M3/4S17 Quantitative Methods in Retail Finance

Chapter 3: Markov transition models

Overview

- At the end of the Credit Scoring 1 course we briefly discussed how Markov transition models could be used for behavioural modelling.
- In this chapter we shall look at Markov transition models in more detail.

We shall cover the following topics:-

1. Markov transition models;
2. Estimation;
3. Testing first-order assumptions;
4. Roll-rate model.

Markov transition models

Markov transition models (or *Markov chains*) are a dynamic approach to modelling processes with changes of state.

They are valuable in credit scoring since they allow us to model changes in the state of an account over time. For instance,

- Modelling the number of account periods of delinquency.
- Changes in behavioural score.

Markov transition models are especially useful for modelling revolving credit with highly variable credit usage.

- For instance, for tracking credit card use.

First-order Markov transition model

Some definitions:

- Let $X_0, X_1, X_2, \ldots$ be a sequence of discrete random variables taking values from $\{1,2,\ldots,K\}$ for some fixed K.
- The sequence is a ***finite-valued first-order Markov chain*** if

$$P(X_{n+1} = j \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = i\,) = P(X_{n+1} = j \mid X_n = i\,)$$

  for all n, $x_0, \ldots, x_{n-1}$ and i,j such that $1 \le i \le K$ and $1 \le j \le K$.

- Define the ***transition probability*** $p_n(i,j) \triangleq P(X_n = j \mid X_{n-1} = i\,)$.
  - The transition probability represents the probability of moving from one state i to another state j.
- The ***transition matrix*** P_n is defined as a $K \times K$ matrix such that $P_n[i,j] \triangleq p_n(i,j)$.

- If we make a prior assumption that there is no transition from state i to state j in the nth period then we fix $p_n(i,j) = 0$ and call this a ***structural zero***.

Propagation of transition probabilities

- Transitional probabilities can be propagated so a conditional probability of state at some period n can be estimated from a starting state n-r.
- This is important if we want to use Markov transition models for forecasting.

Theorem

$$P(X_n = j \mid X_{n-r} = i) = (P_{n-r+1}P_{n-r+2} \dots P_n)[i,j]$$

for any $1 \leq r \leq n$.

Proof is in the Appendix at the end of this chapter.

The special case enables forecasts from state 0:

$$P(X_n = j \mid X_0 = i) = (P_1 P_2 \dots P_n)[i, j]$$

Let π_n be the marginal distribution of X_n:

$$\pi_n = (P(X_n = 1\,), \dots, P(X_n = K\,)\,)^T$$

Then, since $P(X_n = j\,) = \sum_{i=1}^{K} P(X_n = j \mid X_0 = i\,) P(X_0 = i)$,

$$\pi_n = \pi_0 (P_1 P_2 \dots P_n)$$

Stationary Markov chains

- A Markov chain is ***stationary*** if $p_n(i,j) = p(i,j)$ for all n, i and j, for some transition probability p.

- A ***stationary distribution*** for transition matrix P is a distribution π^* such that $\pi^* = \pi^* P$.

- In practice, most Markov chains converge (with n) to a stationary distribution.

  (Markov chains which have a periodicity in state change do not necessarily converge, but we do not cover this material in this course).

Example 3.1

Consider a two state stationary Markov chain for behavioural score change (state 1=high score, 2=low score) with transition matrix

$$P = \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$$

Suppose we start with an individual having high score.

1. What are the distributions after one and two periods?
2. What is the stationary distribution which this process converges to?

Solution

$\pi_0 = (1 \quad 0)$

Therefore, after one period:

$$\pi_1 = \pi_0 P = (1 \quad 0)\begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix} = (0.95 \quad 0.05)$$

And, after two periods,

$$\pi_2 = \pi_0 P^2 = (0.95 \quad 0.05)\begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix} = (0.9075 \quad 0.0925)$$

Convergence:

Let $\pi^* = (a \quad b)$ with $a + b = 1$.
Then $(a \quad b) = (a \quad b)\begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$.
So $a = 0.95a + 0.1b$ and $b = 0.05a + 0.9b$.
Therefore $\pi^* = (2/3 \quad 1/3)$.

Estimation of the transition matrix

Use maximum likelihood estimation (MLE) for each $p(i,j)$.

Estimate parameters $\boldsymbol{\theta} = \left(\theta_{ij}\right)_{i,j=1}^{K}$ where each $\theta_{ij} = p(i,j)$.

Given a sequence of n realizations $x_0, x_1, \ldots, x_n$, the probability of this realization is given as

$$P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \left(\prod_{t=1}^{n} P(X_t = x_t | X_0 = x_0, \ldots, X_{t-1} = x_{t-1})\right) P(X_0 = x_0)$$

$$= P(X_0 = x_0) \prod_{t=1}^{n} P(X_t = x_t | X_{t-1} = x_{t-1})$$

$$= P(X_0 = x_0) \prod_{t=1}^{n} p(x_{t-1}, x_t)$$

$$= P(X_0 = x_0) \prod_{i=1}^{K} \prod_{j=1}^{K} [p(i,j)]^{n_{ij}}$$

$$\Rightarrow P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = P(X_0 = x_0) \prod_{i=1}^{K} \prod_{j=1}^{K} [\theta_{ij}]^{n_{ij}}$$

where $n_{ij} \triangleq \left|\{t: x_{t-1} = i, x_t = j \text{ for } t \in \{1, \dots, n\}\}\right|$

(ie number of observed moves from i to j).

Therefore, the log-likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n | \boldsymbol{\theta}) \\ &= \log P(X_0 = x_0) + \sum_{i=1}^{K} \sum_{j=1}^{K} n_{ij} \log \theta_{ij} \end{aligned}$$

with constraint $\sum_{j=1}^{K} \theta_{ij} = 1$.

Therefore, choose some r such that $1 \leq r \leq K$ and substitute

$$\theta_{ir} = 1 - \sum_{j \in \{1,\ldots,K\} \setminus \{r\}} \theta_{ij}$$

to get

$$L(\boldsymbol{\theta}) = \log P(X_0 = x_0) + \sum_{i=1}^{K} \left[\left(\sum_{j \in \{1,\ldots,K\} \setminus \{r\}} n_{ij} \log \theta_{ij} \right) + n_{ir} \log \left(1 - \sum_{j \in \{1,\ldots,K\} \setminus \{r\}} \theta_{ij} \right) \right]$$

Then find the derivative with respect to each θ_{ij} where $j \neq r$ and set to zero to find the maxima:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_{ij}} = \frac{n_{ij}}{\theta_{ij}} + \frac{n_{ir}}{\theta_{ir}} \times -1 = 0$$

Treat i as fixed, hence $\frac{\widehat{\theta}_{ir}}{n_{ir}}$ is a constant for all j. Therefore,

$$\hat{\theta}_{ij} = n_{ij} \frac{\hat{\theta}_{ir}}{n_{ir}} \propto n_{ij}$$

But, the choice of r is arbitrary so for consistency the result must hold generally for all j.

In particular, the MLE is

$$\hat{p}(i,j) = \hat{\theta}_{ij} = \frac{n_{ij}}{\sum_{l=1}^{K} n_{il}}$$

Notice that this result easily generalizes to the case when we have multiple sequences of realizations (eg more than one borrower), so long as we assume independence between each sequence.

Side note 1: If $n_{ij} = 0$, then $\hat{\theta}_{ij} = 0$, but it looks like the log-likelihood on slide 12 is undefined in this case, since $\log \hat{\theta}_{ij} = -\infty$. However, notice that as $n_{ij} \to 0$, $n_{ij} \log \hat{\theta}_{ij} \to 0$, so in fact it is fine.

Example 3.2

Consider three states (1=good customer, 2=delinquent, 3=default) for a stationary process. Transition probabilities are given as:

- from good customer to delinquent is 0.05;
- from delinquent to good customer is 0.1;
- from delinquent to default is 0.02.

It is impossible to move from being a good customer to default in one period. Also, it is impossible to move out of default.

1. What is the transition matrix?
2. How many structural zeroes are there in the matrix?

Solution

$$P = \begin{pmatrix} 0.95 & 0.05 & 0 \\ 0.1 & 0.88 & 0.02 \\ 0 & 0 & 1 \end{pmatrix}$$

There are 3 structural zeroes.

Testing the first-order assumption

The first-order Markov process assumption is a strong assumption and needs to be tested.
To do this we model using a *second-order* Markov chain to see if this is statistically significantly different from a first-order Markov chain.

Definition
A sequence is a **finite-valued second-order Markov chain** if

$$P(X_{n+1} = j \mid X_0 = x_0, \dots, X_{n-1} = k, X_n = i\,) = P(X_{n+1} = j \mid X_n = i\,, X_{n-1} = k)$$

for all n and i,j,k such that $1 \le i \le K$, $1 \le j \le K$ and $1 \le k \le K$.

We denote the second order transition probability as

$$p_n(k,i,j) \triangleq P(X_n = j \mid X_{n-1} = i\,, X_{n-2} = k)$$

The MLE for a second-order Markov chain is found in a similar way to the first-order MLE, therefore second-order transition probabilities can be estimated.

Specifically,

$$\hat{p}(k,i,j) = \hat{\theta}_{kij} = \frac{n_{kij}}{m_{ki}}$$

where $n_{kij} \triangleq \left|\left\{t:, x_{t-2} = k, x_{t-1} = i, x_t = j \text{ for } t \in \{2, \dots, n\}\right\}\right|$

(number of observations moving from state k to i to j)

and $m_{ki} \triangleq \left|\left\{t:, x_{t-2} = k, x_{t-1} = i \text{ for } t \in \{2, \dots, n\}\right\}\right| = \sum_{l=1}^{K} n_{kil}$

(number of observations moving from state k to i in the first $n-2$ periods).

Using the Pearson chi-square test of independence

We set up a null hypothesis that corresponds to the first-order Markov chain for a state i. That is, we test whether the transition probability to any state j is the same for all k:

$$H_0(i): p(1,i,j) = p(2,i,j) = \cdots = p(K,i,j) = p(i,j) \text{ for all } j \in \{1,\ldots,K\}.$$

The Pearson goodness-of-fit χ^2 statistic is then used to test $H_0(i)$.

- Let $n_k \triangleq \left|\left\{t:, x_t = k \text{ for } t \in \{1,\ldots,n\}\right\}\right|$
  (the number of observed state k).

- The observed number of times state k is followed by i then j is
$$O_{kj} = n_{kij} = m_{ki}\hat{p}(k,i,j).$$

- The expected number of times state k is followed by state i and then state j, **given the null hypothesis**, is
$$E_{kj} = n_k\hat{p}(k,i)\hat{p}(i,j) \approx m_{ki}\hat{p}(i,j).$$

Here we make use of the fact that when n is large,

$$\hat{p}(k,i) = \frac{n_{ki}}{\sum_{l=1}^{K} n_{kl}} \approx \frac{m_{ij}}{n_k}$$

since an estimate based on transitions from $n-1$ periods will be approximately the same as that from $n-2$ periods.

- So the Pearson chi-square test is

$$\chi^2 = \sum_{k \in S_1} \sum_{j \in S_2} \frac{\left(O_{kj} - E_{kj}\right)^2}{E_{kj}}$$

$$\approx \sum_{k \in S_1} \sum_{j \in S_2} \frac{\left(m_{ki}\hat{p}(k,i,j) - m_{ki}\hat{p}(i,j)\right)^2}{m_{ki}\hat{p}(i,j)}$$

$$= \sum_{k \in S_1} \left[m_{ki} \sum_{j \in S_2} \frac{\left(\hat{p}(k,i,j) - \hat{p}(i,j)\right)^2}{\hat{p}(i,j)} \right]$$

where

- $S_1 \subseteq \{1, \ldots, K\}$ is the set of states which do *not* have a structural zero moving *to* state i, and
- $S_2 \subseteq \{1, \ldots, K\}$ is the set of states which do *not* have a structural zero moving *from* state i.

- So if there are no structural zeroes, χ^2 is just a nested sum over all states.

- This is a chi-square test over a contingency table of K^2 elements. However, since each sum (inner and outer) must sum to number of observations, lose one degree of freedom within square: ie degrees of freedom df = $(K-1)^2$.

If the null hypothesis is rejected (low p-value), then the first-order Markov chain assumption is rejected.

- Degrees of freedom when structural zeroes are in the transition matrix:
  - Let z_1 be the number of structural zeroes in the transition from k to i.
  - Let z_2 be the number of structural zeroes in the transition from i to j.
  - Then df=$(K-1-z_1)(K-1-z_2)$.

  where n_{zero} is the number of structural zeroes in the transition matrix.

Illustration of number of degrees of freedom:

		j			
		1		Structural zero i to j	K
	1			0		All rows must sum to same number m_{ki}
k	Structural zero k to i	0	0	0	0	
				0		
K		All columns must sum to same number			

Only the blue regions contribute to degrees of freedom.

Example 3.3

Consider the first-order Markov chain in Example 3.2.

For state $i = 1$, we have the following estimates for a second-order Markov chain:

- $\hat{p}(1,1,1) = 0.97, \quad \hat{p}(2,1,1) = 0.9$
- $\hat{p}(1,1,2) = 0.03, \quad \hat{p}(2,1,2) = 0.1$

All other second-order transition probabilities, relating to default, are structurally zero.

We also have the following numbers of observations in the data:
$m_{1,1} = 100$, $m_{2,1} = 12$ and $m_{3,1} = 0$.

Test the first-order Markov chain null hypothesis $H_0(1)$ using the Pearson chi-square test at a significance level of 5%.

Solution

There are only four transition probabilities in the sum due to the structural zeroes. Therefore,

$$\begin{aligned}\chi^2 &= m_{1,1}\left[\frac{\left(\hat{p}(1,1,1)-\hat{p}(1,1)\right)^2}{\hat{p}(1,1)}+\frac{\left(\hat{p}(1,1,2)-\hat{p}(1,2)\right)^2}{\hat{p}(1,2)}\right]\\ &+m_{2,1}\left[\frac{\left(\hat{p}(2,1,1)-\hat{p}(1,1)\right)^2}{\hat{p}(1,1)}+\frac{\left(\hat{p}(2,1,2)-\hat{p}(1,2)\right)^2}{\hat{p}(1,2)}\right]\\ &= 100\left[\frac{(0.97-0.95)^2}{0.95}+\frac{(0.03-0.05)^2}{0.05}\right]+12\left[\frac{(0.9-0.95)^2}{0.95}+\frac{(0.1-0.05)^2}{0.05}\right]\\ &\approx 1.47\end{aligned}$$

From Example 3.2,

- it is impossible to move from 3 to 1, hence $z_1 = 1$ and
- it is impossible to move directly from 1 to 3, hence $z_2 = 1$.

Therefore df=$(K-1-z_1)(K-1-z_2)=1$.

With 1 degree of freedom, this has a p-value approximately 0.22.
Therefore the null hypothesis is not rejected.

Hence, it is reasonable to assume a first-order Markov chain.

Extensions to Markov transition models

An obvious omission from the Markov chain formulation is the lack of predictor variables.

There are two ways to include borrower details in the model:

1. Include behavioural variables within the state space.

2. Segment the population on static variables and build segmented Markov transition models.

Both methods suffer from similar problem:

1. Increasing the state space means more transition probabilities need to be estimated and this will mean reduced estimation efficiency.

2. Segmentation will mean several distinct Markov chains, each based on a reduced training sample.

3. Neither method allows for continuous data, unless it is discretized, and there is a limit to the number of categorical variables that can be used in states or separate models.

Example 3.3

Suppose we want to include credit usage, in terms of monthly spend in a model for behavioural score (Low or High).

- First discretize credit usage into levels:
  eg three levels: monthly spend < £200, ≥£200 and < £1000, ≥£1000,

- Then, form 6 states, instead of 2:

Behavioural score	Monthly spend	State
Low	< £200	1
Low	≥£200 and < £1000	2
Low	≥£1000	3
High	< £200	4
High	≥£200 and < £1000	5
High	≥£1000	6

Example 3.4

Research suggests two broad categories of credit card usage: the movers and stayers.

- ***Movers*** are those whose credit card usage is erratic; having periods of heavy credit card usage then quiet periods.

- ***Stayers***, by contrast, tend to be steady, and stay in the same state over long periods.

We could build a static behavioural model to broadly categorize borrowers into one of the two categories.

Then separate Markov transition models could be built separately for the two segments.

Roll-rate model

A roll-rate model is a type of Markov transition model but the focus is on the number of accounts or value of loans that rolls over from one level of delinquency to another over several months.

- Consider K states where 0 corresponds to no delinquency, states >0 correspond to increasing levels of delinquency and K corresponds to loan default with write-off.
- Let A be a vector of initial number of accounts or value of loans.
- Let P be a $K \times K$ transition matrix.

Then the vector of values in each state at month t is given by AP^t.

Example 3.5

Let $K = 3$.

Let $A = (50000, 10000, 5000, 1000)$, in GB£.

Let $P = \begin{pmatrix} 0.98 & 0.02 & 0 & 0 \\ 0.3 & 0.4 & 0.3 & 0 \\ 0.1 & 0.05 & 0.55 & 0.3 \\ 0 & 0 & 0 & 1 \end{pmatrix}$.

Let first month ($t = 0$) be January 2013.

Then roll-rate table (*projection*) for six months is computed as:-

Month	Computation	State			
		0	1	2	3
Jan 13	A	50000	10000	5000	1000
Feb 13	AP	52500	5250	5750	2500
Mar 13	AP^2	53600	3438	4737	4225
Apr 13	AP^3	54033	2684	3637	5646
May 13	AP^4	54121	2336	2805	6737
Jun 13	AP^5	54020	2157	2244	7579

Chapter review

We have covered the following topics:-

1. Markov transition models;
2. Estimation;
3. Testing first-order assumptions;
4. Roll-rate model.

Appendix: Proof of Theorem on Slide 6

This is background material on Markov transition models and is not examinable.

Theorem (Chapman-Kolmogorov, essentially)

$$P(X_n = j \mid X_{n-r} = i) = (P_{n-r+1}P_{n-r+2} \dots P_n)[i,j]$$

for any $1 \leq r \leq n$.

Proof is by induction (via two Lemmas):

Lemma 1.

$$P(X_n = j | X_{n-1} = i, A) = p_n(i,j)$$

for any $A = (X_i = x_i)_{i \in R}$ where $R \subseteq \{0, \dots, n-2\}$.

Note: the notation $(X_i = x_i)_{i \in S}$ represents the concatenation of conditions $X_i = x_i$ across all indexes $i \in S$.

Proof

By the law of total probability,

$$P(X_n = j | X_{n-1} = i, A) = \sum_{B \in S} P(X_n = j, B | X_{n-1} = i, A)$$

where S is the set of all possible $B = (X_i = x_i)_{i \in \{0,\dots,n-2\} \setminus R}$ across all possible values of x_i's (hence members of S represent exclusive events and are exhaustive).

Then

$$P(X_n = j|X_{n-1} = i, A) = \sum_{B \in S} P(X_n = j|X_{n-1} = i, A, B)\, P(B|X_{n-1} = i, A)$$
$$= \sum_{B \in S} P(X_n = j|X_{n-1} = i)\, P(B|X_{n-1} = i, A)$$

by the first-order Markov assumption, since A and B together contain events for all X_0 to X_{n-2}.

Therefore

$$P(X_n = j|X_{n-1} = i, A) = P(X_n = j|X_{n-1} = i) \sum_{B \in S} P(B|X_{n-1} = i, A)$$
$$= P(X_n = j|X_{n-1} = i)$$

since S is defined as the set of all possible events B.

Lemma 2.

$$P(X_{n+r} = j \,|X_n = k, X_{n-1} = i) = P(X_{n+r} = j \,|X_n = k)$$

Proof by induction:

When $r = 1$, applying lemma 1,

$$P(X_{n+1} = j \,|X_n = k, X_{n-1} = i) = P(X_{n+1} = j \,|X_n = k)$$

Suppose true for $r = s \geq 1$. Then, for $r = s + 1$,

$$P(X_{n+s+1} = j \,|X_n = k, X_{n-1} = i) = \sum_{v=1}^{K} P(X_{n+s+1} = j \,, X_{n+s} = v | X_n = k, X_{n-1} = i)$$

$$= \sum_{v=1}^{K} P(X_{n+s+1} = j \,|X_{n+s} = v, X_n = k, X_{n-1} = i)\, P(X_{n+s} = v | X_n = k, X_{n-1} = i)$$

$$= \sum_{v=1}^{K} P(X_{n+s+1} = j \,|X_{n+s} = v)\, P(X_{n+s} = v|X_n = k)$$

applying lemma 1 and induction.

$$= \sum_{v=1}^{K} P(X_{n+s+1} = j \,|X_{n+s} = v, X_n = k)\, P(X_{n+s} = v|X_n = k)$$

applying lemma 1 again (in reverse).

Therefore

$$P(X_{n+s+1} = j \,|X_n = k, X_{n-1} = i) \quad = \sum_{v=1}^{K} P(X_{n+s+1} = j\,, X_{n+s} = v|X_n = k)$$
$$= P(X_{n+s+1} = j \,|X_n = k)$$

as required.

Proof of main result, by induction:

When $r = 1$: by definition,

$$P(X_n = j \mid X_{n-1} = i) = P_n[i, j].$$

Now suppose the theorem is true for $r = s \geq 1$.

Then for $r = s + 1$, using law of total probability,

$$\begin{aligned} P(X_n = j \mid X_{n-s-1} = i) &= \sum_{k=1}^{K} P(X_n = j\,, X_{n-s} = k \mid X_{n-s-1} = i) \\ &= \sum_{k=1}^{K} P(X_n = j \mid X_{n-s} = k, X_{n-s-1} = i) P(X_{n-s} = k \mid X_{n-s-1} = i) \\ &= \sum_{k=1}^{K} P(X_n = j \mid X_{n-s} = k) P(X_{n-s} = k \mid X_{n-s-1} = i), \text{ by lemma 2} \\ &= \sum_{k=1}^{K} P_{n-s}[i,k](P_{n-s+1}P_{n-s+2} \dots P_n)[k,j], \text{ by induction step} \\ &= (P_{n-s}P_{n-s+1}P_{n-s+2} \dots P_n)[i,j] \end{aligned}$$

as required.