Markov Chain Models for Delinquency: Transition Matrix Estimation and Forecasting

Scott D. Grimshaw 1*, William P. Alexander 2

1 Department of Statistics, Brigham Young University, Provo, UT
2 Works Progress Analytics, Boise, ID

email: Scott D. Grimshaw (grimshaw@byu.edu)

* Correspondence to Scott D. Grimshaw, Department of Statistics, Brigham Young University, Provo, UT 84602

KEY WORDS: Delinquency movement matrix, Dirichlet-Multinomial posterior, empirical Bayes, loss forecasts, portfolio valuation, roll rates.

ABSTRACT

A Markov chain is a natural probability model for accounts receivable. For example, accounts that are “current” this month have a probability of moving next month into “current”, “delinquent” or “paid-off” states. If the transition matrix of the Markov chain were known, forecasts could be formed for future months for each state. This paper applies a Markov chain model to subprime loans that appear neither homogeneous nor stationary. Innovative estimation methods for the transition matrix are proposed. Bayes and empirical Bayes estimators are derived where the population is divided into segments or subpopulations whose transition matrices differ in some, but not all entries. Loan-level models for key transition matrix entries can be constructed where loan-level covariates capture the nonstationarity of the transition matrix. Prediction is illustrated on a $7 billion portfolio of subprime fixed first mortgages and the forecasts show good agreement with actual balances in the delinquency states.
1 Introduction

Cyert et al. [1] proposed a discrete-time Markov chain model for estimating loss on accounts receivable. The intuition and appeal behind a Markov chain model for accounts receivable is that an account moves through different delinquency states each month. For example, an account in the “current” state this month will be in the “current” state next month if a payment is made by the due date and will be in the “30 days past due” state if no payment is received. Another valuable feature is that the Markov chain model maintains the progression and timing of events in the path from “current” to “loss.” For example, an account in the “current” state doesn’t suddenly become a “loss.” Instead, an account must progress monthly from the “current” state to the “30 days past due” state to the “90 days past due” state and so on until foreclosure activities are completed and the collateral assets are sold to pay the outstanding debt.

The transition matrix in the Markov chain represents the month-by-month movement of loans between delinquency classifications or states. Barkman [2] observes the transition matrix is often of interest as an accounting summary that evaluates loan quality or loan collection practice. The matrix elements are commonly referred to as “roll-rates” since they denote the probability that an account will move from one state to another in one month. The transition matrix is sometimes referred to as the “roll-rate matrix” or the “delinquency movement matrix” (DMM).

Another application of the Markov chain model in credit risk is introduced in Jarrow et al. [3]. Institutional investors can use a continuous-time Markov chain model that incorporates credit ratings to assess the risk of structured finance securities. The states of the Markov chain are the bond rating and the transition matrix reflects the likelihood of maintaining a rating or migrating to another rating level. The transition matrix is called the “migration matrix” in Gupton et al. [4] for CreditMetrics. The estimation of the continuous-time Markov chain transition probabilities is introduced in Fleming [5] and more recently in Monteiro et al. [6]. While the issue of correlation between issuers discussed in that research may be applicable to a portfolio of mortgages, a mortgage is a simple accounts receivable discrete-time Markov chain model with no arbitrage or hedging opportunities that require more complicated model features.

The statistical problem of interest is to estimate the transition matrix using a sample of observed monthly loan movements between delinquency states. The estimation is complicated by the frequent observation in studies that the Markov chain is neither homogeneous
nor stationary. Betancourt [7] concluded repayment for Freddie Mac data on prime mortgages was neither homogeneous nor stationary, and estimated transition matrices produce poor forecasts. This paper proposes two innovative estimation methods for the transition matrix based on two observations that improve forecasting precision.

First, it is common to divide a portfolio of loans into segments, where all loans within a segment are similar and expected to have the same transition matrix. This practice may be based on characteristics of the financial product or on data mining. For example, fixed-rate loans are different in many ways from adjustable rate mortgage (ARM) loans so it seems reasonable to assume they will have different transition matrices. Cyert and Thompson [8] propose segments based on a credit score. However, there may not be enough observed transitions within each segment to provide accurate estimates of all transition probabilities.

Second, it is reasonable to incorporate covariates that may change a few of the transition probabilities for a given loan as it progresses through the repayment period. For example, the data suggests that loans are more likely to remain current as they age. This nonstationarity could be modeled with the covariate ‘number of months since last delinquency’ or ‘number of months since origination.’ Other covariates that may model differences between loans over the repayment period, are credit quality, repayment history, and loan age.

To address the first issue, segment transition probabilities are constructed by pooling data from loans in the same segment and borrowing strength from data in other segments. To address the second issue, using loan-level models for key transition probabilities allows the incorporation of covariates that result in different transition probabilities for different loans over the repayment period. Before describing the estimators, the notation and details of using a Markov chain model to produce forecasts by delinquency state for a portfolio is presented in Section 2. In Sections 3 and 4 Bayes and empirical Bayes estimates are described for applications where the transition matrices for several segments differ in some, but not all entries. Section 5 demonstrates how estimated loan-level models can be used in a few key transition probabilities. Note that different covariates can be used in the different loan-level models, as is demonstrated in an example where the “current” to “30 days past due” model uses a repayment behavior covariate, but the “current” to “paid” model uses the interest rate covariate. The empirical Bayes estimates and loan-level models for current-to-30 days past due and current-to-paid are applied to an example of forecasting a $7 billion portfolio of subprime fixed first mortgages in Section 6. One of the important questions regarding the applicability of the Markov chain model to forecasts is the effect of sampling variation in the transition probability estimates on the forecasts, since forecasts are products of transition
matrices. This question is addressed in Section 7 in a simulation study. This paper concludes with a discussion of the value of the Markov chain model for delinquency and ideas for future research in Section 8.

2 Using a Markov Chain Model to Produce Forecasts of Outstanding Balance in Each Delinquency State

Let \( \{X_n\} \) denote a Markov chain where \( X_n \) is the delinquency state of a loan in month \( n \). Let \( \pi(n) \) denote the unconditional probability distribution of a loan in month \( n \), and is a vector whose entries correspond to the different Markov chain delinquency states. If the delinquency state for month \( n \) is known, then \( \pi_i(n) \) is a row vector with a one indicating this month’s delinquency state for loan \( i \) and zeros elsewhere. The transition matrix moving from month \( n \) to month \( n + 1 \) of the Markov chain is denoted by \( P(n,n + 1) \), a matrix containing the probabilities of movement between delinquency states.

If the transition matrix is known, a forecast of the delinquency state probability distribution for next month can be formed given the previous month’s delinquency state probability distribution. That is, for loan \( i \), the delinquency state probability distribution of month \( n + 1 \) is computed from \( \pi_i(n + 1) = \pi_i(n)P_i(n,n + 1) \) if \( \pi_i(n) \) and \( P_i(n,n + 1) \) are known. See Matis et al. [9] for an example of forecasting cotton yields using a Markov chain model with changing transition matrix. Associated with loan \( i \) is an outstanding balance at month \( n \) denoted by \( w_i(n) \). The “one month ahead” forecast outstanding balance by delinquency state of loan \( i \) is the vector \( w_i(n + 1) = w_i(n) \cdot \pi_i(n + 1) \).

Consider a simple example of a Markov chain model for a loan where the states are defined as “current,” “delinquent,” “loss,” “paid.” Notice that if the transition matrix from the 24th month since origination \( (n = 24) \) to the 25th month since origination \( (n = 25) \) was

\[
P_i(24,25) = \begin{bmatrix}
0.95 & 0.04 & 0.00 & 0.01 \\
0.15 & 0.75 & 0.07 & 0.03 \\
0.00 & 0.00 & 1.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 1.00
\end{bmatrix}
\]

the matrix elements represent the probability of moving from a row state to a column state. The value 0.04 in the first row second column is the probability of moving from “current” this month to “delinquent” next month. Notice that the loss and paid state are absorbing
states since eventually all loans terminate in one of these two states.

Suppose loan $i$ is in the current state at $n = 24$ so that the delinquency state probability distribution is $\pi_i(24) = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$. The forecasted delinquency state probability distribution of loan $i$ for next month is

$$\pi_i(25) = \pi_i(24) \cdot P_i(24, 25) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.95 & 0.04 & 0.00 & 0.01 \\ 0.15 & 0.75 & 0.07 & 0.03 \\ 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix} = \begin{bmatrix} 0.95 & 0.04 & 0.00 & 0.01 \end{bmatrix}.$$ 

Suppose loan $i$ was originated as a $50,000 30$-year fixed-rate $12\%$ APR (1\% per month) loan with a monthly payment of $514.30$. If the borrower has made timely monthly payments for 24 months, then this month ($n = 24$) the loan is in the current state with an outstanding balance of $49,614.11$. The “one month ahead” forecast outstanding balance by delinquency state of loan $i$ next month ($n = 25$) is

$$w_i(25) = w_i(24) \cdot p_i(25) = w_i(24) \cdot p_i(24) \cdot P_i(24, 25) = (49,614.11) \cdot \begin{bmatrix} 0.95 & 0.04 & 0.00 & 0.01 \end{bmatrix} = \begin{bmatrix} 47,133.41 & 1,984.56 & 0 & 496.14 \end{bmatrix}.$$ 

While this forecast is on a specific loan, since the loan will be in only one of the delinquency states next month depending on what the borrower chooses to do, the forecast will clearly be wrong. However, if a collection of loan-level forecasts (for example, in a portfolio or a pool of securitized loans) is accumulated then the forecast balance by delinquency state should be close to the actual balance by delinquency state. That is, if $i = 1, \ldots, N$ indexes a portfolio of loans where each loan is $n_i$ months from origination this month, then the “one month ahead” portfolio delinquency balance forecast is

$$\sum_{i=1}^{N} w_i(n_i + 1) = \sum_{i=1}^{N} w_i(n_i) \cdot p_i(n_i + 1) = \sum_{i=1}^{N} w_i(n_i) \cdot p_i(n_i) \cdot P_i(n_i, n_i + 1).$$ 

If the Markov chain is stationary and homogeneous then all loans will have the same
transition matrix and the calculation will not need to be performed on each loan. However, the emphasis of this paper is on applications in which loans are not homogeneous and transition probabilities are known to change as the loan ages.

One appealing feature of the Markov chain model is that the forecast maintains the timing of losses. This is demonstrated by the forecast of the delinquency state distribution in two months. Notice that for loan $i$ at month $n$, the “two month ahead” delinquency state probability distribution is $p_i(n + 2)$. Since

$$
\pi_i(n + 2) = \pi_i(n) P_i(n, n + 2) = \pi_i(n) P_i(n, n + 1) P_i(n + 1, n + 2) = \pi_i(n + 1) P_i(n + 1, n + 2),
$$

the “two month ahead” delinquency state probability distribution is the “one month ahead” distribution multiplied by the transition matrix from $n + 1$ to $n + 2$. The “two month ahead” forecast outstanding balance by delinquency state of loan $i$ is $w_i(n + 2) = w^*_i(n + 1) \cdot p_i(n + 2)$, where $w^*_i(n + 1)$ differs from $w_i(n + 1)$ because movement from a current state to a current state or from a delinquent state to a current state next month indicates that loan payments were made and the outstanding balance is reduced by one month’s principal payment.

Continuing the example of loan $i$ that is in the current state at $n = 24$, the ‘two month ahead’ delinquency state probability distribution is

$$
\pi_i(26) = \pi_i(25) P_i(25, 26)
$$

\[
= \begin{bmatrix}
0.95 & 0.04 & 0.00 & 0.01 \\
0.15 & 0.75 & 0.07 & 0.03 \\
0.00 & 0.00 & 1.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 1.00
\end{bmatrix}
\]

\[
= \begin{bmatrix}
0.9180 & 0.0585 & 0.0028 & 0.0207
\end{bmatrix}
\]

Notice $P_i(25, 26)$ differs from $P_i(24, 25)$ indicating that if the loan is still current at this point in the loan age it is more likely to remain current and less likely to become delinquent.

To demonstrate the details of the outstanding balance, continue the example of loan $i$, a $50,000 30-year fixed-rate 12\%$ APR loan whose borrower has made timely monthly payments for 24 months. While $w_i(24) = 49,614.11$, assuming timely payment last month then $w^*_i(25) = 49,595.94$. This results in the ‘two month ahead’ forecast outstanding balance by delinquency state is

$$
w_i(26) = w^*_i(25) \cdot p_i(26)
$$
It should be noted that there are other ways to handle the loan amortization component. One is to recognize that delinquent loans will remain at the previous outstanding balance plus interest. Another is to value loans based on the market value of a delinquent loan with a given contract value and rate. This detail can be applied at the loan level before aggregating the forecast for a collection of loans.

Throughout this simple example, only a few states are defined for the Markov chain. In applications, the choice of the number and identity of the states is determined by the monthly delinquency steps to loss and the desired reporting detail. Table 1 presents an example transition matrix that demonstrates this point. In this example, loans are declared loss after 120 days of delinquency. The ‘0’ state indicates current loans (borrower’s whose monthly payment was received by the due date). The ‘30,’ ‘60,’ ‘90,’ and ‘120’ states are for that are delinquent (1-30, 31-60, 61-90, 91-120 days past due). In this application, once a loan becomes more than 150 days past due it is carefully analyzed to determine if it should be declared a loss. The evaluation of a seriously delinquent loan may take several months while loss mitigation and foreclosure options are pursued. There is no value to creating additional monthly delinquency states just because the loan advances farther into delinquency. The ‘150+’ state contains all these seriously delinquent loans. Another interesting feature of the states in this example is the 0, 30, 60, 90, 120, 150+ states are duplicated, with one collection falling under the ‘Never Bankrupt’ label and the other called ‘Bankrupt.’ This was required since the lender required separate forecasts for borrowers who had filed personal bankruptcy protection, and the model would allow borrowers to transition from the ‘Never Bankrupt’ group to the ‘Bankrupt’ group.

Other definitions of states that may prove useful in other applications. For example, if loans enter foreclosure proceedings after 60 days past due, choose the states: current, 30 days past due, 60 days past due, enter foreclosure, 1, 2, 3+ mo in foreclosure, loss, paid. Another example is when differences are known between current accounts based on the repayment history, choose the states: 1, 2, 3, 4, 5, 6+ mo since last delinquency, delinquent, loss, paid.
Embedded in the transition matrix in Table 1 is a wealth of information surrounding both the quality of the loans and the practices of the loan collection or servicing. For example the probabilities associated with the 0, 30 and 60 states reflect behavior of loans in early stages of delinquency. While only 94.6% of loans stay current, aggressive action on loans 30 days past due brings 29.5% of them back to current. It is interesting to note that for borrowers in bankruptcy the principles are the same but the probabilities are lower. Frequently the transition probabilities reflecting highly delinquent loans becoming less delinquent is an accounting summary used to evaluate loss mitigation efforts such as payment plans (in which the borrower agrees to make increased payments for a time to bring the loan current). Also of interest are the probabilities that transition to the ‘Paid’ state. These are the single month mortality (SMM) used in modeling prepayment. As an aside, the conditional prepayment rate (CPR) is the annualized SMM. Notice that current loans have higher prepayment probability than delinquent loans, which is to be expected as borrowers improve their credit by demonstrating payment on a subprime mortgage and then exercising the opportunity to refinance to a less risky mortgage.

### 3 Bayesian Estimation

The goal of segmentation is to create groups in which some of the transition probabilities differ. While it may be easy to find segments with differences in a few transition probabilities, it is difficult to identify segments with differences in all the entries of the transition matrix. For those transition probabilities that are the same between segments, the segment estimator is inefficient compared to the estimator based on the entire sample. Another practical difficulty is that a segment must contain enough loans to estimate less frequent transition probabilities like the “120 days past due” to “150 days past due” and “150 days past due” to “current.” Some valuable segmentation schemes may not be possible because some segments may have few observations for some transition probabilities.

Anderson and Goodman [10] show that the maximum likelihood estimator of the transition probabilities \( p(j, k) \) when the Markov chain is stationary is the fraction of the number of observed transitions from state \( j \) to state \( k \) over the total number of observations beginning in state \( j \). There is confusion in the literature about whether the transition matrix should be estimated by the movement of dollars or accounts between delinquency states. Cyert \textit{et al.} [1] and van Kuelen \textit{et al.} [11] propose dollar estimation, while others use frequency estimation. To demonstrate the distinction between dollar and frequency estimation, consider the data
Table 1. Example Transition Matrix for an Application with Mortgage Delinquency. The states allow monthly transitions to delinquency and provide the required reporting detail. Transition probabilities are estimated from a sub-prime portfolio of first lien fixed-rate 30-year mortgages secured by residential real estate assuming stationarity (likely inappropriate).

<table>
<thead>
<tr>
<th>FROM:</th>
<th>Never Bankrupt</th>
<th></th>
<th>Bankrupt</th>
<th></th>
<th>Loss</th>
<th></th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>30</td>
<td>60</td>
<td>90</td>
<td>120</td>
<td>150+</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0.946</td>
<td>0.042</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>30</td>
<td>0.295</td>
<td>0.602</td>
<td>0.079</td>
<td>0.006</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>60</td>
<td>0.248</td>
<td>0.249</td>
<td>0.300</td>
<td>0.150</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>90</td>
<td>0.168</td>
<td>0.159</td>
<td>0.136</td>
<td>0.255</td>
<td>0.179</td>
<td>0.010</td>
<td>0.011</td>
</tr>
<tr>
<td>120</td>
<td>0.055</td>
<td>0.014</td>
<td>0.007</td>
<td>0.045</td>
<td>0.749</td>
<td>0.021</td>
<td>0.001</td>
</tr>
<tr>
<td>150+</td>
<td>0.127</td>
<td>0.047</td>
<td>0.024</td>
<td>0.013</td>
<td>0.011</td>
<td>0.648</td>
<td>0.001</td>
</tr>
<tr>
<td>BK 0</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>BK 30</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>BK 60</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>BK 90</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>BK 120</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>BK 150+</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Loss</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Paid</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
DMM Forecasting

on the movement of 401 loans with an outstanding balance of $27,358,745 that begin the month in the ‘30 days past due’ state. The dollar estimation is based on the distribution of outstanding balances, meaning that the estimate for the ‘30 days past due’ would use

<table>
<thead>
<tr>
<th></th>
<th>Current</th>
<th>30 days past due</th>
<th>60 days past due</th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days</td>
<td>$14,116,167</td>
<td>$4,248,055</td>
<td>$8,866,529</td>
<td>$127,994</td>
</tr>
</tbody>
</table>

to produce the transition probability estimates for the ‘30 days past due’ row

<table>
<thead>
<tr>
<th></th>
<th>Current</th>
<th>30 days past due</th>
<th>60 days past due</th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days</td>
<td>0.5160</td>
<td>0.1553</td>
<td>0.3241</td>
<td>0.0047</td>
</tr>
</tbody>
</table>

In comparison, the frequency estimation is based on the the distribution of loans, so the tally is based on loans not dollars, and the data is

<table>
<thead>
<tr>
<th></th>
<th>Current</th>
<th>30 days past due</th>
<th>60 days past due</th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days</td>
<td>215</td>
<td>57</td>
<td>127</td>
<td>2</td>
</tr>
</tbody>
</table>

and would produce the estimate

<table>
<thead>
<tr>
<th></th>
<th>Current</th>
<th>30 days past due</th>
<th>60 days past due</th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 days</td>
<td>0.5362</td>
<td>0.1421</td>
<td>0.3167</td>
<td>0.0050</td>
</tr>
</tbody>
</table>

While dollar estimation appears to have an impressive ‘sample size’ over 27 million, in fact this is not 27 million $1 loans. Using the notation that \( w_i(n) \) is the outstanding balance in month \( n \) for the \( i \)th loan, when the \( i \)th loan moves in the Markov chain model all the \( w_i(n) \) dollars move. It can be shown that the frequency and dollar estimators are both unbiased for the transition matrix, but the dollar estimator is inefficient since its variance is greater than or equal to the frequency estimator variance because of the variation in the \( w_i(n) \) for a given portfolio. It should be emphasized that the choice of frequency or dollar estimation does not depend on the desire to forecast the balance in each delinquency state since either method estimates the transition matrix and the outstanding balance is produced by incorporating the \( w_i(n) \) in the Markov chain forecast as demonstrated in the previous section.

Bayesian methods offer a formal approach for incorporating an expert’s beliefs into the analysis. In estimating the transition matrix of a Markov chain for loan repayment, the prior could represent an expert’s belief about the future loan servicing practices or a change from the observed economic or lending environment. Lee et al. [12] demonstrate the conjugate prior for a homogeneous Markov chain and the resulting Bayes transition matrix estimator.

Consider estimating \( P_1, P_2, \ldots, P_L \), which denote the stationary transition matrices for \( L \)
segments. The data available to estimate the segment transition matrices are the observed monthly account movements. Let \( f_\ell(j, k) \) denote the number of accounts in segment \( \ell \) that started the month in state \( j \) and moved to state \( k \) the next month. Model the row vector of observed monthly movements in segment \( \ell \), denoted by

\[
f_\ell(j) = [f_\ell(j, 1), f_\ell(j, 2), \ldots, f_\ell(j, K)]
\]
as a multinomial distribution. That is, the \( n_\ell(j) = \sum_k f_\ell(j, k) \) accounts that start the month in state \( j \) move to the possible delinquency states according to the \( j \)th row of the \( \ell \)th segment’s transition matrix, denoted by

\[
p_\ell(j) = [p_\ell(j, 1), p_\ell(j, 2), \ldots, p_\ell(j, K)],
\]
where \( \sum_k p_\ell(j, k) = 1 \). This model gives the likelihood function

\[
\mathcal{L}(f_\ell(j) \mid p_\ell(j)) = \frac{n_\ell(j)!}{f_\ell(j, 1)! \cdots f_\ell(j, K)!} p_\ell(j, 1)^{f_\ell(j, 1)} \cdots p_\ell(j, K)^{f_\ell(j, K)}.
\]

The conjugate prior for the multinomial is the Dirichlet distribution, whose parameters are denoted by \( \alpha(j) = [\alpha(j, 1), \alpha(j, 2), \ldots, \alpha(j, \ell)] \) with \( j \) indexing the different rows of the transition matrix. Suppose \( p_\ell(j) \) has prior distribution

\[
\pi(p_\ell(j) \mid \alpha(j)) = \frac{\Gamma[M(j)]}{\Gamma[\alpha(j, 1)] \cdots \Gamma[\alpha(j, K)]} p_\ell(j, 1)^{\alpha(j, 1)+1} \cdots p_\ell(j, K)^{\alpha(j, K)+1},
\]
where \( \Gamma(\cdot) \) is the gamma function and \( M(j) = \sum_k \alpha(j, k) \). This prior has the beta distribution as a special case, corresponding to only two delinquency states when the binomial distribution would be applicable.

A non-informative prior is nonsensical since transition probabilities are not equally likely to occur. A reasonable practice for each row of the transition matrix is to allow the expert to distribute \( M(j) \) loans according to their knowledge of loan quality and servicing. These values are \( \alpha(j, k) \). To demonstrate, Table 2 presents the \( \alpha(j, k) \) assigned by an expert for a subprime mortgage portfolio. In the top table, the expert believes most loans will remain current, and of those that become delinquent the loan servicing is aggressive and effective, with loans more likely to become less delinquent or prepay. In the bottom table, the expert has represented repayment behavior in harsher economic times where borrowers who become delinquent won’t have the resources to return to current and few refinancing
options to avoid foreclosure. At first glance it may be surprising to see non-zero values in Table 2 for transitions such as ‘90 days past due’ to ‘30 days past due,’ but these transitions correspond to borrowers who make a payment only large enough to pay part of the amount due but not enough to become current.

It is useful in the interpretation and computation to reparameterize the Dirichlet in the spirit of a reparameterization of the beta distribution. Define \( \mu(j,k) = \alpha(j,k)/M(j) \) and write the prior distribution as

\[
\pi(p_\ell(j) \mid \mu(j), M(j)) = \frac{\Gamma[M(j)]}{\Gamma[M(j)\mu(j,1)] \cdots \Gamma[M(j)\mu(j,K)]} \cdot p_\ell(j,1)^{M(j)\mu(j,1)+1} \cdots p_\ell(j,K)^{M(j)\mu(j,K)+1}.
\]

In this reparameterization,

\[
E[p_\ell(j,k)] = \mu(j,k) \quad \text{and} \quad V[p_\ell(j,k)] = \frac{\mu(j,k)[1 - \mu(j,k)]}{M(j) + 1}.
\]

The \( \mu(j,k) \) parameters represent the mean transition probability from state \( j \) to state \( k \) over all the segments.

The posterior distribution is

\[
p(p_\ell(j) \mid f_\ell(j)) = \frac{\Gamma[n_\ell(j) + M(j)]}{\Gamma[f_\ell(j,1) + M(j)\mu(j,1)] \cdots \Gamma[f_\ell(j,K) + M(j)\mu(j,K)]} \cdot p_\ell(j,1)^{f_\ell(j,1)+M(j)\mu(j,1)-1} \cdots p_\ell(j,K)^{f_\ell(j,K)+M(j)\mu(j,K)-1},
\]

which has a Dirichlet distribution. Under squared-error loss, the Bayes estimator of \( p_\ell(j,k) \) is given by

\[
p_\ell^B(j,k) = \frac{f_\ell(j,k) + M(j)\mu(j,k)}{n_\ell(j) + M(j)} = \left[ \frac{n_\ell(j)}{n_\ell(j) + M(j)} \right] \left[ \frac{f_\ell(j,k)}{n_\ell(j)} \right] + \left[ \frac{M(j)}{n_\ell(j) + M(j)} \right] \mu(j,k).
\]

4 Empirical Bayes Estimation

This section defines empirical Bayes (EB) estimators for segmented transition matrices. While some may prefer a fully Bayesian estimator, in the current financial environment with
Table 2. Elicited Dirichlet $\alpha(j, k)$ Priors for the Transition Matrix for a Subprime Mortgage Portfolio. The top table is an optimistic scenario with successful loan servicing where most borrowers are believed to stick in delinquency states or get better and the timing of losses can be managed. The bottom table is a pessimistic scenario where most borrowers proceed to higher delinquency states and have few refinancing options to avoid foreclosure. The states of the Markov chain model are: ‘0’ for current accounts; ‘30,’ ‘60,’ ‘90,’ ‘120’ denote 1 to 30 days past due, 31 to 60 days past due, 61 to 90 days past due, 91 to 120 days past due, respectively; ‘150+’ denotes more than 150 days past due; ‘Loss’ denotes accounts where the account is closed due to nonpayment; and ‘Paid’ denotes accounts where the borrower paid off the mortgage.

<table>
<thead>
<tr>
<th>FROM</th>
<th>Days past due:</th>
<th>Loss</th>
<th>Paid</th>
<th>M(j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4,850,000 125,000 0 0 0 0 0 0 0 25,000 5,000,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>2,500 6,400 1,000 0 0 0 0 0 100 10,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>1,000 1,500 1,400 1,000 0 0 50 50 5,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>525 375 600 720 600 0 150 30 3,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>100 40 40 100 1,400 100 200 20 2,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150+</td>
<td>1,000 200 200 200 200 7,000 1,000 200 10,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss</td>
<td>0 0 0 0 0 0 1 0 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paid</td>
<td>0 0 0 0 0 0 0 1 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FROM</th>
<th>Days past due:</th>
<th>Loss</th>
<th>Paid</th>
<th>M(j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4,625,000 350,000 0 0 0 0 0 0 25,000 5,000,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1,000 2,500 6,500 0 0 0 0 0 50 10,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0 0 1,000 4,000 0 0 0 0 5,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>0 0 0 300 2,700 0 0 0 3,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>0 0 0 0 100 1,900 0 0 2,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150+</td>
<td>0 0 0 0 0 6,500 3,500 0 10,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss</td>
<td>0 0 0 0 0 0 1 0 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paid</td>
<td>0 0 0 0 0 0 0 1 1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
 Basel II requirements for model validation, it may appeal to auditors that the EB transition matrix is a weighted average of the segment’s transition matrix and the grand mean transition matrix estimated ignoring segmentation. Others may appreciate the argument in Efron’s 2004 Fisher lecture that empirical Bayes is a thoughtful compromise between frequentist and Bayesian perspectives. Regardless of the motivation, one interesting feature of the EB transition matrix estimator is that the weights are dictated by the value of segmentation in the data.

If the Markov chain were stationary, the EB estimator derived in Billard and Meshkani [13] for the transition matrix could be applied. However, Betancourt [7] found poor forecasts on Freddie Mac prime mortgages using a homogeneous transition matrix and tried segmenting into five groups based on origination loan-to-value ratio (LTV) to improve the short-term forecast performance. As discussed in the previous section, the data mining for segments that improve forecasts is enticing but one would hardly expect to find segments whose transition matrices differ for every possible transition. Instead, the exploration of segments needs to maintain the properties of the Bayes estimator where truly different transition probabilities between segments are well estimated but transition probabilities that are the same between segments borrow strength across segments.

As described in Carlin and Louis [14], the basic empirical Bayes approach uses the observed data to estimate the prior parameters. The parametric EB estimator uses the standard Bayes estimator where the prior parameters \((\mu(j), M(j))\) are replaced by the estimates \((\hat{\mu}(j), \hat{M}(j))\), which maximizes the marginal likelihood \(L(f_1(j), \ldots, f_L(j) | \mu(j), M(j))\), viewed as a function of \((\mu(j), M(j))\).

The marginal likelihood is

\[
L(f_1(j), \ldots, f_L(j) | \mu(j), M(j)) = \int \cdots \int f(f_1(j), \ldots, f_L(j), p_1(j), \ldots, p_L(j) | \mu(j), M(j)) \, dp_1(j) \cdots dp_L(j)
\]

\[
= \prod_\ell \int f(f_\ell(j), p_\ell(j) | \mu(j), M(j)) \, dp_\ell(j)
\]

\[
= \prod_\ell \frac{n_\ell(j)!}{f_\ell(j, 1)! \cdots f_\ell(j, K)!} \cdot \frac{\Gamma[M(j)]}{\Gamma[M(j)\mu(j, 1)] \cdots \Gamma[M(j)\mu(j, K)]}
\cdot p_\ell(j, 1)^{f_\ell(j, 1)+M(j)\mu(j, 1)-1} \cdots p_\ell(j, K)^{f_\ell(j, K)+M(j)\mu(j, K)-1} \, dp_\ell(j)
\]

\[
= \prod_\ell \frac{n_\ell(j)}{f_\ell(j, 1)! \cdots f_\ell(j, K)!} \cdot \frac{\Gamma[M(j)]}{\Gamma[n_\ell(j) + M(j)]}
\]
When maximized and viewed as a function of \((\mu(j), M(j))\) it is easier to consider the log marginal likelihood, given by
\[
\ln L(\mu(j), M(j)) = \sum_\ell [\ln n_\ell(j)! - \ln f_\ell(j, 1)! - \cdots - \ln f_\ell(j, K)!] \\
+ L \ln(\Gamma[M(j)]) - \sum_\ell \ln(\Gamma[n_\ell(j) + M(j)]) \\
+ \sum_\ell \ln(\Gamma[f_\ell(j, 1) + M(j)\mu(j, 1)]) - L \ln(\Gamma[M(j)\mu(j, 1)]) \\
+ \cdots + \sum_\ell \ln(\Gamma[f_\ell(j, K) + M(j)\mu(j, K)]) - L \ln(\Gamma[M(j)\mu(j, K)]).
\]

The estimates \((\hat{\mu}(j), \hat{M}(j))\), which maximize the log marginal likelihood, can be found using non-linear optimization involving the digamma function for the derivative of the log of the gamma function; however, it seems intuitive that \(p(j, k) = \sum_\ell f_\ell(j, k)/\sum_\ell n_\ell(j)\) should be extremely close to the optimal estimator \(\hat{\mu}(j, k)\), if it isn’t exactly equal. The parameter \(\mu(j, k)\) represents \(E[p_\ell(j, k)]\), and \(p(j, k)\) is the proportion of accounts in all segments that started the month in state \(j\) and moved to state \(k\). Since \(p(j, k)\) is simple to compute, the optimization simplifies from a nonlinear optimization over \(K + 1\) space to a one-dimensional optimization.

Conditional on \(\hat{\mu}(j, k) = p(j, k)\), the log marginal likelihood becomes
\[
\ln L(M(j)) = \sum_\ell [\ln n_\ell(j)! - \ln f_\ell(j, 1)! - \cdots - \ln f_\ell(j, K)!] \\
+ L \ln(\Gamma[M(j)]) - \sum_\ell \ln(\Gamma[n_\ell(j) + M(j)]) \\
+ \sum_\ell \ln(\Gamma[f_\ell(j, 1) + M(j)p(j, 1)]) - L \ln(\Gamma[M(j)p(j, 1)]) \\
+ \cdots + \sum_\ell \ln(\Gamma[f_\ell(j, K) + M(j)p(j, K)]) - L \ln(\Gamma[M(j)p(j, K)]).
\]

If there is at least one state where \(p(j, k) > 0\) and at least one segment \(\ell\) such that \(n_\ell(j) \geq 2\) with at least one state with \(f_\ell(j, k) \geq 2\) then the log marginal likelihood is concave. The proof uses the digamma function, \(\psi(x) = \frac{d}{dx} \ln \Gamma(x)\) and the property of the digamma function that \(\psi(x + 1) = \psi(x) + 1/x\). Taking the derivative with respect to
\[ \frac{d}{d M(j)} \ln \mathcal{L}(M(j)) = L \cdot \psi[M(j)] - \sum_{\ell} \psi[n_{\ell}(j) + M(j)] \\
+ \sum_{\ell} \psi[f_{\ell}(j, 1) + M(j)p(j, 1)] \cdot p(j, 1) - L \cdot \psi[M(j)p(j, 1)] \cdot p(j, 1) \\
+ \cdots + \sum_{\ell} \psi[f_{\ell}(j, K) + M(j)p(j, K)] \cdot p(j, K) - L \cdot \psi[M(j)p(j, K)] \cdot p(j, K) \\
= \sum_{\ell} \{ \psi[M(j)] - \psi[n_{\ell}(j) + M(j)] \} \\
+ \sum_{k} \sum_{\ell} p(j, k) \cdot \{ \psi[f_{\ell}(j, k) + M(j)p(j, k)] - \psi[M(j)p(j, k)] \} \\
= \sum_{\{\ell: n_{\ell}(j)=0\}} \{ \psi[M(j)] - \psi[n_{\ell}(j) + M(j)] \} \\
+ \sum_{\{\ell: n_{\ell}(j)=1\}} \{ \psi[M(j)] - \psi[n_{\ell}(j) + M(j)] \} \\
+ \sum_{\{\ell: n_{\ell}(j)\geq2\}} \{ \psi[M(j)] - \psi[n_{\ell}(j) + M(j)] \} \\
+ \sum_{\{k: p(j, k)>0\}} \left\{ \sum_{\{\ell: f_{\ell}(j, k)=0\}} p(j, k) \cdot \{ \psi[f_{\ell}(j, k) + M(j)p(j, k)] - \psi[M(j)p(j, k)] \} \\
+ \sum_{\{\ell: f_{\ell}(j, k)=1\}} p(j, k) \cdot \{ \psi[f_{\ell}(j, k) + M(j)p(j, k)] - \psi[M(j)p(j, k)] \} \\
+ \sum_{\{\ell: f_{\ell}(j, k)\geq2\}} p(j, k) \cdot \{ \psi[f_{\ell}(j, k) + M(j)p(j, k)] - \psi[M(j)p(j, k)] \} \right\} \\
= 0 \\
+ \sum_{\{\ell: n_{\ell}(j)=1\}} \left\{ \psi[M(j)] - \psi[M(j)] - \frac{1}{M(j)} \right\} \\
+ \sum_{\{\ell: n_{\ell}(j)\geq2\}} \left\{ \psi[M(j)] - \psi[M(j)] - \frac{1}{M(j)} - \frac{1}{M(j) + 1} \\
- \cdots - \frac{1}{M(j) + n_{\ell}(j) - 2} - \frac{1}{M(j) + n_{\ell}(j) - 1} \right\} \]
\[
\begin{align*}
+ & \sum_{\{k: \overline{p}(j,k) > 0\}} \left\{ 0 \right\} \\
+ & \sum_{\{\ell: f_{\ell}(j,k)=1\}} \overline{p}(j,k) \cdot \left\{ \psi[M(j)\overline{p}(j,k)] + \frac{1}{M(j)\overline{p}(j,k)} - \psi[M(j)\overline{p}(j,k)] \right\} \\
+ & \sum_{\{\ell: f_{\ell}(j,k) \geq 2\}} \overline{p}(j,k) \cdot \left\{ \psi[M(j)\overline{p}(j,k)] + \frac{1}{M(j)\overline{p}(j,k)} + \frac{1}{M(j)\overline{p}(j,k) + 1} + \cdots \right. \\
& \quad + \left. \frac{1}{f_{\ell}(j,k) + M(j)\overline{p}(j,k) - 2} \right. \\
& \quad + \left. \frac{1}{f_{\ell}(j,k) + M(j)\overline{p}(j,k) - 1} \right. \\
& \quad - \psi[M(j)\overline{p}(j,k)] \right\} \\
= & \sum_{\{\ell: n_{\ell}(j)=1\}} \left\{ -\frac{1}{M(j)} \right\} \\
+ & \sum_{\{\ell: n_{\ell}(j) \geq 2\}} \left\{ -\frac{1}{M(j)} - \frac{1}{M(j) + 1} \right. \\
& \quad - \cdots - \frac{1}{M(j) + n_{\ell}(j) - 2} - \frac{1}{M(j) + n_{\ell}(j) - 1} \right\} \\
+ & \sum_{\{k: \overline{p}(j,k)>0\}} \left\{ \sum_{\{\ell: f_{\ell}(j,k)=1\}} \overline{p}(j,k) \cdot \left\{ \frac{1}{M(j)\overline{p}(j,k)} \right\} \\
& \quad + \sum_{\{\ell: f_{\ell}(j,k) \geq 2\}} \overline{p}(j,k) \cdot \left\{ \frac{1}{M(j)\overline{p}(j,k)} + \frac{1}{M(j)\overline{p}(j,k) + 1} + \cdots \right. \\
& \quad \left. + \frac{1}{f_{\ell}(j,k) + M(j)\overline{p}(j,k) - 2} \right. \\
& \quad + \left. \frac{1}{f_{\ell}(j,k) + M(j)\overline{p}(j,k) - 1} \right\} \right\} \\
= & \sum_{\{k: \overline{p}(j,k)>0\}} \left\{ \sum_{\{\ell: f_{\ell}(j,k)=1\}} \overline{p}(j,k) \cdot \left[ \frac{1}{M(j)\overline{p}(j,k)} - \frac{1}{M(j)} \right] \right\} \\
& \quad + \sum_{\{\ell: f_{\ell}(j,k) \geq 2\}} \left\{ \overline{p}(j,k) \cdot \left[ \frac{1}{M(j)\overline{p}(j,k)} - \frac{1}{M(j)} \right] \\
& \quad + \overline{p}(j,k) \cdot \left[ \frac{1}{M(j)\overline{p}(j,k) + 1} - \frac{1}{M(j) + 1} \right] \right. \\
& \quad + \cdots \right. \\
& \quad \left. + \cdots \right\}
\end{align*}
\]
if there is at least one state $k$ such that $\bar{p}(j, k) > 0$ and at least one segment $\ell$ such that $n_\ell(j) \geq 2$ with $f_\ell(j, k) \geq 2$.

A fairly efficient algorithm that takes advantage of the fact that the log marginal likelihood is easy to evaluate, is to begin with a small value of $M(j)$, such as one, and evaluate $\ln \mathcal{L}(M(j))$. The search continues by incrementing $M(j)$ by a small amount $\Delta$, for example one, evaluating $\ln \mathcal{L}(M(j))$ at the new value of $M(j)$, and comparing the new value to the previous value. When a value of $M(j)$ decreases $\ln \mathcal{L}(M(j))$ the previous value is $\hat{M}(j)$.

After computing the estimates $(\hat{\mu}(j), \hat{M}(j))$, the EB estimator of each transition probability in the $j$th row of the transition matrix for segment $\ell$ is given by

$$
\hat{p}_\ell(j, k) = \left[ \frac{n_\ell(j)}{n_\ell(j) + M(j)} \right] \left[ \frac{f_\ell(j, k)}{n_\ell(j)} \right] + \left[ \frac{\hat{M}(j)}{n_\ell(j) + \hat{M}(j)} \right] \hat{\mu}(j, k)
$$

$$
= \left[ \frac{n_\ell(j)}{n_\ell(j) + M(j)} \right] \hat{p}_\ell(j, k) + \left[ \frac{\hat{M}(j)}{n_\ell(j) + \hat{M}(j)} \right] \bar{p}(j, k).
$$
Notice this is a weighted average of the segment’s transition probability, denoted by $\hat{p}_\ell(j, k) = \frac{f_\ell(j, k)}{n_\ell(j)}$, and the grand mean transition probability, denoted by

$$\bar{p}(j, k) = \frac{\sum_\ell f_\ell(j, k)}{\sum_\ell n_\ell(j)}.$$ 

This weighted average estimator is very attractive. At one extreme is the estimator $\bar{p}(j, k)$, which assumes all segments have the same value for the transition probability. If this is true then $\bar{p}(j, k)$ is the most efficient estimator. At the other extreme is the segmented estimator $\hat{p}(j, k)$, which assumes each segment may have a different value for the transition probability. If this is true then the only unbiased estimator is $\hat{p}(j, k)$ since it is based solely on the loans in a given segment.

Reality is likely to be somewhere in between these two extremes. For example, when a segmentation is proposed for the entire transition matrix there will be some probabilities that are different between the segments but there will also be many probabilities which are the same for all segments. The EB transition matrix estimate chooses the weight based on what the data in the segmentation indicates.

The weight in the segmented EB transition matrix estimator depends on the relative size of $\hat{M}(j)$ and $n_\ell(j)$. When $\hat{M}(j)$ is small the variance between the segment transition probabilities is large. This indicates that the segmentation provides evidence of differences in the transition probabilities and the weight should favor the segment estimator $\hat{p}(j, k)$. When $\hat{M}(j)$ is large the variance between the segment probabilities is small. When all the segment transition probabilities in a row are identical, the conditional likelihood of $M(j)$ is maximized taking $M(j) \rightarrow \infty$. The case $M(j) \rightarrow \infty$ corresponds to the Dirichlet prior, yielding a point mass at $\hat{\mu}(j) = \bar{p}(j)$. In addition to being an interesting diagnostic of a proposed segmentation scheme, this result yields an upper bound for $M(j)$ in the maximization of the conditional likelihood. Choose as the upper bound a value of $M(j)$ which makes the weight on the $\bar{p}(j, k)$ component of the EB estimator sufficiently close to one.

One practical feature of the EB transition matrix estimator is that segmentations can be explored with some protection against the practical worry that there won’t be enough data to estimate some transition probabilities. The EB transition matrix estimator will return the grand mean transition matrix estimate, estimated from the data in all segments, if a segmentation is not useful or doesn’t have sufficient data. The Bayes Information Criterion could be used to compare different segmentation schemes.
5 Loan-Level Models

While many papers demonstrate the transition matrix is non-stationary, some of the estimation approaches are quite extreme. Diggle et al. [15] describes a general approach to modeling longitudinal categorical data as a Markov chain where the transition matrices $P(n, n + 1)$ have multinomial logistic models for each row with $n$ as a covariate. Once a multinomial model is introduced for each row of the transition matrix it is possible to include other covariates. Smith and Lawrence [16] model all transition probabilities, but this extreme is unnecessary for modeling mortgage repayment since only a few key transition probabilities have non-constant probabilities. For example, the left panel of Figure 1 presents the proportion of loans in a similar credit risk group that were “150+ days past due” in one month and moved to “loss” the next month at different loan ages (months since origination). While there is some variation, the scatter.smooth line indicates this particular transition probability appears constant over the life of a loan and it is acceptable to pool data regardless of loan age to form an estimate. In contrast, the right panel of Figure 1 presents the proportion of loans that were “current” in one month and moved to “30 days past due” the next month for a similar credit risk group at different loan ages. The pattern is that the default risk changes over loan lifetime, with the risk being very low after origination, reaching a peak and then plateauing (or possibly declining for older loans). The “current” to “30 days past due” transition probability changes over loan lifetime and there must be some way to include this feature in $P_i(n, n + 1)$. This section describes how loan-level models containing important explanatory variables can be estimated for a few key transition probabilities and inserted into the transition matrix.

Suppose $P_i(n, n + 1) = P(x_i; \theta)$, meaning that changes in the transition matrix over $n$ as well as credit risk, loan characteristics, and collateral information on the $i$th loan are represented by a set of explanatory variables. The extreme of estimating every row of the transition matrix with a multinomial model, as done in Smith and Lawrence [16], has three major concerns. First, not every transition probability merits modeling; for example, in the row corresponding to loans “current” this month, while the probabilities of moving to “30 days past due” and moving to “prepay” should be modeled, there is no value to modeling the “current” to “60 days past due” (a movement that appears to occur only on accounts whose previous month’s payment check bounced) or the “current” to “loss” probabilities—these calculations are unnecessary and would suffer the bias of overfitting even if a Bayesian approach to estimation such as described in Sung et al. [17] and Healy and Engler [18].
Figure 1: The left panel is the proportion of loans 150+ days past due one month and loss
the next month over loan lifetime. This transition probability appears constant over time.
The right panel is the proportion of loans current one month and 30 days past due the next
month over loan lifetime. This transition probability clearly changes over time.

Second, the explanatory variables may differ between transition probabilities; for example,
again considering the row for loans “current” this month, the explanatory variables for the
“current” to “30 days past due” probability are measures of credit and repayment history,
outstanding loan balance, and market value of the collateral, but the explanatory variables for
the “current” to “prepay” probability are an evaluation of the borrower’s interest rate and the
market interest rate available if the borrower were to refinance. Third, it is overly complicated
to model all the transition probabilities except in overly simple models; for example, the
simple model of Smith and Lawrence [16] requires 8 estimated transition probabilities, but
a Markov model with “current,” “30,” “60,” “90,” “120 or more days past due,” “loss,” and
“paid” states has 35 transition probabilities.

This paper proposes that the multinomial probability for any row of the transition matrix
can be estimated by a collection of binary logistic models using the results from Begg and
Gray [19]. Let \( p(j, k) \) for \( k = 1, \ldots, K \) denote the probabilities of moving from the \( j \)th state
this month to the \( k \)th state next month. Writing the multinomial logistic for the \( j \)th row as
separate binary logistic models yields the collection

\[
\Pr\{\text{state } k \text{ next month} \mid \text{state } j \text{ this month}\} = \exp(x_j' \beta_{j,k}), \quad k = 1, \ldots, K - 1.
\]
To introduce a loan-level model for the transition probability $p(j, k^*)$ with explanatory variables $x_{j,k^*}$, use a sample of loans that are in state $j$ this month and move next month to either state $k^*$ ($Y = 1$) or state $K$ ($Y = 0$). The logistic model estimated from this sample is denoted by $x_{j,k^*}' \hat{\beta}_{j,k^*}$. If loan $i$ has explanatory variables $x_{i,j,k^*}$, then

$$
\begin{align*}
P\{\{\text{state } k^*\} \mid \{\text{state } j \text{ this month and state } k^* \text{ next month}\} \\
\cup \{\text{state } j \text{ this month and state } K \text{ next month}\}\} &= \frac{\exp(x_{j,k^*}' \hat{\beta}_{j,k^*})}{1 + \exp(x_{j,k^*}' \hat{\beta}_{j,k^*})}.
\end{align*}
$$

It is important to stress that this is not the transition probability and more is required than simply replacing $p(j, k^*)$ with the prediction from this binary logistic prediction. With respect to the multinomial logistic model for row $j$,

$$
\begin{align*}
P\{\{\text{state } k^* \text{ next month}\} \mid \{\text{state } j \text{ this month}\}\} = \exp(x_{j,k^*}' \hat{\beta}_{j,k^*}).
\end{align*}
$$

For any transition probabilities without a model, the binary logistic model is simply an intercept. That is, suppose the transition from state $j$ this month to state $k_o$ will use the intercept only model. Then

$$
\begin{align*}
P\{\{\text{state } k_o \text{ next month}\} \mid \{\text{state } j \text{ this month}\}\} = \exp(\hat{\beta}_{j,k_o}) = \tilde{p}(j,k_o).
\end{align*}
$$

where the estimated transition probability is denoted by $\tilde{p}(j,k)$. To combine all the binary logistic models (those with and without explanatory variables) into the multinomial logistic model for the $j$th row, first compute the sum

$$
C = 1 + \sum_{k^*} \exp(x_{j,k^*}' \hat{\beta}_{j,k^*}) + \sum_{k_o} \tilde{p}(j,k_o),
$$

where the first sum is over all states that have loan-level models and the second sum is over all states that are intercept only, but are not state $K$. Then, the elements of the $j$th row of
the transition matrix are

\[
p(j, k) = \begin{cases} 
  p(j, k^*; \mathbf{x}_{j,k^*}, \hat{\beta}_{j,k^*}) = \frac{\exp(\mathbf{x}_{j,k^*}' \hat{\beta}_{j,k^*})}{C} & \text{for states with loan-level models} \\
  p(j, k_o) = \frac{\tilde{p}(j, k_o)}{\tilde{p}(j, K)} & \text{for states that are intercept only} \\
  p(j, K) = \frac{1}{C} 
\end{cases}
\]

6 Example

Any company in the business of making loans will require a forecast of future cash flows on their loans. The demands placed on the forecast are particularly heavy for companies who securitize their loans. In essence, a group of loans is set aside in a trust. The cash flow from these loans (principal and interest) is used to support bonds sold to investors. The proper valuation of these bonds will depend critically on the amount and timing of losses and prepayments over the life of the loans. Securitized loans form an unchanging pool. These pools are small enough to expect noticeable variation in loss and prepayment rates even within a stable environment.

The loans for this example are from a $7 billion portfolio in the subprime home equity market. These borrowers generally have weaker or damaged credit, which prevents them from qualifying for loans in the prime market. Not surprisingly, loss rates in the subprime sector are greater than in the prime sector. These loans are fixed-rate, with first liens secured by residential real estate originated between 1994 and 2002. Table 3 contains summary statistics on these 97,124 loans. The proprietary credit score in Table 3 is a custom scorecard that combines characteristics of the collateral and information from a recent credit report. Characteristics about the collateral, such as the size of the lot and home, the year built, and a general assessment of the quality of the home are used in the proprietary credit score, where more valuable collateral motivates borrowers to make payments. Every four months a new credit report is pulled for each loan from one of the bureaus. The counts of satisfactory trades, inquiries the last six months, derogatories, and recent delinquencies, as well as information from the different types of credit card outstanding balances and credit limits are useful predictors of default. There is also some information about changes from the previous credit report that are early predictors of changes in repayment behavior. Repayment behavior on
Table 3. Summary statistics of a $7 billion portfolio of subprime home equity loans secured by residential real estate.

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest Rate (%)</td>
<td>6.00</td>
<td>7.50</td>
<td>8.75</td>
<td>9.5</td>
<td>18.50</td>
</tr>
<tr>
<td>Loan Amount ($)</td>
<td>25,800</td>
<td>55,933</td>
<td>71,519</td>
<td>88,725</td>
<td>331,015</td>
</tr>
<tr>
<td>Loan-to-Value Ratio (%)</td>
<td>1.00</td>
<td>70.33</td>
<td>90.39</td>
<td>94.94</td>
<td>100.00</td>
</tr>
<tr>
<td>Proprietary Credit Score (0=bad, 1=good)</td>
<td>0.00</td>
<td>0.52</td>
<td>0.64</td>
<td>0.76</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Percentage of Credit Report Derogatories

- 12.86  Filed for Bankruptcy
- 11.72  At Least One NSF Check
- 41.24  At Least One Major Derogatory

these loans is observed between March 2001 and February 2002. This time window is narrow since many loans in the portfolio are purchased from other lenders without data on earlier repayment history and because of the assumption that the future was representative of the recent past.

The basic states of the Markov chain model are “current,” “30,” “60,” “90,” “120 or more days past due,” “loss,” and “paid.” The five delinquency states are determined by comparing the due date to the last day of the month. The 120 or more days past due state reflects seriously delinquent borrowers who may remain in this state while loss mitigation or foreclosure proceedings are considered. Loss and paid are absorbing states which are achieved when loans are closed either because the account is declared uncollectable or the borrower pays the outstanding principal (most likely due to borrower refinancing).

In order to use “number of delinquent months” as an explanatory variable in loan-level models for transition probabilities, the “current” and delinquent states must be exploded to “current and \( k \) delinquent months” for example. The resulting transition matrix will have many zero values since transitions like “30 days past due and 3 delinquent months” to “60 days past due and 3 delinquent months” are impossible. In the loan-level models, it appears 10+ is a practical upper bound for the number of delinquent months. The resulting Markov chain model then has 57 states.

The lending environment has been far from stable within the subprime home equity industry. As competition in the industry heated up in the last half of the 1990’s, underwriting standards declined and the use of higher-risk loan acquisition channels increased. Later, a renewed emphasis on servicing replaced the emphasis on volume, causing further
non-stationarity in loss rates. One approach to modeling the changing subprime lending environment is to estimate a different transition matrix for loans originating in each quarter. That is, choose quarterly originations as the segmentation variable. This is problematic since older loans have been observed longer than new loans and the rare transition probabilities must have sufficient sample sizes for every segment. The EB transition matrix estimator is appealing since it shrinks each segment transition probability toward the mean observed for all loans. The amount of shrinkage depends on the sample size and the degree to which the loans originated in a quarter differ from all subprime loans.

Figure 2 demonstrates how the EB estimates differ from the segment estimates. The transition from “30” to “60 days past due” is frequently observed in subprime loans, and the fact that these probabilities vary from 0.20 to 0.40 reflects the importance of early and active management of delinquent subprime loans. The diameter of the circles centered at each segment estimate in Figure 2 represents sample size, with large samples having large circles. Notice that segments with large samples have little shrinkage, and the EB estimates are very close to the segment estimates. The “90 days past due” to “current” event is more unusual, since it represents loans that have progressed into serious delinquency but then return to “current” through a large payment that covers the outstanding principle and interest. The EB estimates demonstrate more shrinkage, particularly since the segment estimates vary from 0.00 to 0.55 and the sample sizes within a segment are often small.

The next feature to incorporate into the transition matrix is the loan-level modeling of a few critical transition probabilities. Because of the large number of loans in the “current” row, precise transition probability estimates of the “30 days past due” and “paid” states seem crucial to the forecast. There are some explanatory variables that would be common to both models, but the mortgage repayment literature suggests default and refinance are driven by different factors. The models presented here are simplifications and it is very likely that many more explanatory variables would be included in the loan-level models.

First, consider the 1,014,362 observations on loans in the data set which were “current” one month and moved to either “current” \((Y = 0)\) or “30 days past due” \((Y = 1)\) the next month. A stratified sample is constructed for modeling containing all 7,626 observations at \(Y = 1\) and a random sample of 7,626 observations at \(Y = 0\). The logistic regression model is estimated with 90% of the stratified sample using the number of delinquent months observed in repayment of this loan \((nd1q)\), the loan-to-value percentage at origination \((ltv)\), the proprietary credit score \((score)\), and the indicator variables for the origination year. The estimated coefficients are given in Table 4. Figure 3 clarifies the piecewise linear ap-
proximations for ndlq and ltv, as well as the differences over time in origination where the increasing likelihood of default in recent origination years reflects the recent decline of underwriting standards and use of higher-risk loan acquisition channels.

Two diagnostics of the quality of the model’s fit are the KS plot and the Decile plot for the 10% holdout sample in Figure 4. The KS plot shows the separation of the empirical distribution function of the estimated probabilities for the binary outcomes. The horizontal axis is the corresponding percentile of the entire population. The distribution function of the accounts that stayed “current” the next month is relatively linear since they represent the majority of the population. A large distance between the two curves indicates a well-performing model since the estimated probabilities for the two groups are well separated. The KS score, the maximum difference between the two distribution functions, is 54, which is respectable compared to delinquency models constructed in other credit scoring applications. Credit scoring models can be measured by separating “goods” and “bads,” but prediction is a requirement of a loan-level transition probability model. The Decile plot is constructed by calculating the estimated probability of becoming “30 days past due” next month in the holdout sample and grouping the probabilities into deciles. In each decile, the observed proportion of loans moving from “current” this month to “30 days past due” next month is computed. The Decile plot is the estimated probability and observed proportion for each decile. The actual probabilities appear well predicted since the values fall closely on a unit

Figure 2: Comparison of the empirical Bayes and segment estimates where segments are quarterly loan origination. The right panel is for the transition from 30 to 60 days past due. The left panel is for the transition from 90 days past due to current.
Figure 3: Piecewise linear approximation for the effect of number of delinquent months observed in repayment of this loan (ndlq) and loan-to-value ratio at origination (ltv), as well as the effect of origination year in the logistic model for loans “current” one month and either “current” \( Y = 0 \) or “30 days past due” \( Y = 1 \) next month.

Figure 4: KS plot and Decile plot demonstrating holdout sample performance of the logistic model for loans “current” one month and either “current” \( Y = 0 \) or “30 days past due” \( Y = 1 \) next month. The KS plot indicates the separation in the population “goods” and “bads,” and the Decile plot indicates the prediction performance.
Table 4. Estimated logistic regression coefficients for a loan-level estimate of the “current” to “30 days past due” transition probability. Both ndlq and ltv use a piecewise linear approximation at the given knots to approximate the behavior of the gam model.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.8829</td>
<td>0.0000</td>
</tr>
<tr>
<td>ndlq at 0 mo</td>
<td>1.5539</td>
<td>0.0783</td>
</tr>
<tr>
<td>ndlq at 1 mo</td>
<td>2.3097</td>
<td>0.0726</td>
</tr>
<tr>
<td>ndlq at 5 mo</td>
<td>2.7968</td>
<td>0.0677</td>
</tr>
<tr>
<td>ndlq at 10+ mo</td>
<td>3.1895</td>
<td>0.0555</td>
</tr>
<tr>
<td>ltv less than 50%</td>
<td>-0.1377</td>
<td>0.0413</td>
</tr>
<tr>
<td>ltv at 80%</td>
<td>-0.1080</td>
<td>0.0705</td>
</tr>
<tr>
<td>ltv at 100%</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>score</td>
<td>-4.6405</td>
<td>0.0873</td>
</tr>
<tr>
<td>Originated 1994</td>
<td>0.0322</td>
<td>0.0748</td>
</tr>
<tr>
<td>Originated 1995</td>
<td>-0.1200</td>
<td>0.0553</td>
</tr>
<tr>
<td>Originated 1996</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Originated 1997</td>
<td>-0.0414</td>
<td>0.0467</td>
</tr>
<tr>
<td>Originated 1998</td>
<td>0.1338</td>
<td>0.0446</td>
</tr>
<tr>
<td>Originated 1999</td>
<td>0.2918</td>
<td>0.0450</td>
</tr>
<tr>
<td>Originated 2000</td>
<td>0.5540</td>
<td>0.0506</td>
</tr>
<tr>
<td>Originated 2001–02</td>
<td>0.7567</td>
<td>0.0855</td>
</tr>
</tbody>
</table>
slope line through the origin. One feature of interest in the Decile plot is the large spread of average probability from the highest to lowest decile, indicating a very high risk group in the population for becoming delinquent.

The “current” to “paid” loan-level model is based on a stratified sample of the available 1,015,008 observations where loans were “current” one month and moved to either “current” ($Y = 0$) or “paid” ($Y = 1$) the next month. The stratified sample consists of all 8,272 observations at $Y = 1$ and a random sample of 8,272 observations at $Y = 0$. Table 5 contains the estimated logistic coefficients using 90% of the stratified sample. An important explanatory variable in prepayment is interest rate ($\text{apr}$), which is included in addition to number of months since loan origination ($\text{age}$), proprietary credit score ($\text{score}$), and indicator variables for the origination year used in the previous model. Figure 5 provides the partial logistic effects of the piecewise linear approximations for $\text{apr}$ and $\text{age}$, as well as the differences in origination over time where loans made recently are more likely to prepay. Figure 6 contains the KS plot indicating typical separation for prepayment models and the Decile plot indicating good prediction for a very small probability, both for the 10% holdout sample.

The inclusion of exogenous covariates is intriguing because they have demonstrated improved prediction in previously observed months. For example, the spread between the mortgage interest rate and LIBOR would measure when the borrower had an incentive to refinance to a lower rate. An improvement on the refinance incentive for subprime borrowers was demonstrated in Alexander et al. [20] who used a covariate that measured the borrower’s time-varying incentive to refinance when interest rates fall. The lender in their paper assigned each loan an internal grade at origination. The grade is based on loan-to-value ratio, credit score, mortgage payment to income ratio, and mortgage payment history. Loans with an A grade are the safest and a C- grade are the riskiest. In order to compute a borrower’s incentive to refinance the assumption was that the borrower would refinance into the market rate for the same grade. That is, if a borrower’s original loan was 15 basis points above the average B+ rate for this lender then the refinance rate was computed to be 15 basis points above the average B+ interest rate for this lender the current month. Another example would be the time-varying state unemployment rate that would measure when a borrower is likely to default.

The problem with including many exogenous covariates is that they are time-varying, with the consequence that the Markov chain model requires future values of the exogenous covariate for the prediction period to be provided. That is, while state unemployment rate
Figure 5: Piecewise linear approximation for the effect of interest rate (\( \text{apr} \)) and months since origination (\( \text{age} \)), as well as the effect of origination year in the logistic model for loans “current” one month and either “current” (\( Y = 0 \)) or “paid” (\( Y = 1 \)) next month.

Figure 6: KS plot and Decile plot demonstrating holdout sample performance of the logistic model for loans “current” one month and either “current” (\( Y = 0 \)) or “paid” (\( Y = 1 \)) next month. The KS plot indicates the separation in the population “goods” and “bads,” and the Decile plot indicates the prediction performance.
Table 5. Estimated logistic regression coefficients for a loan-level estimate of the “current” to “paid” transition probability. Both apr and age use a piecewise linear approximation at the given knots to approximate the behavior of the gam model.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.8016</td>
<td>0.0000</td>
</tr>
<tr>
<td>apr at &lt;6%</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>apr at 9%</td>
<td>2.1326</td>
<td>0.0832</td>
</tr>
<tr>
<td>apr at 15+%</td>
<td>2.6943</td>
<td>0.1245</td>
</tr>
<tr>
<td>age at &lt;12 mo</td>
<td>-5.3184</td>
<td>0.2778</td>
</tr>
<tr>
<td>age at 48 mo</td>
<td>-3.9319</td>
<td>0.1529</td>
</tr>
<tr>
<td>age at 120+ mo</td>
<td>-3.5462</td>
<td>0.3991</td>
</tr>
<tr>
<td>score</td>
<td>3.3856</td>
<td>0.1183</td>
</tr>
<tr>
<td>Originated 1994</td>
<td>-0.1504</td>
<td>0.1858</td>
</tr>
<tr>
<td>Originated 1995</td>
<td>-0.0060</td>
<td>0.1034</td>
</tr>
<tr>
<td>Originated 1996</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Originated 1997</td>
<td>0.1082</td>
<td>0.0960</td>
</tr>
<tr>
<td>Originated 1998</td>
<td>0.2912</td>
<td>0.1222</td>
</tr>
<tr>
<td>Originated 1999</td>
<td>0.4808</td>
<td>0.1587</td>
</tr>
<tr>
<td>Originated 2000</td>
<td>0.7297</td>
<td>0.2062</td>
</tr>
<tr>
<td>Originated 2001–02</td>
<td>0.1348</td>
<td>0.2574</td>
</tr>
</tbody>
</table>
is known for the period where segments or loan-level models are estimated, these unknown future values are now required inputs in the Markov chain model. One pragmatic approach would be to use exogenous covariates with available forward curves or economist forecasts, but this should be done cautiously since the forecasts have uncertainty that is not included in the model.

It is certainly worth entertaining the idea of a few other loan-level models. The “30” to “60 days past due” transition probability may be worth modeling if the loan servicer concentrated attention on early defaults. The “120+ days past due” to “loss” transition probability could be modeled if loss mitigation and foreclosure proceedings played an important role in determining the timing of losses. However, one must be conscious of the available sample size to estimate these models, the performance of the available explanatory variables, and the increasing complexity of the overall model.

One validation of the proposed model is to compare the forecast of the portfolio as composed in March 2001 (the first month of the repayment data) to the actual values through February 2002 (the last month of the repayment data). For each of the 97,124 loans, the forecast is computed as detailed in Section 1 using the EB estimate and loan-level models for the transition matrices. Figure 7 compares the actual and forecast balance for each state of the Markov chain. There is good agreement in all states, particularly the “30 days past due” and “paid” states, which demonstrates the value of the loan-level models. The worst performances from the perspective of $|y - \hat{y}|/y$ are in the “90 days past due” and “loss” states, where the difference is $3$ million to $5$ million relative to a $10$ million to $16$ million balance.

Another validation investigates the precision of multistep ahead forecasts. This is of particular concern since the estimation focuses on the transition probabilities not the state balances. Figure 8 contains boxplots of the multistep ahead forecast error for current and prepayment. To compare current forecasts, consider the percentage of the original portfolio balance in the current state $k$ months ahead. The right panel of Figure 8 indicates the best forecasts are one month ahead with the prediction variance increasing before leveling out at about four months ahead. The comparison of prepayment is based on the conditional prepayment rate (CPR), which measures prepayments as a percentage of the current outstanding balance expressed as a compound annual rate. The left panel of Figure 8 indicates equivalent forecasting of prepayments for all $k$ month ahead forecasts.

Finally, to demonstrate a useful analytic feature of the model consider forecasts generated under possible loan servicing strategies or lending environments. Three examples are
Figure 7: Actual and Forecast using EB transition matrix estimates and loan-level models for “current” to “30 days past due” and “current” to “paid” transition probabilities. Validation is performed using loans on the books March 2001 and followed for 11 months. While “loss” and “paid” are absorbing states, the plot is of each month’s addition.
presented in Figure 9 in the forecast using February 2002 balances to December 2003 and compared to the regular forecast. The blue lines in Figure 9 are associated with a possible effort to address the large balance in the 120+ days past due state. All loans with a custom credit score greater than 0.40 will be targeted by loan servicing to move them to current, loss, or paid. Loss mitigation offers may allow borrowers to return from serious delinquency, borrowers with equity in their home may sell or refinance with another lender, or the evaluation of the borrower’s ability to pay and collateral may determine that the loan should be written off as a loss. For one month, loan servicing is expected to alter the estimated transition matrix so that the transition probabilities from the 120+ days past due state are 0.90 to current, 0.04 to paid, 0.05 to loss, and 0.01 to remain 120+. Notice that the blue lines are what is expected, with an increase in current balance, a drop in 120+ balance, and a one-time large loss balance. One of the appealing features of the Markov Chain model is that the 120+ balances that move to current in the first month are likely to default and model reflects the timing of these moves through delinquency.

The next scenario is a favorable borrowing environment with low interest rates, home appreciation, and more loan products available. Only borrowers with good credit history and equity will be able to take advantage of these opportunities. Since the computations are performed on each loan, the forecast can be modified for only a subset of the portfolio. For example, the green lines in Figure 9 consider the case that borrowers with interest rates greater than 12%, custom credit score greater than the median (0.64), and originated their loan in the last 5 years, are three times more likely to refinance than currently reflected in the current-to-paid model. As expected, the balance in paid is significantly higher and the
Figure 9: Forecast for loans on the books February 2002 under a loan servicing strategy on 120+ days past due loans (blue), favorable borrowing environment (green), and increased default environment (red). The regular forecast (black) is included for comparison. While loss and paid are absorbing states, the plots include the monthly addition.
balance of current loans is much lower. One interesting feature is that the paid balance is declining over time instead of increasing as shown in the regular forecast. This is because borrowers who can will refinance quickly in this borrowing environment leaving a smaller portfolio for later months.

Finally, consider the case where rising interest rates, declining home prices, and a slow economy result in more defaults. The red lines indicate forecasts if borrowers with a delinquency in loan repayment, a loan-to-value ratio greater than 80%, and custom credit score less than the upper quartile (0.76) are 50% more likely to default than the current-to-30 model. As expected, ultimately there is an increase in loss but the Markov Chain model maintains the timing that those losses won’t be observed until the March 2002 increase in 30 days past due works through the 60, 90, and 120+ states.

7 Simulation Study

The sampling distributions of the estimated transition probabilities are well known, but the practical importance of the Markov chain model is in its predictions. A simulation study based on the subprime example in the previous section investigates how sampling variation in transition probability estimates affects prediction precision. Consider a homogeneous stationary Markov chain with states “current,” “30,” “60,” “90,” “120+ days past due,” “loss,” and “paid.” For all simulations, the transition matrix is given in Table 6 and is the mean transition matrix from the subprime example given in Section 6. The simulation investigates the first observation month, the number of months in which loan movements are observed, and the number of loans in the initial portfolio. The results of 10,000 replications for each simulation are presented in Table 7.

The most important question is the performance of the estimator in future forecasts. After estimating the transition matrix, forecasts 12, 24, 36, and 48 months ahead of the number of loans in each state are compared to the simulated path using the true transition matrix. As expected, the median absolute deviation (MAD) comparing the actual and predicted proportion of loans in non-absorbing states usually increases as the forecast horizon increases. The exceptions are transitions to absorbing states, which may be due to smaller proportions of loans moving to these states further into the future.

One interesting feature of estimating the transition matrix is that the individual transition sample size can not be selected, but the movements from a beginning portfolio must be observed. The upper and lower portions of Table 7 compare different numbers of loans
Table 6. Transition matrix for simulation study. The homogeneous Markov chain has states states “current,” “30,” “60,” “90,” “120+ days past due,” “loss,” and “paid.” and was constructed from the mean transition matrix from the subprime example given in Section 6.

\[
P = \begin{bmatrix}
0.983 & 0.008 & 0.001 & 0.000 & 0.000 & 0.000 & 0.008 \\
0.327 & 0.116 & 0.333 & 0.004 & 0.003 & 0.002 & 0.010 \\
0.279 & 0.019 & 0.052 & 0.481 & 0.152 & 0.010 & 0.007 \\
0.161 & 0.005 & 0.012 & 0.043 & 0.748 & 0.025 & 0.006 \\
0.053 & 0.001 & 0.001 & 0.001 & 0.834 & 0.108 & 0.002 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\
0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \\
\end{bmatrix}
\]

at the beginning of the Markov chain. Clearly more loans result in smaller error since the transition probabilities as a whole become better estimated with more data, even if particular transitions have few observations.

A common question in applying the Markov chain model for delinquency is the trade-off between a wide window of monthly data to improve estimation and a narrow window of monthly data to represent current loan servicing practice. The simulation indicates that more data is always better, but six months of data results in adequate prediction performance.

The final comparison is between scenarios where loans are observed from the beginning compared to scenarios where loans are observed after 36 months. That is, can the first few months be used to estimate the transition matrix or does the model require a few months for the loans to distribute among the states? The MAD indicates forecasts are better when loans have been in the Markov chain a few months, but forecasts are acceptable when observing a large portfolio of loans the first six months to estimate transition probabilities. However, Figure 10 indicates that the prediction error is not symmetric in the forecasts based on early observations. It appears underestimating the current state and overestimating the delinquent states is quite common.

8 Conclusion

This paper proposes innovative approaches to estimating the transition matrix for a Markov chain model for accounts receivable. When applied to forecasting balances in delinquency states for a portfolio of subprime loans, the transition matrix is unknown and must be estimated from observed monthly loan transitions. When a portfolio consists of loans from
Table 7. Median absolute deviation of 12, 24, 36, and 48 months ahead forecasts of proportion of loans in non-absorbing states at last observed month that are “current,” “90+ days past due,” “loss,” and “paid” for the simulation study. While “loss” and “paid” are absorbing states, the deviation is computed from the monthly addition.

1,000 Loans Begin in Current State

<table>
<thead>
<tr>
<th>Months</th>
<th>First Observed Month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12 24 36 48</td>
</tr>
<tr>
<td></td>
<td>12 24 36 48</td>
</tr>
<tr>
<td></td>
<td>12 24 36 48</td>
</tr>
<tr>
<td></td>
<td>12 24 36 48</td>
</tr>
<tr>
<td>12</td>
<td>Current 0.545 0.613 0.646 0.664</td>
</tr>
<tr>
<td></td>
<td>90 + 0.442 0.500 0.529 0.539</td>
</tr>
<tr>
<td></td>
<td>Loss 0.094 0.092 0.091 0.089</td>
</tr>
<tr>
<td></td>
<td>Paid 0.197 0.182 0.172 0.166</td>
</tr>
</tbody>
</table>

10,000 Loans Begin in Current State

<table>
<thead>
<tr>
<th>Months</th>
<th>First Observed Month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12 24 36 48</td>
</tr>
<tr>
<td></td>
<td>12 24 36 48</td>
</tr>
<tr>
<td></td>
<td>12 24 36 48</td>
</tr>
<tr>
<td></td>
<td>12 24 36 48</td>
</tr>
<tr>
<td>12</td>
<td>Current 0.172 0.190 0.201 0.205</td>
</tr>
<tr>
<td></td>
<td>90 + 0.138 0.158 0.164 0.169</td>
</tr>
<tr>
<td></td>
<td>Loss 0.029 0.028 0.026 0.024</td>
</tr>
<tr>
<td></td>
<td>Paid 0.062 0.058 0.054 0.051</td>
</tr>
</tbody>
</table>
Figure 10: Boxplots of prediction error for 12, 24, 36, and 48-month ahead forecasts of proportion of loans in non-absorbing states at last observed month that are current, 90+ days past due, loss, and paid for simulation study with 1,000 loans beginning in the current state and observed for 6 months. While loss and paid are absorbing states, the deviation is computed from the monthly addition.
different segments, Bayes and empirical Bayes estimators allow different estimates for those transition probabilities that differ between segments but borrow strength in estimating transition probabilities that are the same between segments. Additional covariates can be used to construct loan-level models for a few key transition probabilities that reflect differences over time and credit quality. The case study on a $7 billion securitized portfolio of subprime home equity loans demonstrates estimated transition matrices and resulting forecasts. Differences in underwriting standards over time can be modeled using empirical Bayes estimators. More precise predictions for the “current” to “30 days past due” and “paid” transitions can be made by incorporating past repayment behavior, interest rate, and home equity as covariates in loan-level models. A simulation study indicates the estimated transition matrix can produce reasonable predictions. Some of the important elements of good prediction include the number of loans observed, the number of months observed, and the first observation of the Markov chain.

Like any model using historical data to make forecasts, the Markov chain model would not have forecasted the rapid decline in subprime mortgage valuations when the data showed servicing was effective and borrowers had equity in their homes that could be used to avoid delinquency. One can speculate that a Bayes estimator of the transition matrix using a prior as shown in the pessimistic scenario in Table 2 foretelling a change in borrower repayment that dominated the historical data would have indicated the possibility of a rapid change to the high subprime default rates observed in the last year. More realistically, efficient use of the data in transition probability estimation would provide accurate revised forecasts using only a few months of data. That is, even with only three or four months of repayment behavior data during the current recession, segmented transition matrices could be estimated and loan-level models for a few key transition probabilities could be constructed with useful covariates. These estimates applied within the mechanics of the Markov chain model would produce revised short-term forecasts that would be far superior to those still based on predominantly older data or models that needed longer horizon data before a revised forecast could be constructed.

ACKNOWLEDGEMENTS

We are grateful to the Editor, Associate Editor, and referees of *Applied Stochastic Models in Business and Industry* for helpful comments on an earlier draft and for suggestions that substantially improved the paper.
References


20. Alexander WP, Grimshaw SD, McQueen GR, Slade, BA. Some loans are more equal than others: third party originations and defaults in the subprime industry. *Real Estate Economics* 2002; **30**:667–697.