Forecasting Mortgage Delinquency: An Information Theoretic Approach

Prepared for Serigne Diop (by George Panterov)

December 11, 2012

Vero Capital Management

220 5th Ave, 19th Floor

New York, NY

Abstract

In this model we develop two information-theoretic models for forecasting mortgage delinquency that naturally incorporate loan-level data as well as prior expert opinions about the probabilities of interest. Both models are estimated using the maximum entropy framework and each model has its own advantages and disadvantages. We use a proprietary data from Vero Capital Management that contains transition information on 6,204 mortgage loans made to small and medium enterprises. We incorporate covariates to make the state in which the loan is in each period conditional on the economic environment of that period and on the specific characteristics of the loan itself.

Introduction

Each mortgage can transition through several possible states over the course of its life. When the borrower makes frequent payments the loan is in "Current" state. When the borrower is late with a payment the loan transitions in "30 days past due" state. If next month the borrower again doesn't make a payment then the loan transitions into "60 days past due" and so on. There are two absorbing states in this framework. The borrower can either pay the loan in full or go into a foreclosure (the lender declares the loan a loss). Because of this states-transitions framework a popular modeling approach in this literature has been to estimate Markov Chain models for portfolios of mortgages. This has been done by Cyert et al. (1962) and Betancourt (1999). More recently a Bayesian extension to the traditional Markov Chain framework was developed and applied to a portfolio of sub-prime mortgages by Grimshaw and Alexander (2011). Our model is similar in spirit to the ones proposed by the studies mentioned above, but our estimation framework is based on the maximum entropy approach.

Markov chain models are a natural choice for modeling and forecasting mortgage delinquencies because they handle the state dependent transition probabilities. Suppose for example that we are interested in modeling only 5 delinquency states for our mortgage loan portfolio. Let these states be: "Current", "30 days past due", "60 days past due", "Loss" and "Paid". Then the transition probability matrix from time t to $t+1$, P_t might be something like :

$$P_t = \begin{bmatrix} 0.6 & 0.3 & 0 & 0.03 & 0.07 \\ 0.35 & 0.45 & 0.1 & 0.05 & 0.05 \\ 0.2 & 0.3 & 0.35 & 0.05 & 0.1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where the entries indicate the probability of transition from the row state to the column state. For example, entry $P(1,2)$ indicates that the probability of a loan, which is "Current" (row number 1) to transition into state "30 days past due" (column number 2) is 0.3. A current loan cannot transition automatically in "60 days past due" state so $P(1,3) = 0$. Similarly, the probability of loan that is "30 days past due" to stay in the same state next period is $P(2,2) = 0.45$. The "Loss" and "Paid" states are absorbing, which means that once a loan is in one of the two states it will remain there. Note also that the entries in each row sum to one which simply means that a loan must be in one of the four possible states. If we accurately estimate this transition probability matrix for a loan portfolio we will know the expected state of the portfolio for the next period.
The state of each loan in time t can be represented by a Boolean vector y_{it} whose j-th entry is one if the loan is in state j at time t. Therefore the expected state of a loan at time $t+1$ is: $y_{i,t+1} = y_{it} \times P_t$. For example, if loan i is in state $j = 2$, at time t then $y_{it} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix}$ and we have that the expectation for next period is:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.6 & 0.3 & 0 & 0.03 & 0.07 \\ 0.35 & 0.45 & 0.1 & 0.05 & 0.05 \\ 0.2 & 0.3 & 0.35 & 0.05 & 0.1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.35 & 0.45 & 0.1 & 0.05 & 0.05 \end{bmatrix}$$

Since each loan is going to be in one of the four states at time $t+1$, our individual loan forecast will necessarily be wrong. However if we aggregate the values across all loans in the portfolio we will be able to answer questions like how many loans in the portfolio are expected to be "Current" next period, how many "30 days past due" etc. We can also answer questions like what is the dollar amount of outstanding balances that are expected to be in each state at time $t+1$. Therefore the expected number of loans in each state in time $t+1$ is given by $P_t \sum_i y_{it}$. Of course the challenge is in developing an econometric model that can estimate this transition matrix in a consistent way for N loans over T periods and also allow us to incorporate useful loan-specific information (e.g. credit scores of borrowers, unemployment rate in region where loan is made etc.).
Grimshaw and Alexander (2011) propose several methods for estimating the transition probabilities for a portfolio of mortgage loans. One approach is to divide the portfolio of loans into segments of similar mortgages and calculate a stationary transition matrix for each segment. To do this they develop a Bayesian model for estimating the transition probabilities. One major advantage of the Bayesian approach is that it allows the researcher to incorporate prior beliefs for some entries of the transition matrix when there is not enough data. For example, if the researcher segments the portfolio of mortgages, she may not have enough data in each segment to estimate some of the rarer transitions such as "120 past due" to "Current" (i.e. the borrower made 5 payments in the same period). This data problem could be alleviated by having an expert provide prior beliefs (probabilities) for the entries in the transition matrix. This gives the researchers a considerable flexibility because it allows them to consistently incorporate both "hard" data and expert beliefs when data is sparse. One draw back of the model developed by Grimshaw and Alexander is that it doesn't allow for loan-level characteristics to be incorporated into the estimation. The heterogeneity in the loans is handled by the segmentation of the portfolio into "similar" segments which is often not ideal.

One advantage of the maximum entropy framework developed in this study is in the ease with which it handles loan-specific and economy-level characteristics in addition to handling prior beliefs on the transition probability matrix.

In our second approach we develop a multinomial model that estimates the probability that a mortgage will transition in a particular state given its current state and a vector of covariates. One advantage of this approach is that it is much easier to develop a loan-level forecast for a particular type of loan than it is with the Markov chain model. This could be useful to the practitioner who wants to extrapolate information from one portfolio and apply it to a another portfolio with a different composition. However, we don't introduce a way to incorporate prior information with the multinomial model.

In what follows we describe the data, present the maximum entropy frameworks for the Markov and the multinomial models and the empirical results and we conclude.

Data

Our data consists of 6,204 mortgage loans provided by Vero Capital Management. We observe the monthly status of each of the loans for the period between September 2011 and June 2012. Our covariates include the loan characteristics at the time the loan was issued. For example we have data on the original FICO score of the loan, the loan to value ration (LTV), the original balance and coupon and others. In addition we also have some macro data like HPA and the unemployment rate for each month in the sample.

The possible states of the loans are:

Table 1. Observed transition states

state	description
C	Current
30	30 days past due
60	60 days past due
90	90 days past due
120	120 days past due
REO	Real Estate Owned
F	Foreclosure

We present a summary of some of the loan-specific covariates in our data:

Table 2.1. Summary statistics for loan-level covariates

Variable	Mean	Standard Deviation
Original LTV	61.7	22.1
Original FICO	728.51	130.24
Original Balance	229689.18	157367.99
Original Coupon	5.439	0.73

and the summary for the states for each period in our sample:

Table 2.2. Number of loans in each state in each period

	6/12	5/12	4/12	3/12	2/12	1/12	12/11	11/11	10/11	9/11
C	6040	6057	6053	6048	6040	6034	6043	6050	6073	6073
30	74	56	50	69	75	81	85	87	70	70
60	11	11	26	17	23	33	22	17	15	19
90	10	15	12	13	16	11	12	9	8	4
120	9	9	9	9	8	4	5	4	3	4
REO	54	51	47	41	40	36	34	33	31	31
F	6	5	7	7	2	5	3	4	4	3

We only focus on the four loan characteristics summarized in Table 2. Part of the reason of this somewhat limited model is that our purpose in this model is to focus on the framework and approach and less so on the results. A practitioner could easily extend the variables included in the model.

Markov Chain Model

In this section we present the Markov Chain model within the maximum entropy framework. In what follows we rely almost exclusively on the framework developed in Golan (2008).
Following Golan (2008) a stationary first order Markov model can be formulated as follows:

$$\sum_{i=1}^{N} y_{itj} = \sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj} y_{i,t-1,k} \tag{1}$$

where for loan $i = 1, 2, ..., N$ in period $t = 1, 2, ..., T$ the indices j and k represent the states in periods t and $t-1$. y_{itj} is a K-dimensional Boolean vector for each loan that has 1 in its j-th position if loan i is in state j at time t. For example if there are 3 possible states ($K = 3$) and loan 3 is in state 2 at time 0 then $y_{3,2,0} = [0, 1, 0]$. p_{kj} are the elements of the $K \times K$ transition probability matrix P where $\sum_{j=1}^{K} p_{kj} = 1$ (row elements must sum to one) ensures that $\mathbf{p}$ is a proper probability distribution.
This formulation of the transition probabilities requires that the expected number of loans in each state at time $t-1$ exactly equals the actual number of loans in each state at time t. Although we expect the two quantities to be close, they may not always be exactly equal. We account for this in what follows.
Let's generalize this formulation for multiple periods T and let's remove the requirement that the cross moments equal each other exactly by introducing a noise term:

$$\sum_{t=2}^{T}\sum_{i=1}^{N} y_{itj} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj} y_{i,t-1,k} + \sum_{t=1}^{T-1}\sum_{i=1}^{N} \epsilon_{itj} \tag{2}$$

where the additive noise $\epsilon \in [-1, 1]$ has a zero mean and is defined on a support ν that is $M \geq 2$ dimensional. Therefore we can rewrite the above as

$$\sum_{t=2}^{T}\sum_{i=1}^{N} y_{itj} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj} y_{i,t-1,k} + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{m=1}^{M} w_{itjm}\nu_{jm} \tag{3}$$

where $\mathbf{w}$ is the probability distribution over the support ν such that $\sum_{m=1}^{M} w_{itjm}\nu_{jm} = \epsilon_{itj}$ and $\sum_{m=1}^{M} w_{itjm} = 1$. The noise ϵ alleviates the requirement that the cross moments of the data are exactly equal.
So far we have defined a stationary Markov chain model that doesn't incorporate any loan-specific information in the transition probabilities and also doesn't incorporate any prior beliefs. Suppose that we have loan-specific characteristics Z_t that may influence the probabilities p_{ij} (e.g. credit score of borrower, collateral, economic characteristics of region where loan is made etc.) In traditional estimation frameworks we usually need to specify the functional dependence between the P and Z. But in most cases it is impossible for the researcher to know this exact functional form. Thus, in the information-theoretic framework, which this paper develops, we incorporate loan-level characteristics through the cross moments of the data:

$$\sum_{t=2}^{T}\sum_{i=1}^{N} y_{itj} z_{its} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj} y_{i,t-1,k} z_{its} + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{m=1}^{M} w_{itjm}\nu_{jm} z_{its} \tag{4}$$

where $s = 1, 2, ...S$ denotes the explanatory variable for each loan i at time t. This gives as a system of equations for each delinquency state j and explanatory variable s (a $K \times S$ system of equations). We want to recover the transition probabilities P and the noise probabilities W.
The idea behind the information-theoretic approach is that we want to recover the probabilities of interests that are consistent with the data at hand and at the same time carry the smallest amount of information. The rationale behind this approach is that we want to impose as little structure on the data as possible (we only use what we observe – the moments and cross moments of the data). To do this we minimize an objective function, which is our information measure, subject to the constraints, which come from the observed data and the proper probability requirements. Therefore, an important

question in this framework is what should the objective function be? Jayens (1957a,b) proposed a constrained optimization problem that involves maximizing the Shannon (1948) entropy:

$$H(\mathbf{p}) = -\sum_{k=1}^{K}\sum_{j=1}^{K} p_{kj}\ln p_{kj} \tag{5}$$

which reaches its peak when the p-s are uniformly distributed (lowest amount of information). This is the objective function for the maximum entropy framework (ME). Therefore maximizing Shannon's entropy subject to the data constraints will give us the probability distribution that is consistent with the observed data and at the same time has the "flattest" distribution (least amount of structure possible). In the case at hand the objective function to be maximized would be:

$$H(\mathbf{p},\mathbf{w}) = -\sum_{k=1}^{K}\sum_{j=1}^{K} p_{kj}\ln p_{kj} - \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{m=1}^{M} w_{itjm}\ln w_{itjm} \tag{6}$$

which is the objective function for the generalized maximum entropy (GME) framework.
The last element to incorporate in this model is prior information. Suppose we want to elicit some expert (prior) opinions about the probabilities of interest and let these priors be $\mathbf{p^0}$ and $\mathbf{w^0}$. This modification can easily be handled within the information-theoretic approach. Instead of maximizing Shannon's entropy measure we can minimize a distance measure between the pairs of probability distributions $\mathbf{p},\mathbf{w}$ and $\mathbf{p^0},\mathbf{w^0}$. The information distance between the two pairs of distributions can be measured by the Generalized Cross Entropy (GCE):

$$D(\mathbf{p},\mathbf{w}||\mathbf{p^0},\mathbf{w^0}) = \sum_{k=1}^{K}\sum_{j=1}^{K} p_{kj}\ln\left(\frac{p_{kj}}{p^0_{kj}}\right) + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{m=1}^{M} w_{itjm}\ln\left(\frac{w_{itjm}}{w^0_{itjm}}\right) \tag{7}$$

discussed at length in Golan (2008). Therefore we minimize (7) subject to the data constraints (4) and the requirement that all probabilities are proper (i.e. sum to one). The solution of this constrained optimization problem gives us the probability distribution that is closest to our prior beliefs (closeness is defined by our cross entropy measure) and consistent with the data observed (or more exactly with the moments of the data). If the prior distribution is uniform (i.e. no information) then minimizing the GCE is the same as maximizing Shannon's entropy (6) subject to the constraints:

$$\min_{\mathbf{p},\mathbf{w}} D(\mathbf{p},\mathbf{w}||\mathbf{p^0},\mathbf{w^0})$$

subject to the constraints (4) and the proper probability distribution. Note that this constrained optimization problem is difficult to solve even with modern numerical methods due to the large dimensions of the problem. In order to make the optimization more tractable we can transform the (primal) constrained optimization problem into the (dual) unconstrained optimization problem. The Lagrangian for this problem is

$$\begin{aligned}
\mathcal{L} &= \sum_{k=1}^{K}\sum_{j=1}^{K} p_{kj}\ln\left(\frac{p_{kj}}{p^0_{kj}}\right) + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{m=1}^{M} w_{itjm}\ln\left(\frac{w_{itjm}}{w^0_{itjm}}\right) \\
&+ \sum_{s=1}^{S}\sum_{j=1}^{K}\lambda_{sj}\left[\sum_{t=2}^{T}\sum_{i=1}^{N} y_{itj}z_{its} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{kj}y_{i,t-1,k}z_{its} + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{m=1}^{M} w_{itjm}\nu_{jm}z_{its}\right] \\
&+ \sum_{k=1}^{K}\mu_k\left[1-\sum_{j=1}^{K} p_{kj}\right] + \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{K}\rho_{itj}\left[1-\sum_{m=1}^{M} w_{itjm}\right]
\end{aligned} \tag{8}$$

where $\boldsymbol{\lambda}, \boldsymbol{\rho}$ and $\boldsymbol{\mu}$ are the Lagrange multipliers associated with the constraints. The solution for the probabilities is (for the step-by-step solution see Golan (2008)):

$$p_{kj} = \frac{p_{kj}^0 exp(\sum_{t=1}\sum_{i,s} y_{itk} z_{its} \lambda_{sj})}{\sum_j p_{kj}^0 exp(\sum_{t=1}\sum_{i,s} y_{itk} z_{its} \lambda_{sj})} \equiv \frac{p_{kj}^0 exp(\sum_{t=1}\sum_{i,s} y_{itk} z_{its} \lambda_{sj})}{\Omega_k(\boldsymbol{\lambda})} \tag{9}$$

and

$$w_{itjm} = \frac{w_{itjm}^0 exp(\sum_s z_{its} \nu_{jm} \lambda_{sj})}{\sum_m w_{itjm}^0 exp(\sum z_{its} \nu_{jm} \lambda_{sj})} \equiv \frac{w_{itjm}^0 exp(\sum_s z_{its} \nu_{jm} \lambda_{sj})}{\Phi_{itj}(\boldsymbol{\lambda})} \tag{10}$$

We then plug the two solutions (9) and (10) in the objective function (7) to obtain the unconstrained (dual) optimization problem:

$$l(\boldsymbol{\lambda}) = \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{i,s} y_{itj} z_{its} \lambda_{sj} - \sum_k \ln \Omega_k(\boldsymbol{\lambda}) - \sum_{i,t,j} \ln \Phi_{itj}(\boldsymbol{\lambda}) \tag{11}$$

Note that this is now an unconstrained optimization problem whose parameters space consists of the dimensions of the Lagrange multipliers $\boldsymbol{\lambda}$, which is usually of much lower dimension than that of the probabilities of interests. We use numerical optimization routines to solve for the Lagrange multipliers and then plug back the results in (9) and (10) to recover the probabilities of interest. Note that the transition probabilities are not stationary, i.e. they will depend on the loan level characteristics z_{its} for each period t. Note that they also depend on the prior probabilities $\mathbf{p^0}, \mathbf{w^0}$.

When this model is applied in practice there are usually many covariates the practitioner can choose to incorporate in the model that might have potential effects on the delinquency probabilities. This of course raises the issue of bias and variance trade offs and overfitting the model. This is why, in order to apply this framework effectively, we need an adequate method of model selection. One way in which we can evaluate model fit is through the normalized entropy $S(\mathbf{p}) = \frac{\sum_{k,j} p_{kj} \log(p_{kj})}{\sum_{k,j} p_{kj}^0 \log(p_{kj}^0)}$. $S(\mathbf{p})$ is bounded between 0 and 1, with 1 representing complete ignorance and 0 reflecting perfect knowledge. One interpretation of $S(\mathbf{p})$ is that it captures the information in the new data relative to the prior knowledge. This gives us a way to compare models. Suppose that we want to see if the addition of a new covariate is informative. If $S(\mathbf{p}_{new}) \geq S(\mathbf{p})$ then we can conclude that this covariate doesn't contribute any new information.

In what follows we apply this Markov Chain model to our data with several different sets of covariates and we present the results. We adopt a uniform prior for the transition probabilities. This implies that we start off with a belief that all transitions are equally likely. This is clearly wrong. For example we know that some transitions have very low or zero probabilities (e.g. "Current" to "90 days past due"). Nevertheless we continue with the adoption of the uniform prior for two reasons: 1) It is equivalent to the GME (as opposed to the GCE) framework in (6) and 2) it is consistent with our goal to focus on the framework rather than on the application, we let the practitioners choose their own prior beliefs.

In table 3 we present the results of the estimation of 8 different Markov Chain models with uniform priors for both $\mathbf{p}$ and $\mathbf{w}$.

Table 3. Markov Chain models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ORIGINAL_FICO	x					x		x
ORIGINAL_BALANCE	x	x	x					
ORIGINAL_LTV	x	x	x	x				x
HPA	x	x			x	x		
UNEMPLOY	x	x	x	x	x	x	x	x
ORIGINAL_COUPON	x	x	x	x				
Normalized Entropy $S(\mathbf{p})$	0.8879	0.8743	0.8754	0.8502	0.8532	0.8545	0.8493	0.8475

Consistent with our definition of $S(\mathbf{p})$, we select the best model based on the lowest value for the normalized entropy (i.e. the most informative model). This is model (8) with covariates Unemployment, Original LTV and Original FICO. The transition probability matrix for this model is:

Table 4. Transition probability matrix for model (8)

	C	30	60	90	120	REO	F
C	0.941	1.55e-3	1.579e-3	~0	~0	~0	~0
30	0.205	0.1907	0.1909	0.1068	6.54e-2	0.1936	4.729e-2
60	0.1617	0.1585	0.1585	0.1364	0.1796	0.1594	0.1074
90	0.1564	0.1527	0.1528	0.1392	0.127	0.1534	0.1977
120	0.1497	0.1468	0.1487	0.141	0.1365	0.149	0.129
REO	0.181	0.1742	0.1742	0.125	0.0934	0.177	0.0745
F	0.148	0.147	0.14705	0.1415	0.1363	0.1473	0.1328

Notice that the entries in the matrix, for which we don't have many data points such as 120 days past due are dominated by the uniform priors which clearly does not reflect reality. This is so because there isn't enough data to "move" the estimates away from the priors. This illustrates the dangers of choosing bad priors. A more reasonable prior in this case would be one that does not assign equal probabilities to all cells in the transition matrix but instead reflects the fact that some transitions of occur with very small probabilities. We don't do this here because our goal is to emphasize on the method rather than the results. And in any event a good prior should be provided by an expert.

One advantage of the Markov Chain model is that it easily incorporates prior information and additional covariates. However, one drawback is that it is more difficult to analyze transition probabilities for a particular loan or a subset of loans. Since, there is one transition probability matrix, this model is best suited for analyzing a portfolio as a whole. In the next section we address this difficulty by developing a multinomial loan-level model.

Multinomial Model

In this section we develop a multinomial model following Glennon and Golan (2003) who apply this framework to study loan discrimination. Our work is still within the generalized maximum entropy framework developed above. A major advantage of the multinomial model over the Markov one is that it is much easier to extract a loan-level forecasts for delinquency states. This means that the practitioner can use the results from one portfolio of mortgages and apply it to another one. However, in this presentation of the model we do not incorporate prior beliefs on the transition probabilities which is a downside that the Markov model does not suffer from. However, the lack of priors in this formulation does not cause the problem of having very inaccurate entries in the transition probability matrix due to inappropriately chosen priors. We proceed with a description of the theoretical model after which we present the results of the model run using our data.

For each loan i there are J possible outcomes (states) denoted by the binary variables $y_{i1}, y_{i2}, ..., y_{iJ}$. Therefore y_{ij} would equal 1 if outcome j was realized for loan i and zero otherwise. We can stack the outcome data in the $N \times T$-by-J boolean matrix $\mathbf{Y}$ where N is the number of loans in the sample and T is the number of periods for which we observe the loans. Notice that in this formulation we treat each loan in each period as a separate observation. The outcomes also depend on the K covariates $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{iK}]'$ which can include relevant loan characteristics such as credit score, national and local unemployment rates etc. Traditionally the model is formulated in the following way:

$$p_{ij} \equiv Prob(y_{ij} = 1|\mathbf{x}'_i, \beta_j) = G(\mathbf{x}'_i\beta_j) > 0$$

where β_j is a $(K \times 1)$ vector of unknown parameters and $G(.)$ is a function linking the unknown probabilities $\mathbf{p}$ with the covariates $\mathbf{x}$. This naturally gives rise to the following likelihood function:

$$L = \prod_{j=1}^{J} p_{1j}^{y_{1j}} p_{2j}^{y_{2j}} ... p_{Nj}^{y_{Nj}}$$

which can be transformed in the easier to work with logarithmic form:

$$\ln(L) = \sum_i \sum_j y_{ij} ln(p_{ij})$$

The traditional parametric method to recover the unknown probabilities p_{ij} is to assume some functional form for $G(.)$ and to maximize the likelihood with respect to the unknown parameters β. The most popular functional forms are the standard Gaussian cumulative distribution function or the logistic cumulative density function. For example if $G(.)$ is the logistic cdf then:

$$p_{ij} = \frac{\exp(\sum_k \beta_{jk} x_{ik})}{1 + \sum_{j=2}^{J} \exp(\sum_k \beta_{jk} x_{ik})} \equiv \frac{\exp(\sum_k \beta_{jk} x_{ik})}{\Omega_i(\beta)}$$

if we plug this back into the logarithmic likelihood function above we get:

$$\ln(L) = \sum_i \sum_j \sum_k y_{ij} \beta_{jk} x_{ik} - \sum_i \ln \Omega_i(\beta) \tag{12}$$

which after optimization gives us the unknown parameters β.

The traditional approach outlined above requires that the researcher specify the correct functional form $G(.)$. This is however rarely if ever known a priori. This makes the traditional model sensitive to the correct specification of the underlying data generating process. The maximum entropy approach developed by Jaynes (1957a, b) allows us to relax this requirement. Instead of specifying beforehand the functional form linking the probabilities p_{ij} and the data x_i we will select the "flattest" or least informative probability distribution that is consistent with data in our sample. To do this we need an objective function that measures the information in the probability distribution. We again select the criterion proposed by Shannon (1948) which is the "Shannon" entropy.

The model can then be reformulated as:

$$y_{ij} = p_{ij} + e_{ij} \tag{13}$$

where e_{ij} are the natural noise components. Since we want to work with probabilities we have to restate the noise e as a probability (the $\mathbf{p}$ are already in probability form). We can do this by defining a probability support $\mathbf{v}$ and a probability distribution $\mathbf{w}$. Then $e_{ij} \equiv \sum_d v_d w_{ijd}$ and $\sum_d w_{ijd} = 1$ where the errors support v is centered around zero and is contained in the interval $[-1, 1]$. This informational-based formulation doesn't incorporate any of the covariates $\mathbf{x}$. We can introduce them by adopting the approach from the instrumental variable literature. Following Courchane et al. (2000) the model can be reformulated as:

$$\sum_i y_{ij} x_{ik} = \sum_i x_{ik} p_{ij} + \sum_i x_{ik} e_{ij} \tag{14}$$

The additional covariates $\mathbf{x}$ are incorporated in the model through the cross moments. The generalized maximum entropy approach (GME) relies on finding the probability distributions $\mathbf{p}$ and $\mathbf{w}$ that satisfy the above constraint and the same time are least informative i.e. the probability distribution that satisfies the data at hand and has the lowest number of imposed assumptions. The measure of information is often chosen to be Shannon's entropy. Therefore we can recover the unknown probabilities by maximizing:

$$\max_{\mathbf{p},\mathbf{w}} \left\{ H(\mathbf{p}, \mathbf{w}) = -\sum_{ij} p_{ij} \ln p_{ij} - \sum_{ij} w_{ijd} \ln w_{ijd} \right\}$$

subject to the constraint in the previous equation.

Although modern numerical optimization techniques can solve problems of this type, this is a computationally intensive task which is prone to many errors. In order to reduce the dimensionality of the problem we can derive the dual formulation which transforms the constrained optimization into an unconstrained one. This is done by forming the Lagrangian and restating the constrained optimization problem above in terms of the Lagrange multipliers λ. After some manipulation the dual formulation is:

$$L(\lambda) = -\sum_{ijk} y_{ij} x_{ik} \lambda_{kj} + \sum_{i} \ln \Omega_i + \sum_{ij} \ln \Phi_{ij} \tag{15}$$

where $\Omega_i(\lambda) = 1 + \sum_{j=2}^{J} \exp(-\sum_k \lambda_{jk} x_{ik})$ and $\Phi_{ij}(\lambda) = \sum_d \exp(-\sum_k x_{ik} \lambda_{jk} v_d)$.

Note that if we don't incorporate the noise e in the model, then the likelihood function becomes

$$L(\lambda) = -\sum_{ijk} y_{ij} x_{ik} \lambda_{kj} + \sum_{i} \ln \Omega_i \tag{16}$$

which is equivalent to the multinomial logit model discussed earlier with $\beta = -\lambda$. However, not allowing the data to be noisy could be unjustified and too strict.

When applying this framework to the problem at hand we have to take into account the fact the probability of a loan transitioning to a particular state is not only dependent on the particular set of covariates for that loan but also on the current state of the loan. For example a loan that is 30 days past due has a much greater chance of transitioning to a 60 days past due than a loan that is current. In order to account for this fact we estimate the probability of transition conditional on the current state of the loan as well as the covariates. In order to do this, every time we condition on a specific state j we subset the sample so that only loans that were in state j are retained. In this way we can populate a transition matrix for a given set of covariates.

We study model fit, as in the Markov Chain model in the previous section by calculating the normalized entropy for the model. In the multinomial model the normalized entropy is $S(\mathbf{p}) = \left[\sum_{ij} p_{ij} \log p_{ij}\right] / (\log(J) \times N)$ with the similar interpretation as in the Markov formulation from the previous section: 0 indicates perfect knowledge and 1 perfect ignorance. The difference between the two definitions of the normalized entropy is that in the multinomial framework we don't measure the distance from a prior probability distribution because we don't have one.

In this setting there will be a different transition matrix row for each set of covariates Z. Each row of the matrix represents a different run of the model. In order to present the results we need to select some set of covariates because our predictions are not at the portfolio level but at the loan level. So, somewhat arbitrarily we choose to work with the following "representative" loan:

Table 5. Representative model

BALANCE	COUPON	FICO	LTV	HPA	UNEMPLOYMENT
1.314e+5	6.448	681.65	13.765	-0.1677	7.6

It is trivial to select a different set of covariates. Next we present a transition probability matrix for the "representative" loan. Each row of the matrix represents the probability of transition into the column state conditional on the representative loan being in the row state. As mentioned earlier, we do this by subsetting the data and examining only loans who were in the row state in the previous period. For example, the first row in Table 6 was estimated by only including loans that were "current" in the previous period. The reported normalized entropy $S(\mathbf{p})$ is for the probability distribution of the non-zero entries. Here are the results

Table 6.1. GME transition probabilities for representative loan (all covariates)

	C	30	60	90	120	REO	F	$S(\mathbf{p})$
C	0.699	0.300	0	0	0	0	0	0.941
30	0.3666	0.33683	0.2965	0	0	0	0	0.9931
60	0.0926	0.199	0.2526	0.455	0	0	0	0.9655
90	0.0208	0.3087	0.2589	0.215	0.1964	0	0	0.6659
120	0.0055	4.29e-8	0.07166	0.8979	0.0105	0	0.0144	0.3817
REO	0.237	0.0931	0	0	0	0.33991	0.32996	0.9711
F	0.0006	0.0065	0	0.000004	0.9896	0.00332	0.000002	0.4688

The zero entries in Table 6.1 represent transitions that did not occur. For example there were zero transitions from "30 days past due" to "90 days past due." Notice that the probabilities in this transition matrix do not reflect closely the actual data. For example, the first row in the matrix estimates that the probability of a "Current" loan staying "Current" (conditional on the covariates in Table 5) is only 69% while in the data this probability is much higher than that (see Table 2.2). This is because in the GME framework of (13) and (14) we allow for noisy estimates that deviate from the observed moments (in a model that allows the incorporation of priors this could be addressed by imposing a "tighter" prior on the noise probabilities). That is we don't force the expected cross moments to equal the observed ones. Instead we allow them to be flexible and we choose the most uniform distribution possible.

Next we estimate the same model and present the results for the same representative model of Table 5 but estimated within the maximum entropy (ME) framework and not the generalized maximum entropy (GME). As noted earlier, the ME framework is equivalent to estimating the traditional multinomial logit model and it differs from the GME framework in that we don't incorporate the noise element e. This imposes stricter requirements on the probabilities than the GME. Since this framework requires that the cross moments in the data equal exactly, our transition probability estimates follow the actual data much more closely:

Table 6.2. ME transition probabilities for representative loan (all covariates)

	C	30	60	90	120	REO	F	$S(\mathbf{p})$
C	0.994	0.0062	0	0	0	0	0	0.0448
30	0.41	0.451	0.14	0	0	0	0	0.8672
60	0.076	0.481	0.24	0.2	0	0	0	0.8502
90	0.004	0.0003	0.27	0.2	0.523	0	0	0.5746
120	0.017	~0.0	~0.0	0.001	0.41	0	0.573	0.535
REO	0.135	0.006	0	0	0	0.854	0.005	0.09
F	0.005	0.008	0	~0.0	0.08	0.748	0.159	0.542

Conclusion

In this paper we propose two different approaches for estimating mortgage delinquency probabilities. The first model we develop is a Markov Chain model where the transition probabilities are estimated using the generalized maximum entropy framework. The Markov Chain model has often been applied by researchers to similar problems because it naturally conditions the transition probabilities on the current state of the loan/mortgage. In addition we show how to condition on other covariates when estimating the transition probability matrix. Furthermore, the maximum entropy formulation easily incorporates noisy data (so that the cross moments are not forced to be exactly equal to the observed data) and prior beliefs. Allowing for the incorporation of prior beliefs is a major advantage because for many entries of the transition probability matrix, there are not enough observations for obtaining robust estimates.

Our second model is a loan-level model where we estimate the probability of transition separately for each model in the data. We propose a multinomial model which is estimated again using the maximum entropy principles. Although, in its current version, this model doesn't allow for the easy incorporation of prior beliefs, its major advantage is that it is quite easy to produce and interpret the loan-level predictions and apply these results to different portfolios.

References

[1] Betancourt L. (1999)"Using Markov chains to estimate losses from a portfolio of mortgages." Review of Quantitative Finance and Accounting 12, 303-317.

[2] Courchane M., Golan A., Nickerson D., (2000) "Estimation and evaluation of loan discrimination: an informational based approach." Journal of Housing Research 11(1),

[3] Cyert R., Davidson H., Thompson G. (1962)."Estimation of the allowance for doubtful accounts by Markov chains." Management Science 8, 287-303.

[4] Glennon, D., Golan A., (2003). "A Markov model of bank failure estimated using an information-theoretic approach." OCC Economics Working Paper 2003-1

[5] Golan, A., Judge G., Perloff J., (1996) "A maximum entropy approach to recovering information from multinomial response data." Journal of the American Statistical Association. 91(434). 841-853

[6] Golan, A. (2008). "Information and entropy econometrics - A review and synthesis." Foundations and Trends in Econometrics 2(1-2), 1-145.

[7] Grimshaw, S., Alexander, W. (2011). "Markov chain models for delinquency: Transition matrix estimation and forecasting." Applied Stochastic Models in Business and Industry 27 (3), 267-279.

[8] Jaynes, E.T. (1957a). "Information theory and statistical mechanics." Physical Review 106(4), 620-630.

[9] Jaynes, E.T. (1957b). "Information theory and statistical mechanics II." Physical Review 108(2), 171-190.

[10] Shannon, C.E. (1948). "A mathematical theory of communication." Bell System Technical Journal 27, 379-423 & 623-656.