected payoff from the deviation is $\tilde{\pi}^d = [P(x_\ell, r_1) - 1] \sum_{i=h,\ell,b} \mu_i(x_\ell)\gamma(y_h \mid \theta_i, x_\ell) + [P(x_\ell, r_2) - 1] \sum_{i=\ell,b} \mu_i(x_\ell)\gamma(y_\ell \mid \theta_i, x_\ell) - c$.

In equilibrium, whether the loan is sold or retained, the expected payoff is $\hat{\pi}^e = \sum_{i=\ell,b} \mu_i v_i(r_2)$. Given the assumptions on the cost of soft information c, $\hat{\pi}^d > \hat{\pi}^e$. However, since $P(x_\ell, r_2) > P(x_\ell, r_1)$, it follows that $\tilde{\pi}^d < \hat{\pi}^e$. Therefore, the deviation is sub-optimal if and only if α exceeds some $\alpha_1 \in (0, 1)$.

A similar argument applies if the optimal interest rate strategy following the soft information signal is different. For example, suppose the optimal interest rate strategy is to offer r_1 to all borrowers with signals (x_ℓ, y_h) and r_2 to all borrowers with signals (x_ℓ, y_ℓ) and (x_ℓ, y_b). Then, there exists an $\alpha_2 \in (0, 1)$ such that the deviation is sub-optimal if and only if $\alpha_2 \geq \bar{\alpha}$. Now, the optimal interest rate given the soft information signal must be one of r_1 or r_2. Thus, there is a finite number of interest rate strategies to consider. Define $\bar{\alpha}$ to be the maximum of the critical values of α across all such strategies. It then follows that the hard information equilibrium is sustained if and only if $\alpha \geq \bar{\alpha}$.

Finally, we show the hard information equilibrium is unique when $\alpha = 1$. Suppose $\alpha = 1$. We have shown it is sub-optimal to collect soft information when $x = x_h$ or x_b, so conjecture an equilibrium in which soft information is collected when $x = x_\ell$. In such an equilibrium, it must be that $P(x_\ell, r_1) > P(x_\ell, r_2)$, else it is sub-optimal to collect soft information. But if $P(x_\ell, r_1) > P(x_\ell, r_2)$, then, regardless of the soft information signal, the lender should offer interest rate r_1 to all borrowers, breaking the conjectured equilibrium. Hence, the only equilibrium when $\alpha = 1$ is the hard information equilibrium. ■

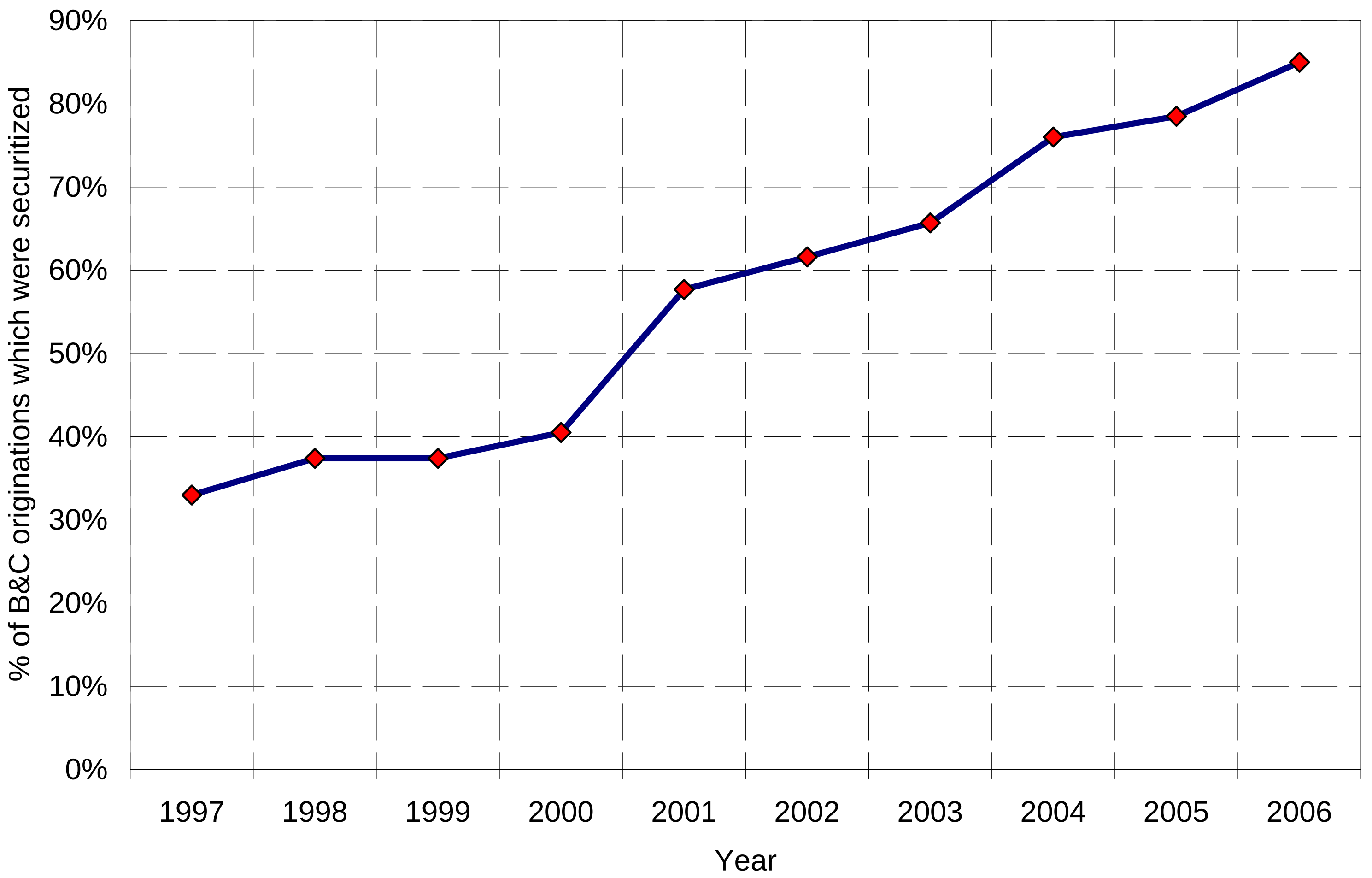


Figure 1: Securitization of B&C Loans

This figure presents the time series trend of securitization of B&C (i.e., subprime) loans from 1997 to 2006. The source for the data is *Inside B&C Lending*, a publication that has extensive coverage of the subprime mortgage market. The figure shows that the percentage of loans securitized has increased sharply from 2000 onwards.

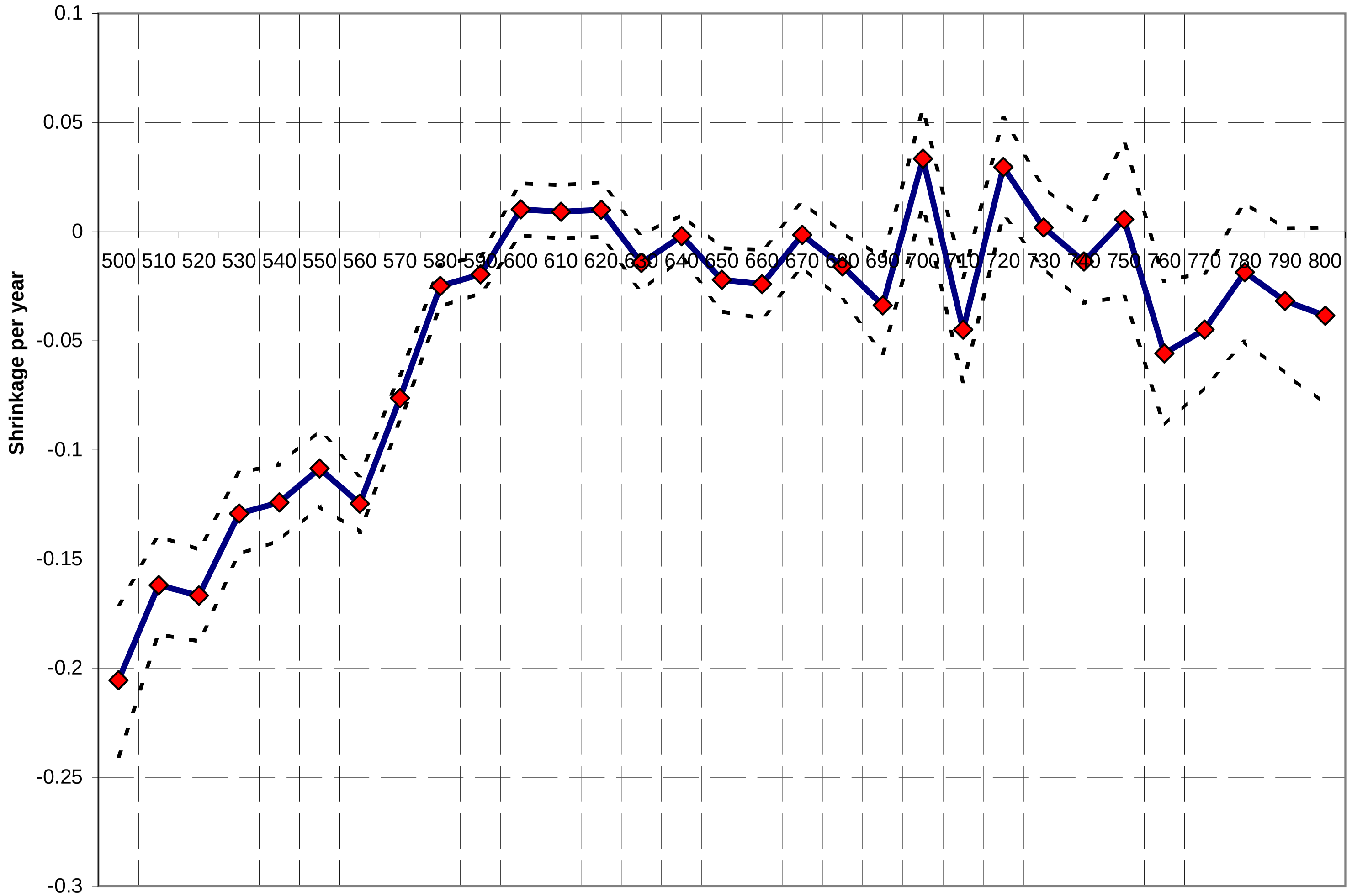


Figure 2: Shrinkage in the Distribution of Interest Rates Across the FICO Spectrum Over Time

This figure presents the slope coefficients from a regression of the dispersion of interest rates (standard deviation) at each FICO score on time for 10 point FICO score buckets (starting FICO of 500), controlling for dispersion in the features of the loan contract. The figure shows that more shrinkage occurs at low FICO scores as compared to high FICO scores.

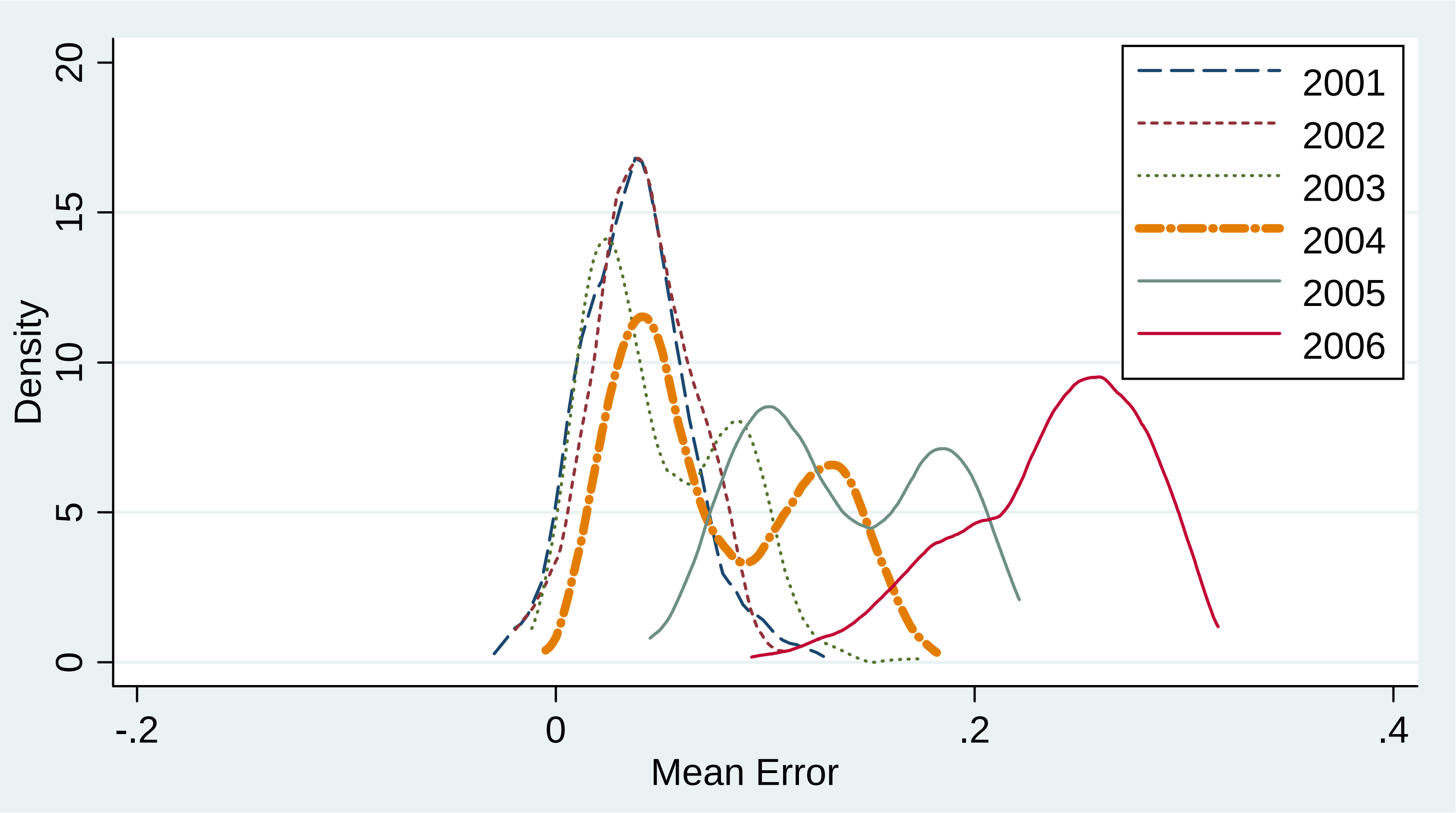


Figure 3: Kernel Density of Mean Prediction Errors Over Time, All Loans

This figure presents the Epanechnikov kernel density of mean prediction errors (*Actual Defaults - Predicted Defaults*) of a baseline model estimated for loans issued in 1997 to 2000. For each subsequent year, we first determine the mean prediction error at each FICO score, and then plot the kernel density of the mean errors. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991). The figure shows that the baseline model underpredicts defaults, and the mean error increases as securitization increases from 2001 to 2006.

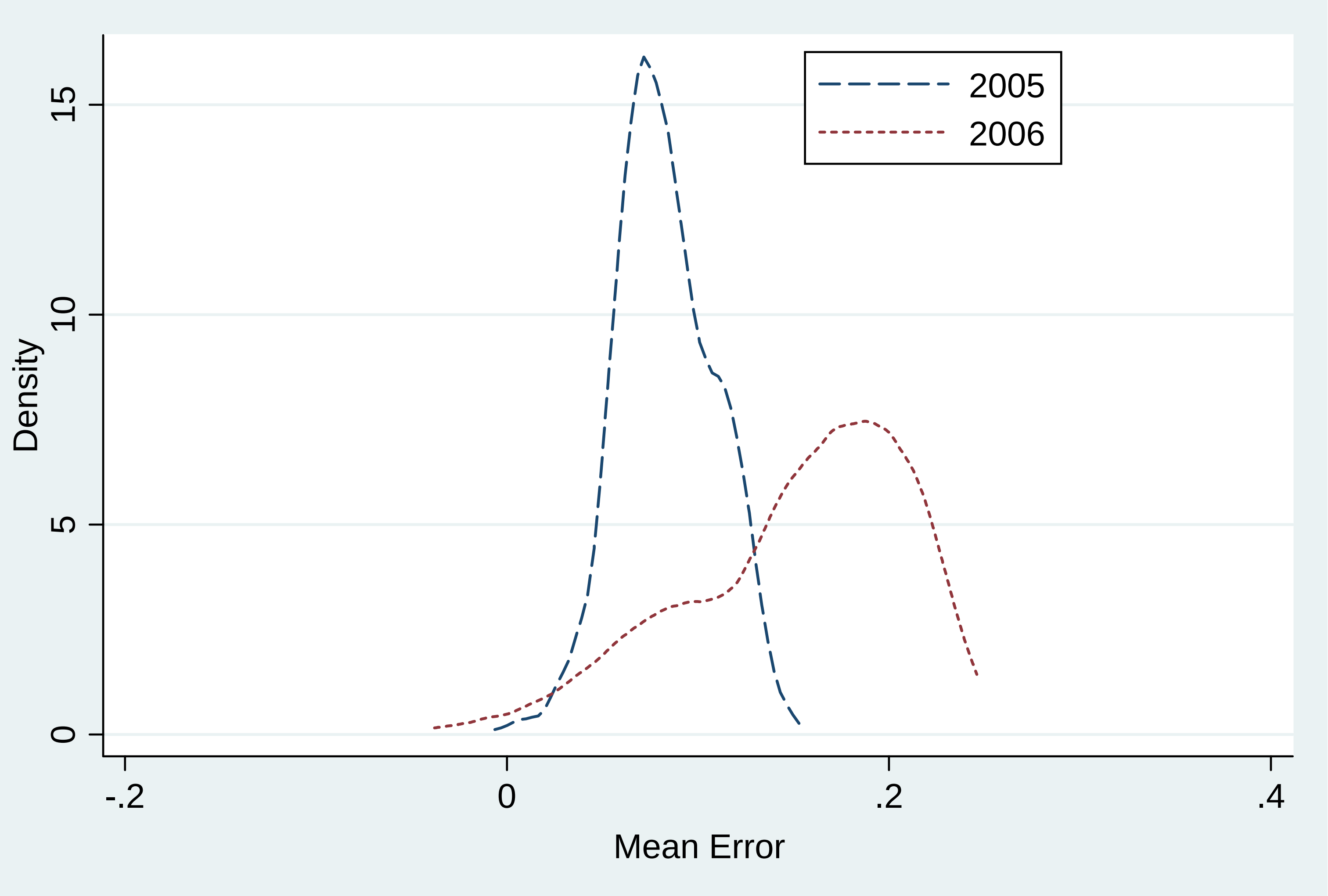
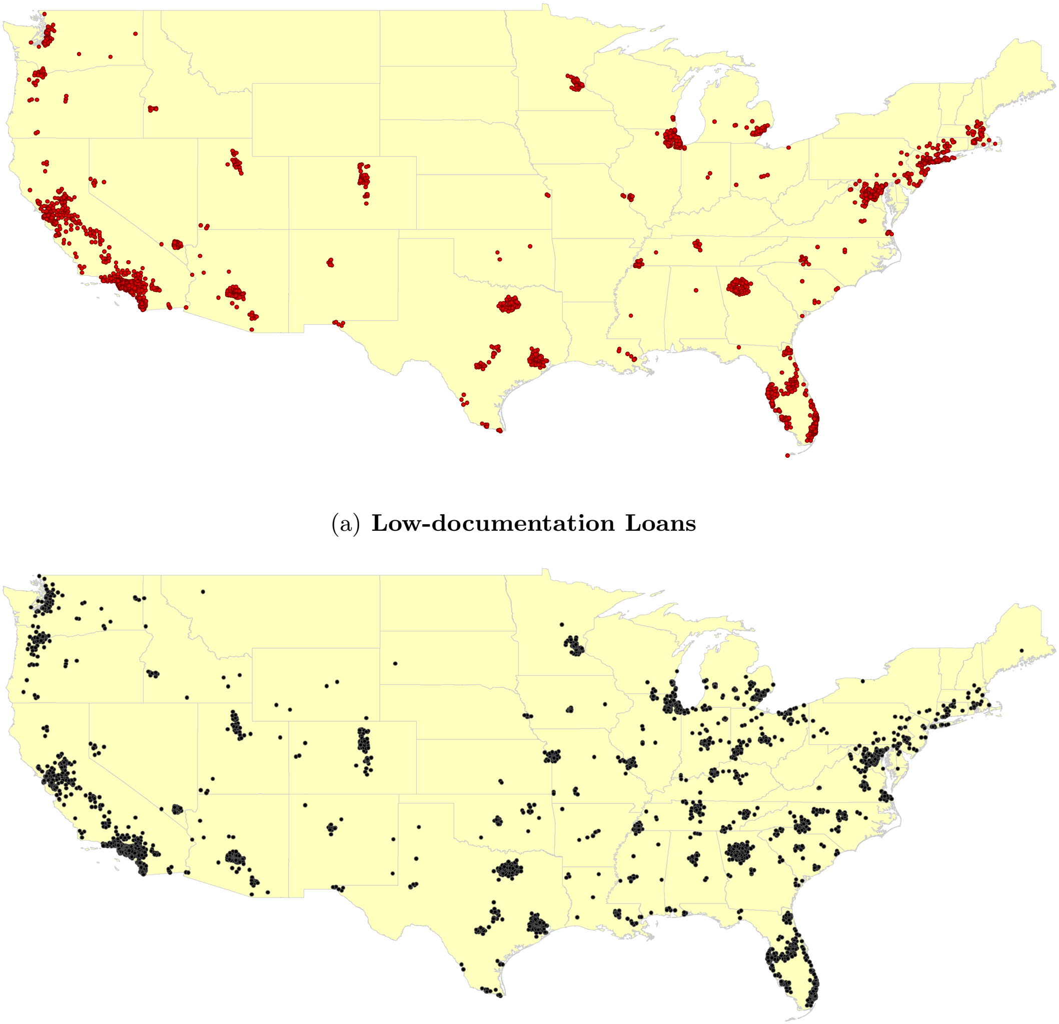


Figure 4: Kernel Density of Mean Prediction Errors Over Time with a Rolling Estimation Window

This figure presents the Epanechnikov kernel density of mean prediction errors (*Actual Defaults - Predicted Defaults*) of a baseline model using a rolling estimation window. The prediction errors for 2005 loans are from a baseline model estimated over 1997 to 2004, and for 2006 loans from a baseline model estimated over 1997 to 2005. For each year, we first determine the mean prediction error at each FICO score, and then plot the kernel density of the mean errors. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991). The figure shows that the baseline model underpredicts defaults in 2005 and 2006.

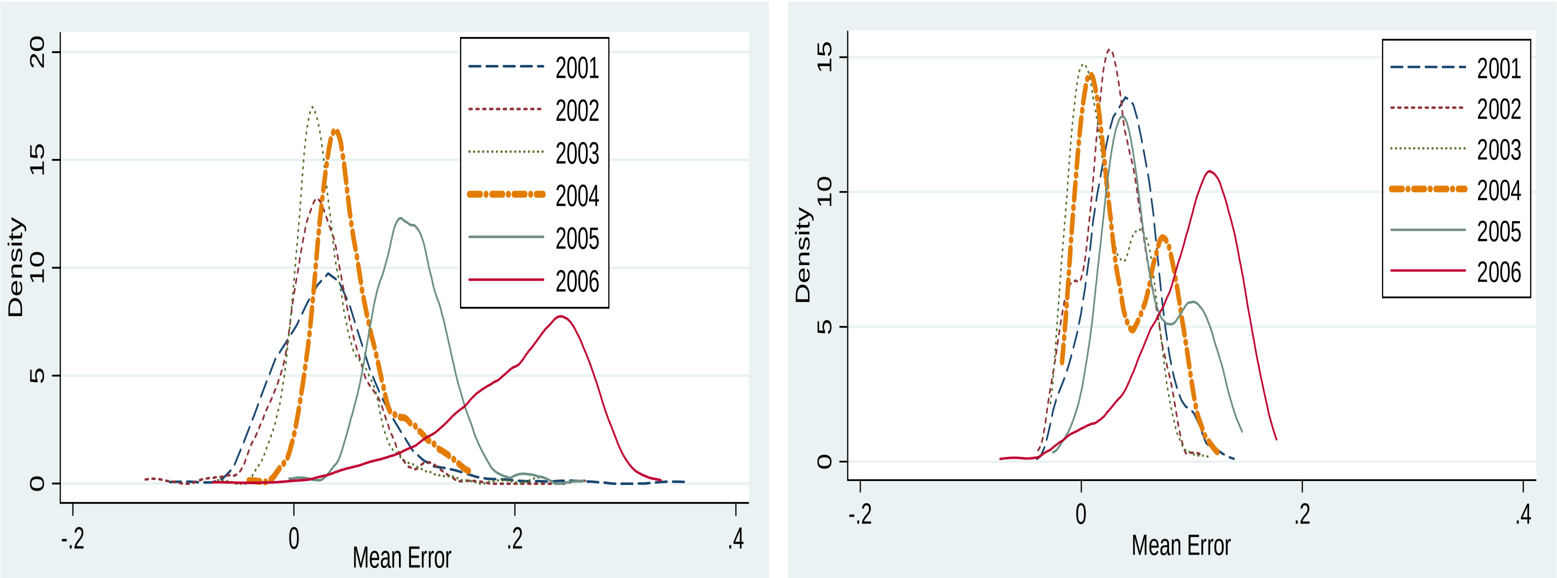


(a) **Low-documentation Loans**

(b) **Full-documentation Loans**

Figure 5: Top 25% of Zip Codes for Subprime Loans, 2001–2006

These figures display the top 25% of zip codes (by number of loans) in which low-documentation (top; figure (a)) and full-documentation (bottom; figure(b)) subprime mortgage loans issued made over the period 1997–2006. These zip codes contribute over 60% of the volume of subprime loans in the respective category. The figure shows that there was substantial overlap of zip codes across the two kinds of loans, with concentrations in places such as California, Florida and the North-East.



(a) **Low-documentation Loans**

(b) **Full-documentation Loans**

Figure 6: Kernel Density of Mean Prediction Errors Over Time for Low-Documentation and Full-Documentation Loans with a Rolling Estimation Window

These figures presents the Epanechnikov kernel density of mean prediction errors (*Actual Defaults - Predicted Defaults*) on low-documentation (figure (a)) and full-documentation (figure(b)) loans of a baseline model using a rolling estimation window. The prediction errors for year $t+1$ are from a baseline model estimated over 1997 to year t. For each year, we first determine the mean prediction error at each FICO score, and then plot the kernel density of the mean errors. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991). Figure (a) shows that for low-documentation loans the baseline model underpredicts defaults and the mean error increases as securitization increases from 2001 to 2006. Figure (b) shows that the baseline model underpredicts defaults by a smaller order of magnitude for full-documentation loans.

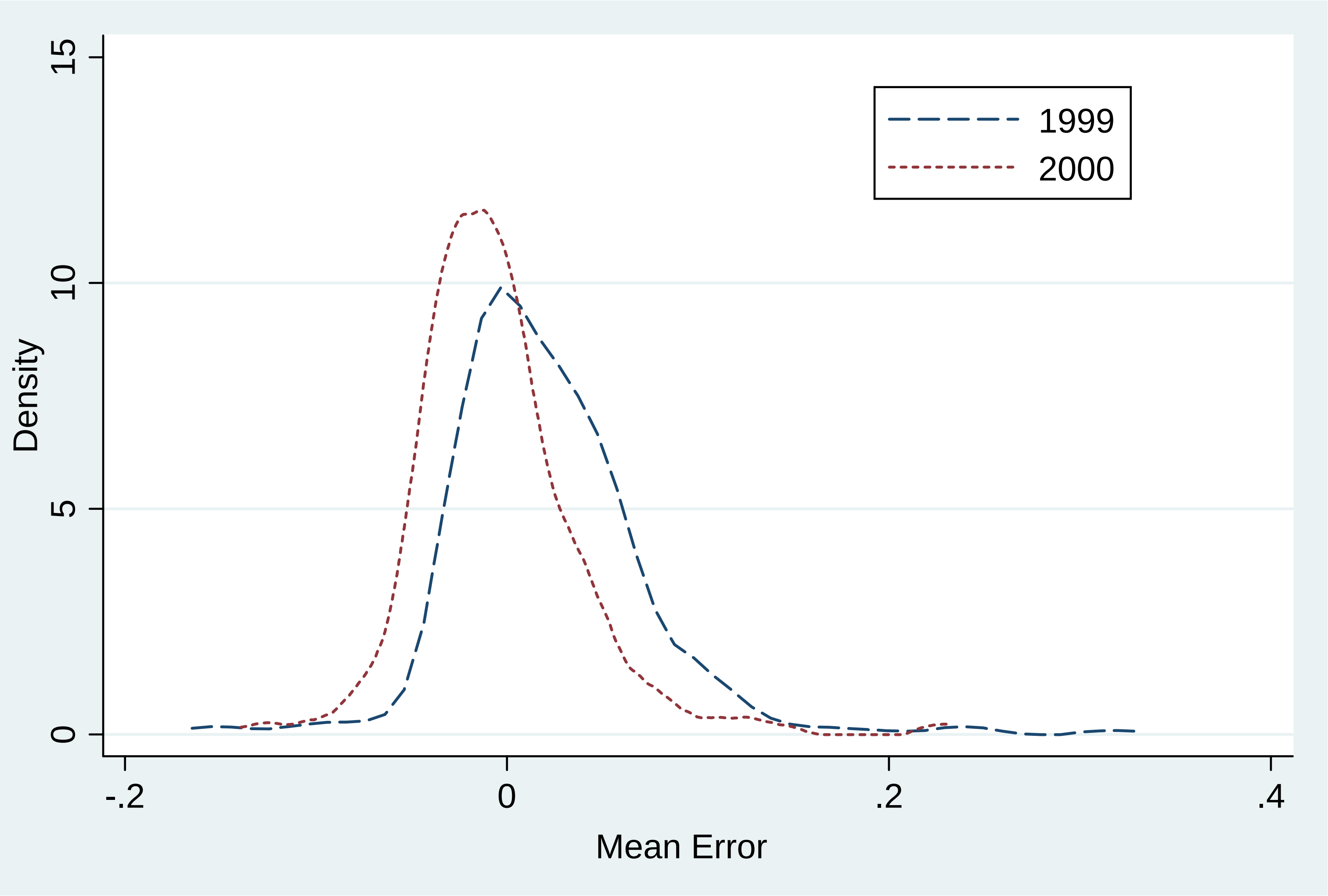


Figure 7: Kernel Density of Mean Prediction Errors in Low Securitization Period for Low-Documentation Loans Only with a Rolling Estimation Window

This figure presents the Epanechnikov kernel density of mean prediction errors (*Actual Defaults - Predicted Defaults*) for low-documentation loans issued in 1999 to 2000. The baseline model for 1999 loans is estimated over 1997 and 1998 and the baseline model for 2000 loans is estimated from 1997 through 1999. For each year 1999 and 2000, we first determine the mean prediction error at each FICO score, and then plot the kernel density of the mean errors. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991). The figure shows that the prediction error from the baseline model is centered around zero for loans issued in a low securitization period (1999 and 2000).

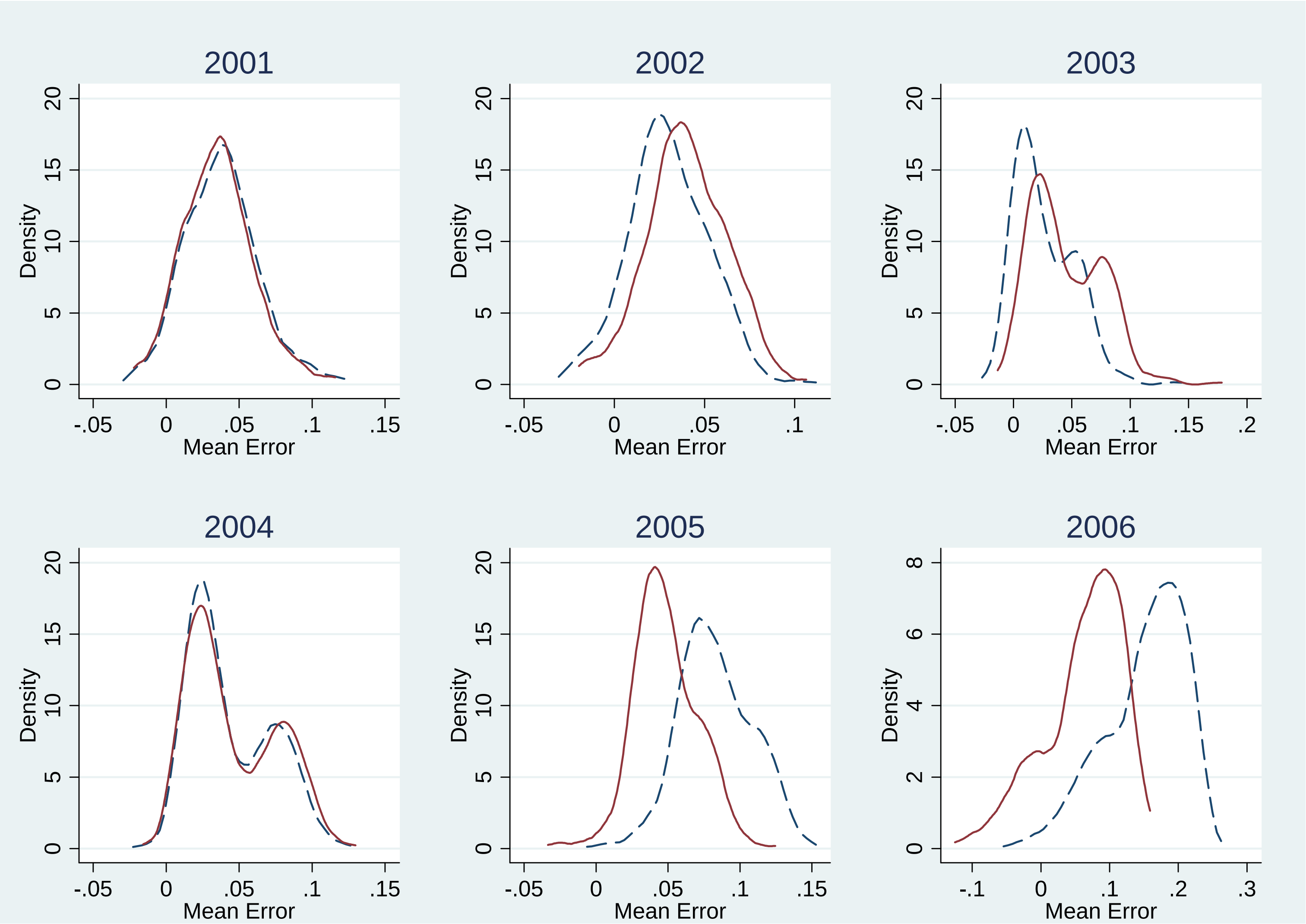
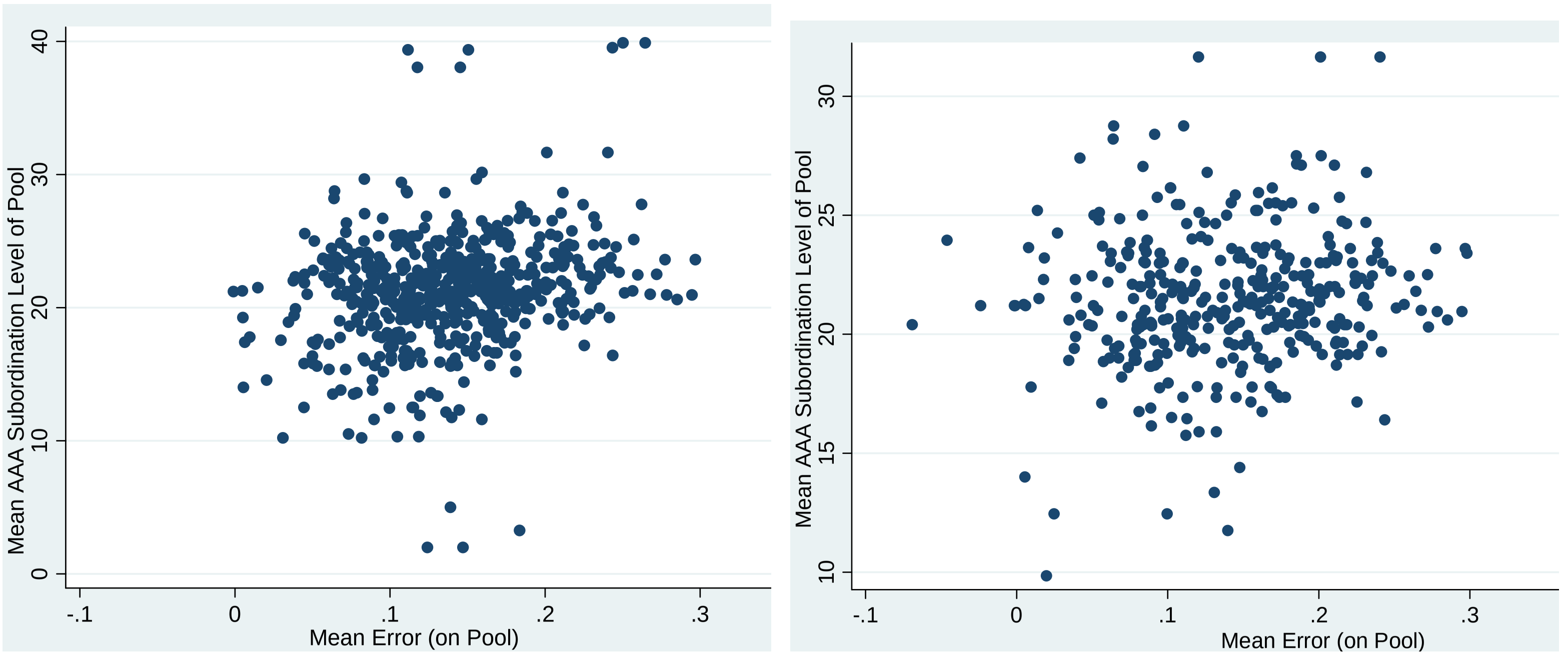


Figure 8: Kernel Density of Mean Prediction Errors With (Solid Lines) and Without (Dashed Lines) Considering House Price Appreciation, using a Rolling Estimation Window

This figure presents the Epanechnikov kernel density of mean prediction errors (*Actual Defaults* - *Predicted Defaults*) on all loans of a baseline model using a rolling estimation window. The prediction errors for year $t+1$ are from a baseline model estimated over 1997 to year t, with and without including house price appreciation (HPA) as an explanatory variable. For each year, we first determine the mean prediction error at each FICO score, and then plot the kernel density of the mean errors. For each year, the dashed line represents the density of errors without HPA and the solid line the density of errors with HPA included. The bandwidth for the density estimation is selected using the plug-in formula of Sheather and Jones (1991). The figure shows that, even after including HPA, the prediction errors remain positive and large in magnitude.



(a) **2005 Loans**

(b) **2006 Loans**

Figure 9: Pool Subordination Level with Mean Prediction Error for 2005 and 2006 Loans

These figures present the scatter plot of mean subordination level of AAA tranches in a pool against the mean prediction error of defaults in that pool for loans issued in 2005 (figure (a)) and 2006 (figure (b)). To highlight whether there is a relationship between subordination levels and prediction errors on default, we consider only pools for which prediction errors (i.e., actual defaults − predicted defaults given the baseline model) are likely to be high: we restrict attention to pools with at least 30% low-documentation loans. The figure suggests that there is no relationship between the prediction errors from the default model and subordination levels of AAA tranches.

Table I: Summary Statistics

This table reports summary statistics of FICO scores, LTV (loan-to-value) ratios and information on the documentation reported by the borrower when taking the loan. Documentation is categorized as full, limited and no. full-documentation loans provide verification of income as well as assets of the borrower. Limited documentation provides no information about the income but does provide some information about the assets. No documentation loans provide no information about income or assets. We combine limited and no documentation loans and call them ‘low-documentation’ loans. See the text for information on sample selection.

Sample Characteristics

Year	Number of	% Low	Mean	Mean
	Loans	Documentation	Loan-To-Value	FICO
1997	24,067	24.9%	80.5	611
1998	60,094	23.0%	81.5	605
1999	104,847	19.2%	82.2	610
2000	116,778	23.5%	82.3	603
2001	136,483	26.0%	84.6	611
2002	162,501	32.8%	85.6	624
2003	318,866	38.9%	87.0	637
2004	610,753	40.8%	86.6	639
2005	793,725	43.4%	86.3	639
2006	614,820	44.0%	87.0	636

Table II: Reliance of Interest Rates on FICO Scores and LTV Ratios

This table reports estimates from the yearly regression of interest rates on FICO and LTV. See the text for information on sample selection.

	β_{FICO}	β_{LTV}	R^2 (in %)	Observations
1997	-0.004***	0.030***	3	24,067
	(.0002)	(.0013)		
1998	-0.007***	0.035***	7	60,094
	(.0001)	(.0008)		
1999	-0.007***	0.020***	8	104,847
	(.0001)	(.0005)		
2000	-0.010***	0.035***	14	116,778
	(.0001)	(.0004)		
2001	-0.012***	0.038***	20	136,483
	(.0001)	(.0004)		
2002	-0.011***	0.071***	18	162,501
	(.0001)	(.0001)		
2003	-0.012***	0.079***	32	318,866
	(.0001)	(.0001)		
2004	-0.010***	0.097***	40	610,753
	(.0001)	(.0001)		
2005	-0.009***	0.110***	48	793,725
	(.0001)	(.0001)		
2006	-0.011***	0.115***	50	614,820
	(.0001)	(.0001)		

Table III: Shrinkage in the Distribution of Interest Rates

We report estimates from regression of yearly standard deviation of interest rates at each FICO score on time. The regressions are estimated separately in buckets of ten FICO points, in the range 500 to 800. The sample period is from 1997–2006. See the text for sample selection.

FICO	β_b	Std. Err.	R^2 (%)
500	-0.212***	(0.019)	53
510	-0.191***	(0.013)	67
520	-0.214***	(0.013)	71
530	-0.179***	(0.011)	71
540	-0.17***	(0.009)	74
550	-0.151***	(0.010)	69
560	-0.146***	(0.008)	75
570	-0.126***	(0.009)	65
580	-0.062***	(0.009)	31
590	-0.052***	(0.008)	25
600	-0.035***	(0.008)	14
610	-0.037***	(0.008)	17
620	-0.035***	(0.007)	17
630	-0.023***	(0.006)	10
640	-0.023***	(0.005)	13
650	-0.043***	(0.007)	23
660	-0.049***	(0.009)	22
670	-0.06***	(0.009)	27
680	-0.047***	(0.008)	22
690	-0.058***	(0.010)	25
700	-0.05***	(0.011)	16
710	-0.059***	(0.012)	19
720	-0.055***	(0.010)	21
730	-0.101***	(0.013)	35
740	-0.085***	(0.012)	33
750	-0.071***	(0.016)	14
760	-0.066***	(0.015)	15
770	-0.045***	(0.013)	9
780	-0.059***	(0.015)	11
790	-0.064***	(0.019)	9
800	-0.065***	(0.032)	3

Table IV: Default Model—Failing to Predict Failure

We report estimates from a baseline default model estimated for low and full-documentation loans issued from 1997 to 2000 in Panel A. Panel B reports the β coefficients from a regression of prediction error on FICO score and LTV ratio for loans issued from each year 2001 to 2006, and also reports the mean prediction errors for each vintage. *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level. See the text for additional information on sample selection.

Panel A: Coefficients of Baseline Model in Low Securitization Regime, 1997–2000

$FICO$	r	LTV	Constant	$I^{Low} \times FICO$	$I^{Low} \times r$	$I^{Low} \times LTV$	I^{Low}	Pseudo R^2 (%)	Observations	Other Controls
-0.009***	0.231***	0.003***	0.570***	0.001***	-0.043***	-0.008***	0.168	7.05	267,511	Yes
(0.0001)	(0.006)	(0.001)	(0.129)	(0.0001)	(0.016)	(0.001)	(0.338)			

Panel B: Prediction Errors during High Securitization Regime.

	β_{FICO}	β_{LTV}	Observations	R^2	Mean Prediction	Actual Defaults
	($\times 10^{-3}$)	($\times 10^{-2}$)		(%)	Error (%)	(%)
2001	-0.123***	0.052***	128,772	0.05	3.96***	16.0
	(0.018)	(.010)				
2002	-0.197***	0.082***	152,057	0.15	4.70***	14.1
	(0.015)	(.010)				
2003	-0.428***	0.077***	308,340	0.61	5.01***	11.9
	(0.010)	(0.010)				
2004	-0.621***	0.061***	596,485	0.97	7.79***	13.9
	(0.008)	(0.004)				
2005	-1.341***	0.143***	788,299	3.90	14.67***	21.1
	(0.030)	(0.007)				
2006	-1.120***	0.190***	608,559	1.60	25.49***	33.2
	(0.012)	(0.005)				

Table V: Default Model—Mean Prediction Errors for Low- and Full-Documentation Loans with a Rolling Estimation Window

We report the mean prediction errors for low and full-documentation loans issued from 2001 through 2006. The estimation uses a rolling window approach with separate baseline models for low-documentation and full-documentation loans. That is, the predictions for year $t + 1$ are based on a model estimated over the years 1 through t, where year 1 is 1997. ***, ** and * represent that differences are significant at the 1%, 5% and 10% levels respectively.

	Low-Documentation	Full-Documentation	Difference (%)
	(%)	(%)	(Low-Documentation − Full-Documentation)
2001	3.40	3.80	-0.40
2002	2.78	2.79	-0.01
2003	3.20	2.21	0.99***
2004	5.17	3.51	1.66***
2005	10.58	5.85	4.73***
2006	20.11	9.84	10.27***

Table VI: Default Model—Low Securitization Years, Low-Documentation Loans Only with a Rolling Estimation Window

We report estimates from a baseline default model estimated for low-documentation loans issued in 1997 and 1998 in Panel A. Panel B reports the β coefficients from a regression of prediction error on FICO score and LTV ratio for loans issued in 1999 and 2000, and also reports the mean prediction errors for each vintage. *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level. See the text for additional information on sample selection.

Panel A: Coefficients of Baseline Model in Low Securitization Regime

	$FICO$	r	LTV	Constant	Pseudo R^2 (%)	Observations	Other Controls
1997-1998	-0.009***	0.249***	-0.008***	0.922	8.11	16,002	Yes
	(0.0005)	(0.034)	(0.003)	(0.695)			
1997-1999	-0.007***	0.259***	-0.003*	-0.354	7.94	33,868	Yes
	(0.003)	(0.022)	(0.001)	(0.436)			

Panel B: Prediction Errors during Low Securitization Regime.

	β_{FICO}	β_{LTV}	Observations	R^2	Mean Prediction	Actual Defaults
	$(\times 10^{-3})$	$(\times 10^{-2})$		(%)	Error (%)	(%)
1999	0.039	0.026	17,866	0.01	0.91	11.0
	(0.038)	(.023)				
2000	0.039	-0.026	24,591	0.01	0.97	11.9
	(0.034)	(.020)				

References

[1] Basel Committee on Banking Supervision (2006), "International Convergence of Capital Measurement and Capital Standards: A Revised Framework, Comprehensive Version," http://www.bis.org/publ/bcbs128.pdf.

[2] Baumol, William J. (1958), "On the Theory of Oligopoly," *Economica* 25(99): 187–198.

[3] Benmelech, Efraim and Jennifer Dlugosz (2008), "The Alchemy of CDO Credit Ratings," Working paper, Harvard University.

[4] Berger, Allen N., Nathan H. Miller, Mitchell A. Petersen, Raghuram G. Rajan, and Jeremy C. Stein (2005), "Does Functional Form Follow Organizational Form? Evidence from the Lending Practices of Large and Small Banks," *Journal of Financial Economics* 76(2): 237–269.

[5] Bolton, Patrick and Xavier Freixas (2000), "Equity, Bonds, and Bank Debt: Capital Structure and Financial Market Equilibrium Under Asymmetric Information," *Journal of Political Economy*, 108(2): 324–351.

[6] Chomsisengphet, Souphala and Anthony Pennington-Cross (2006), "The Evolution of the Subprime Mortgage Market," *Federal Reserve Bank of St. Louis Review*, 88:1, 31-56.

[7] Cole, Rebel, Lawrence Goldberg and Lawrence White (1998), "Cookie-Cutter Versus Character: The Micro Structure Of Small Business Lending By Large And Small Banks," *FRB Chicago Working Paper*.

[8] Dell'Ariccia, Giovanni, Deniz Igan and Luc A. Laeven (2008), "Credit Booms and Lending Standards: Evidence from the Subprime Mortgage Market," Working Paper.

[9] Demyanyk, Yuliya and Otto Van Hemert (2007), "Understanding the Subprime Mortgage Crisis." Working Paper.

[10] Doms, Mark, Fred Furlong and John Krainer (2007), "Subprime Mortgage Delinquency Rates," *FRB San Francisco Working Paper* 2007-33.

[11] Drucker, Steven and Manju Puri (2008), "On Loan Sales, Loan Contracting, and Lending Relationships," *Review of Financial Studies*, forthcoming.

[12] Duffee, Gregory R. and Chunsheng Zhou (2001), "Credit Derivatives in Banking: Useful Tools For Managing Risk?", *Journal of Monetary Economics* 48: 25–54.

[13] Einav, Liran, Mark Jenkins and Jonathan Levin (2008), "The Impact of Information Technology on Consumer Lending," Working paper.

[14] Fishelson-Holstein, Hollis (2005), "Credit Scoring Role in Increasing Homeownership for Underserved Populations," in Retsinas and Belsky, eds., *Building Assets, Building Credit: Creating Wealth in Low-Income Communities*, Washington, D.C.: Brookings Institution Press.

[15] Gerardi, Kristopher, Andreas Lehnert, Shane M. Sherlund and Paul Willen, "Making Sense of the Subprime Crisis," Working Paper, Federal Reserve Bank of Boston and NBER.

[16] Gorton, Gary B. and George G. Pennacchi (1995), "Banks and Loan Sales: Marketing Nonmarketable Assets," *Journal of Monetary Economics*, 35, 389-411.

[17] Gorton, Gary B. and Nicholas S. Souleles (2005), "Special Purpose Vehicles and Securitization," *FRB Philadelphia Working Paper* 05-21.

[18] Gramlich, Edward (2007), "Subprime Mortgages: America's Latest Boom and Bust," Washington, D.C.: *The Urban Institute Press.*

[19] Greenlaw, David, Jan Hatzius, Anil K Kashyap and Hyun Song Shin (2008), "Leveraged Losses: Lessons from the Mortgage Market Meltdown," U.S. Monetary Policy Forum Report No. 2.

[20] Greenspan, Alan (2008), Testimony before House Committee of Government Oversight and Reform, Oct 23, 2008.

[21] Griffin, John M. and Dragon Y. Tang (2008), "What Drove the Mismatch between Initial CDO Credit Ratings and Subsequent Performance?", Working Paper.

[22] Holloway, Thomas, Gregor MacDonald and John Straka (1993), "Credit Scores, Early-Payment Mortgage Defaults, and Mortgage Loan Performance," *Freddie Mac Working Paper.*

[23] Inderst, Roman (2008), "Loan Origination under Soft- and Hard-Information Lending," Working Paper, University of Frankfurt.

[24] Kashyap, Anil, Raghuram Rajan and Jeremy Stein (2008), "Rethinking Capital Regulation," Paper prepared for the Federal Reserve Bank of Kansas City Symposium, Jackson Hole.

[25] Keys, Benjamin J., Tanmoy K. Mukherjee, Amit Seru and Vikrant Vig (2008), "Did Securitization Lead to Lax Screening? Evidence from Subprime Loans," Working Paper, SSRN.

[26] Kroszner, Randall (2007), "Innovation, Information, and Regulation in Financial Markets", Speech on November 30, 2007, Federal Reserve Board.

[27] Liberti, Jose and Atif Mian (2008), "Estimating the Effect of Hierarchies on Information Use", Forthcoming, *Review of Financial Studies.*

[28] Lucas, Robert E., Jr. (1976), "Econometric Policy Evaluation: A Critique," in K. Brunner and A.H. Meltzer, eds., *The Phillips Curve and Labor Markets, Carnegie-Rochester Conferences on Public Policy*, Amsterdam: *North Holland Press.*

[29] Mayer, Christopher and Karen Pence (2008), "Subprime Mortgages: What, Where, and to Whom?" Working Paper No. 14083, NBER.

[30] Mayer, Christopher, Karen Pence and Shane Sherlund (2008), "The Rise in Mortgage Defaults: Facts and Myths," *Journal of Economic Perspectives*, forthcoming.

[31] Mian, Atif and Amir Sufi (2008), "The Consequences of Mortgage Credit Expansion: Evidence from the 2007 Mortgage Default Crisis," Working Paper.

[32] Parlour, Christine A. and Guillaume Plantin (2008), "Loan Sales and Relationship Banking," *Journal of Finance* 63(3): 1291–1314.

[33] Pennacchi, George (1988), "Loan Sales and the Cost of Bank Capital," *Journal of Finance*, 43(2): 375–396.

[34] Petersen, Mitchell A. and Raghuram G. Rajan (2002), "Does Distance Still Matter? The Information Revolution in Small Business Lending," *Journal of Finance*, 57(6), 2533-2570.

[35] Purnanandam, Amiyatosh K. (2008), "Originate-to-Distribute Model and the Sub-Prime Mortgage Crisis," Working Paper, SSRN.

[36] Sheather, S.J. and M.C. Jones (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society, Series B* 53: 683–690.

[37] Standard & Poor's (2007), "Standard & Poor's Enhances LEVELS® 6.1 Model," News release, November 9, 2007, available at www2.standardandpoors.com.

[38] Stein, Jeremy (2002), "Information Production and Capital Allocation: Decentralized versus Hierarchical Firms," *Journal of Finance*, 57(5), 1891-1921