Journal of the Operational Research Society (2013), 1–8

www.palgrave-journals.com/jors/

Creditworthiness dynamics and Hidden Markov Models

L Quirini[1*] and L Vannucci[2]

[1]*Consum.it—Monte dei Paschi di Siena Group, Calenzano, Italy; and* [2]*University of Florence, Florence, Italy*

A dynamic monitoring of credit risky portfolios is described. In the first section, it is shown how a Markov dependence can be used in modelling the borrower's behaviour: a chain of transition probabilities matrices is built in which the states of the dynamic stochastic system are the number of instalments in arrears. In the second part, such a model is generalized in the framework of the Hidden Markov Models to explain how the credit market conditions could affect the borrower's payment process. Numerical examples complete the note.

Journal of the Operational Research Society advance online publication, 1 May 2013
doi:10.1057/jors.2012.181

Keywords: creditworthiness; Hidden Markov Models; simulation; statistical inference

Introduction

The last quarter of 2008 saw the beginning of the worst recession in the western economies after the Great Depression. Since then, credit control has become one of the most important issues for the financial industry.

Under the current macroeconomic environment, the credit companies should pay more attention not only to 'bad' customers but also to those who before the crisis had been evaluated as 'good' customers in terms of credit score ratings. Nowadays, also customers with a high creditworthiness at the beginning of the relationship may show a sudden downgrading in their creditworthiness.

The authors believe that credit companies have to achieve and use, with continuity and awareness, dynamic tools to assess credit risk in order to gain profitability in such complex and fluctuating markets. Among the suitable models for this aim, Markov chains could give valuable results especially for the financial products with a given amortization plan.

In the specialized literature on consumer credit analysis, Markov chains were introduced by the pioneering work of Cyert *et al*, 1962. This line of research has been further developed by several authors (see, for instance, Kallberg and Saunders (1983). For a comprehensive report on the main approaches to this subject until the year 2000 see Thomas *et al* (2002). The issue was also taken up by more recent papers (Malik and Thomas, 2009).

In recent works (Quirini and Vannucci, 2011, 2012), the authors use Markov chains to evaluate the moments of the random present value of the actually paid instalments and they have regarded such a model as a further explanation of their Creditworthiness Index (Quirini and Vannucci, 2010).

In the framework of the Hidden Markov Models (HMM), the work of Giampieri *et al* (2005) is noteworthy, in which an approach to estimate default frequencies was considered.

In this paper, which takes inspiration from this line of research, the aim is to link HMM with the analysis of creditworthiness in retail portfolios by means of Markov chains.

If the macroeconomic conditions are taken into account in assessing the borrowers' creditworthiness, there will be the problem of selection of factors to consider. Regardless of the way this selection is made, it has wide margins of error because the macroeconomic indicators could give false information: good signals make economic agents too optimistic and bad ones too pessimistic (pro-cyclical effects). For instance, in economic growth periods many customers take on more credit, thus producing the conditions for an increment in loan delinquency rates as soon as their expectations worsen.

This study wants to overcome the problem of the identification of external macroeconomic factors that may influence the customer's creditworthiness, and it focusses on inferring these factors from internal data.

In detail, in the first section a Markov chain has been analysed, the parameters of which are used to explain the probabilities that a borrower will be in one of the possible states of the chain in a sequence of periods.

[*]*Correspondence: L Quirini, Consum.it Spa, via Vittorio Emanuele 10, Calenzano, Firenze 50041, Italy.*
E-mail: lorenzo.quirini@consum.it

Such states are identified by the number of instalments in arrears and by a default state that happens when the number of instalments in arrears reaches a given threshold value (eg at least three instalments in arrears).

In the second section, an HMM is built: it is assumed that the set of the different credit market conditions, or states, is finite and that the credit market state changes according to a Markov chain. At each credit market state a particular Markov chain is associated to describe, for each customer, the stochastic process of the number of instalments in arrears. Therefore, it is considered a stochastic Markov process, which depends on a second stochastic Markov process whose states are the not observable credit market conditions. In this section, we have also sketched out a method to estimate HMM from data.

The described model, applied in an operative context, may promptly give to the credit risk managers useful information, at least in probabilistic terms, about the underlying changes in the credit market conditions, enabling them to adopt the necessary and appropriate adjustments in credit control and management.

A behavioural model for the borrower's payments sequence

The model without default

For a financial obligation with an amortization plan beginning at time 0, it is expected that the granted amount c will be refunded by means of a sequence of instalments, $r_1, r_2, \ldots, r_n$, paid at equally spaced times $1, 2, \ldots, n$.

Between the granted amount c and the sequence of instalments, $r_1, r_2, \ldots, r_n$, the following relationship holds:

$$c = r_1(1+x)^{-1} + r_2(1+x)^{-2} + \cdots + r_n(1+x)^{-n}$$

where x is the internal rate of return of the obligation.

In this paper, the case, frequently used in consumer finance in which all the instalments are equal, is analysed. That is:

$$r_1 = r_2 = \cdots = r_n = r$$

The actual instalments are random variables and we assume that

$$R_h = J_h \cdot r$$

J_h being an integer variable, $J_h \in \{0, 1, \ldots, h\}$ and $h = 1, 2, \ldots, n, \ldots$

At each time h, the borrower can pay nothing, $J_h = 0$, or pay a positive multiple of r to recover, if $J_h \neq 1$, some instalments in arrears. With this assumption, in an extreme case, if a borrower before time h has paid nothing and at time h he pays $h \cdot r$ then he returns 'regular', even if it is not true from a financial point of view, unless the contractual internal rate is null.

This situation can be described by means of a simple generalization of a Markov chain. Instead of a fixed set of states, here the set of states is augmented by one state at each time: at time h, for $h = 0, 1, \ldots, n, \ldots$, the number of instalments in arrears can be $0, 1, \ldots, h$.

The proposed model forecasts the number of delinquent loans and their degree of insolvency, expressed in terms of the number of instalments in arrears; the comparison between such forecasts and the observed ones allows risk managers to act promptly.

For a borrower, let S_h, the random state at time $h = 1, \ldots, n, \ldots$ with $S_h \in \{0, 1, \ldots, h-1\}$, which gives the number of instalments in arrears before payment at time h. For S_1, it holds

$$P(S_1 = 0) = 1$$

The transition probabilities matrix, P_h, which is associated with the borrower's behaviour at time h, for $h = 1, 2, \ldots, n, \ldots$, can be considered a matrix with h rows and $h+1$ columns; that is:

$$P_h = \begin{pmatrix} p_{h;0;0} & p_{h;0;1} & 0 & 0 & \cdots & 0 & 0 \\ p_{h;1;0} & p_{h;1;1} & p_{h;1;2} & 0 & \cdots & 0 & 0 \\ & & & & \cdots & & \\ p_{h;h-1;0} & p_{h;h-1;1} & p_{h;h-1;2} & & \cdots & p_{h;h-1;h-1} & p_{h;h-1;h} \end{pmatrix}$$

where $p_{h;i;k}$ is the probability that at time h a financial obligation, which has i ($i = 0, 1, \ldots, h-1$) instalments in arrears just before this time, will have k instalments in arrears after h. The system translates from $S_h = i$ to $S_{h+1} = k$ and the borrower pays in h a total amount equal to $J_h \cdot r = (i - k + 1) \cdot r$, with $k = 0, 1, \ldots, i+1$. $S_{h+1} = 0$ means that after time h the obligation has no instalments in arrears; $S_h = i$ and $S_{h+1} = i+1$ indicate that at time h the borrower has paid nothing and the state of the chain has gone from i instalments in arrears to $i+1$.

Obviously the model assumes the fact that the borrower cannot prepay his financial obligation.

Let for $h = 1, \ldots, n+1, \ldots$ p_{h-1} be the vector with h elements that describes the probability distribution of S_h. It is

$$p_0 = (1) = P(S_1 = 0)$$

and recursively

$$p_h = p_{h-1} P_h$$

for $h = 1, 2, \ldots, n, \ldots$

An example of a model without default

From the operational point of view, the transition probabilities matrices P_h can be estimated by means of internal data and such matrices could depend on h.

For the sake of simplicity, an idealized case is considered. The model uses, in order to describe the borrower's behaviour, two parameters, a and b, both expressing probabilities, hence with conditions $a \geq 0$, $b \geq 0$ and $a + b \leq 1$.

Let for $h = 1, 2, 3, \ldots$

$$p_{h;0;0} = 1 - a; p_{h;0;1} = a$$

and for $h = 2, 3, \ldots$

$$p_{h;i;k} = \begin{cases} 1 - a - b; & \text{if } k = i \\ a; & \text{if } k = i + 1 \\ b; & \text{if } k = i - 1 \\ 0; & \text{otherwise} \end{cases}$$

In this case,

$$p_0 = (1)$$

$$p_1 = (1)(1 - a; a) = (1 - a; a)$$

$$p_2 = (1 - a; a) \begin{pmatrix} 1 - a & a & 0 \\ b & 1 - a - b & a \end{pmatrix} = ((1 - a)^2 + ab; a(1 - a) + a(1 - a - b); a^2)$$

and so on.

This model is a simplified description of the borrower's behaviour; nevertheless, it is suitable to explain his creditworthiness using only two parameters: a is linked to the risk that the borrower cannot pay his financial obligations, while b shows the capability of the borrower to recover one of the instalments in arrears.

Table 1 shows the probability distributions of the number of instalments in arrears, after times h with $a = 0.05$ and $b = 0.04$.

It is noteworthy that, at time 9, 67.66% of the borrowers, on average, will be regular on the payments due.

The model with default

The previous model can be adapted to the case where one of the states of the chain is absorbing and this state represents the default of the borrower.

If it is assumed that the financial obligation becomes a default one, when a given threshold, h^*, in terms of instalments in arrears has been reached (eg $h = 3$), then it is necessary to adjust the transition probabilities matrices putting for $h \geq h^*$

$$p_{h;h^*;h^*} = 1$$

The probability transition matrices after time h^* are all squared ones with $h^* + 1$ rows and $h^* + 1$ columns, being $\{0, 1, \ldots, h^*\}$ the set of the states, with h^* the absorbing (default) one.

Obviously the vectors and the transition probabilities matrices could all be considered, particularly in the default case, with a fixed number of elements over the entire time horizon.

In the model with default, the problem related to the estimation of the recovery rate for the defaulted transactions could be posed, but this problem is not analysed here.

An example for the default case

Table 2 shows the probability distributions of the number of instalments in arrears and of the default state after times h with $a = 0.05$, $b = 0.04$ and $h^* = 3$.

It is noteworthy that, at time 9, 0.74% of the borrowers will have reached, on average, the default state.

The Markovian payment process linked to the credit market conditions that evolve according to a Hidden Markov Process

The model with default

The previous model with default, embedded in hidden credit market conditions, is analysed. In particular, the model assumes that the transition probabilities among the number of instalments in arrears, which are observed at equally spaced time intervals, could depend on the credit market conditions.

Table 1 Probabilities × 10 000 (round off to the nearest integer) of the vector p_h for a=0.05 and b=0.04

h										
0	10 000									
1	9500	500								
2	9045	930	25							
3	8630	1300	69	1						
4	8250	1617	128	5	0					
5	7902	1889	198	11	0	0				
6	7583	2122	275	20	1	0	0			
7	7289	2321	357	32	2	0	0	0		
8	7017	2491	442	47	3	0	0	0	0	
9	6766	2635	529	65	5	0	0	0	0	0

Table 2 Probabilities × 10 000 (round off to nearest integer) of the vector p_h for a=0.05, b=0.04 and h^*=3

h				
0	10 000			
1	9500	500		
2	9045	930	25	
3	8630	1300	69	1
4	8250	1617	128	5
5	7903	1889	197	11
6	7583	2122	274	21
7	7289	2321	355	35
8	7017	2491	440	52
9	6766	2635	525	74

The number of such hidden states (the credit market conditions) is m and it is assumed that the transitions among these states are of Markov type. In detail, let z_0 be the probability distribution at time 0 on the m states, and

$$Q = \begin{pmatrix} q_{h;1;1} & q_{h;1;2} & \cdots & q_{h;1;m} \\ q_{h;2;1} & q_{h;2;2} & \cdots & q_{h;2;m} \\ \cdots & & & \\ q_{h;m;1} & q_{h;m;2} & \cdots & q_{h;m;m} \end{pmatrix}$$

be the probabilities transition matrix among the hidden states at time $h = 1, 2, \ldots$ of order $m \times m$.

For the subsequent probability distributions of the hidden states, the following equation holds for $h = 1, 2, \ldots$

$$z_h = z_{h-1} Q_h$$

The random transitions among the hidden states associated with the credit market conditions help to explain the 'repayment' process with action on the transition probabilities $p_{h,\,i,\,k}$: such probabilities depend on the credit market conditions and not only on h, i, k, and, in symbols, they become

$$p^{(s)}_{h;i;k}$$

$p^{(s)}_{h,i,k}$ represents the probability that at time $h = 1, 2, \ldots$, with the credit market conditions at state $s \in \{1, 2, \ldots, m\}$, a borrower, before time h has i instalments in arrears, has k instalments in arrears after payment in h.

The number of model parameters grows rapidly when both m and h^* increase, even if the model is simplified with the assumption of homogeneity in the Markov chains involved.

Here it is considered a homogeneous model, which is defined by means of $1 + m$ homogeneous chains: one for the hidden credit market conditions, m, for the transition probabilities matrices that describe the payment sequence.

In detail for the homogeneous model, let Q be the transition probabilities matrix of order $m \times m$ associated with the hidden states (the credit market conditions), with $m \times (m-1)$ free parameters.

$P^{(s)}$ with $s = 1, 2, \ldots, m$ be m matrices of order $(h^* + 1) \times (h^* + 1)$, which are the probabilities transition matrices among the states that describe the number of instalments in arrears for the payment sequence; there is one matrix for each of the m hidden states and the number of free parameters is $(m \times h^* \times (h^* + 1))/2$.

z_0 be the vector with m elements of the initial probabilities for the hidden states, which has $m-1$ free parameters.

Thus, for the homogeneous case, the number of free parameters is given by the sum of the previous ones:

$$(m \times (m-1)) + \left(\frac{m \times h^* \times (h^* + 1)}{2}\right) + (m-1)$$

Now we consider the case in which the number of hidden states is $m = 2$ and the default state is reached with $h^* = 3$: so 15 are the free parameters. This assumption makes the computation easier, but it does not lessen the conceptual aspects.

Let the states that describe the credit market conditions be B (for 'bad' state) and G (for 'good' state). The initial probability distribution is

$$z_0 = (d; 1-d) = (P(B); P(G))$$

The homogeneous transition probabilities matrix of order 2×2 between the two hidden states is

$$Q = \begin{pmatrix} 1-f & f \\ g & 1-g \end{pmatrix}$$

With respect to the vectors p_h and the transition probabilities matrices $P^{(s)}$, let

$$p_0 = (1; 0; 0; 0)$$

$$P_h^{(s)} = \left(p_{hij}^{(s)}\right) := P^{(s)}$$

$$= \begin{pmatrix} 1-a_{0;s} & a_{0;s} & 0 & 0 \\ b_{1;s} & 1-a_{1;s}-b_{1;s} & a_{1;s} & 0 \\ c_{2;s} & b_{2;s} & 1-a_{2;s}-b_{2;s} & a_{2;s} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

for $s \in \{B, G\}$ and for $h = 1, 2, \ldots$

Applying basic results of probability calculus, the probability distribution with respect to the number of instalments in arrears at the end of the first period is:

$$p_1 = \left(d \cdot (1 - a_{0;B}) + (1-d) \cdot (1 - a_{0;G})\right); \left(d \cdot a_{0;B} + (1-d) \cdot a_{0;G}; 0; 0\right)$$

or in matrix formulation

$$p_1 = d p_0 P_h^{(B)} + (1-d) p_0 P_h^{(G)}$$

At the end of the first period, the probability distribution for the hidden states is

$$z_1 = z_0 Q = (d; 1-d) \begin{pmatrix} 1-f & f \\ g & 1-g \end{pmatrix}$$
$$= (d(1-f) + (1-d)g; df + (1-d)(1-g)) := (z_{1;B}; z_{1;G})$$

Using z_1, it is possible to evaluate the probability distribution with respect to the number of instalments in arrears at the end of the second period and so on. Using classical properties on conditional probabilities, the k-th

element of the vector p_h for $k = 0, 1, 2, 3$, $p_{h,k}$, is recursively expressed as

$$p_{h;k} = \sum_{j=0}^{3} p_{h-1;j} \cdot (z_{h-1;B} \cdot p_{j;k}^{(BP)} + z_{h-1;G} \cdot p_{j;k}^{(GP)})$$

$$= \sum_{j=0}^{3} z_{h-1;B} \cdot (p_{h-1;j} \cdot p_{j;k}^{(BP)}) + \sum_{j=0}^{3} z_{h-1;G} \cdot (p_{h-1;j} \cdot p_{j;k}^{(GP)})$$

or in matrix formulation

$$p_h = z_{h-1;B} \; (p_{h-1} P_h^{(BP)}) + z_{h-1;G} \; (p_{h-1} P_h^{(GP)})$$

An example

To consider how the model actually functions, the following values of the 15 parameters are given:

$$d = 0.8; f = g = 0.05$$

$$a_{0;B} = a_{1;B} = a_{2;B} = 0.04; \quad b_{1;B} = b_{2;B} = 0.02; c_{2;B} = 0$$

$$a_{0;G} = a_{1;G} = a_{2;G} = 0.02; \quad b_{1;G} = b_{2;G} = 0.01; c_{2;G} = 0$$

According to these valuations, the credit market is in the 'bad' state with a high probability, $d = 0.8$. The hidden states persist with high probabilities, $1 - f = 1 - g = 0.95$. The probability of paying nothing is higher when compared with the probability of recovering instalments in arrears, and this assumption is amplified if the hidden state of the credit market is the 'bad' one. With $c_{2,B} = c_{2,G} = 0$, it is assumed that the borrower cannot recover two instalments in arrears at once.

In Table 3, some results for the model are listed for $h = 0, 1, \ldots, 24$.

With this model, it is possible to observe that the default probability at time 6 is equal to 0.08%, at time 12 is 0.61%, at time 18 is 1.79% and at time 24 is 3.55%; at time 24 the borrower will have a probability of 51.94% to be regular, 33.05% to have one instalment in arrears and 11.46% to have two instalments in arrears.

The simulation of the model

It is noteworthy that the model implies that the borrowers' payment sequences are not independent. In fact, the hidden state (the credit market conditions), which changes according to a Markov chain, affects all the borrowers' behaviours.

It will be of interest to simulate the payment sequence with credit portfolios of a given size in order to analyse, for instance, how the variance of the total number of defaulters, at a given time, might be explained in terms of two components: the systematic one, given by the dynamics of the credit market, and the residual one, given by the different borrowers' behaviours.

Table 3 Vectors z_h and p_h for $h = 0, 1, \ldots, 24$

h	z_{hB}	z_{hG}	p_{h0}	p_{h1}	p_{h2}	p_{h3}
0	0.8000	0.2000	1.0000	0.0000	0.0000	0.0000
1	0.7700	0.2300	0.9640	0.0360	0.0000	0.0000
2	0.7430	0.2570	0.9305	0.0682	0.0013	0.0000
3	0.7187	0.2813	0.8993	0.0971	0.0036	0.0000
4	0.6968	0.3032	0.8700	0.1231	0.0067	0.0002
5	0.6771	0.3229	0.8426	0.1464	0.0106	0.0004
6	0.6594	0.3406	0.8168	0.1675	0.0150	0.0008
7	0.6435	0.3565	0.7924	0.1865	0.0198	0.0013
8	0.6291	0.3709	0.7695	0.2037	0.0249	0.0019
9	0.6162	0.3838	0.7477	0.2192	0.0303	0.0027
10	0.6046	0.3954	0.7271	0.2333	0.0360	0.0036
11	0.5941	0.4059	0.7075	0.2459	0.0417	0.0049
12	0.5847	0.4153	0.6889	0.2574	0.0476	0.0061
13	0.5763	0.4237	0.6711	0.2678	0.0535	0.0076
14	0.5686	0.4314	0.6542	0.2771	0.0594	0.0093
15	0.5618	0.4382	0.6380	0.2855	0.0653	0.0112
16	0.5556	0.4444	0.6225	0.2931	0.0711	0.0133
17	0.5500	0.4500	0.6077	0.2999	0.2999	0.0155
18	0.5450	0.4550	0.5935	0.3060	0.0826	0.0179
19	0.5405	0.4595	0.5799	0.3114	0.0883	0.0204
20	0.5365	0.4635	0.5668	0.3162	0.0938	0.0232
21	0.5328	0.4672	0.5543	0.3205	0.0992	0.0260
22	0.5295	0.4705	0.5422	0.3243	0.1044	0.0291
23	0.5266	0.4734	0.5306	0.3276	0.1096	0.0323
24	0.5239	0.4761	0.5194	0.3305	0.1146	0.0355

As an example, if the portfolio is composed of 1000 customers, the number of defaulters at time 24 will be, on average, equal to $35.5 = 1000 \times 0.0355$. However, the standard deviation of the number of defaulters will not be equal to

$$\sqrt{1000 \times 0.0355 \times (1 - 0.0355)} \approx \sqrt{34.2398} \approx 5.85$$

which will be the result if the independence hypothesis could be assumed.

It is possible to evaluate the variance and standard deviation analytically, but it is more easy to estimate them by means of simulation of the model: the results of such simulations could fruitfully be used for further information as, for instance, quantiles, record values and so on.

In the following table, the results of 400 runs of a simulation of the model are shown: for each run, a set of 1000 customers are considered and the number of defaulters at $h = 24$ are registered. The numerical evidence of the simulation is:

4 1, 5 1, 6 1, 7 3, 8 2, 9 2, 10 1, 11 9, 12 5, 13 4, 14 13, 15 6, 16 10, 17 7, 18 8, 19 4,20 6, 21 9, 22 13, 23 6, 24 4, 25 5, 26 5, 27 8, 28 4, 29 3, 30 2, 31 11 , 32 11, 33 3, 34 11, 35 7, 36 6, 37 7, 38 5, 39 5, 40 8, 41 7, 42 6, 43 3, 44 8 , 45 9, 46 7, 47 10, 48 6, 49 10, 50 9, 51 7, 52 7, 53 6, 54 8, 55 9, 56 6, 57 7 , 58 9, 59 5, 60 4, 61 2, 62 8, 63 7, 64 4, 65 3, 66 2, 67 5, 68 3, 69 2, 70 1 , 71 2, 72 0, 73 0, 74 1, 75 1

In such couples, separated by commas, the first element represents the number of defaulters and the second one gives the number of such occurrences in the 400 runs of the simulation. For example, the first couple, 4 1, means that 4 defaulters in 1 run of the simulation have been observed; 5 1 means that 5 defaulters in 1 run of the simulation have been observed, and so on.

The sum of the second figures of such couples is 400, the number of runs of the simulation.

The sample mean of the number of defaulters obtained by simulation is equal to 35.8950, which is quite near to the expected one. The sample variance, 335.1117, is quite different from the above value of 34.2398, which is true with the independence assumption. The estimated standard deviation has a value of 18.3061, which is more than triple that of 5.8515, which is true with the independence assumption.

From the order statistics, it possible to observe that the estimated 99.5% quantile is 73, the 99.0% quantile is 71 and the 95% quantile is 64.

How to make inferences from empirical data

It is not difficult to suppose that such a model, even with homogeneous probabilities transition matrices, would be able to describe a wide class of borrowers' behaviours in terms of payment sequences using the $P^{(s)}$ matrices for $s = 1, 2, \ldots, m$, the Q matrix and the initial conditions z_0.

Using the simple model described in the previous paragraph, the problem is to estimate z_0, Q, $P^{(B)}$, $P^{(G)}$ starting from the empirical data, and obtain $\hat{z}_0$; $\hat{Q}$; $\hat{P}^{(B)}$; $\hat{P}^{(G)}$.

It is not a huge task collecting the company data set in order to determine the number of customers registered in the different classes, identified by the number of instalments in arrears, at any given time, and also to count the number of transitions among such classes.

On the contrary, it is not an easy task to obtain evidence of the dynamics of the conditions of the market in which the credit company operates. One could imagine using data provided by external vendors or analysing macroeconomic territorial indicators or looking at other informative sources.

In facing such a problem, we have examined a plain case in which the data are only the company ones. They are related to a sequence of n observations of the behaviours of N borrowers. Each borrower, step by step, is classified exactly into one of four classes without considering the frequencies of transition of the borrowers among the four classes.

The four classes are: regular, one instalment in arrears, two instalments in arrears and defaulters. In symbols, the data are only the $n + 1$ vectors:

$$(n_{h;0}; n_{h;1}; n_{h;2}; n_{h;3}); \text{ for } h = 0; 1; 2; \ldots; n$$

where $n_{h,k}$ represents the number of customers that at time h have k instalments in arrears with $k = 0, 1, 2$, and $n_{h,3}$ is the number of defaulters at time h.

For such data, the following conditions hold:

$$n_{0;0} = N; n_{1;0} + n_{1;1} = N; n_{2;0} + n_{2;1} + n_{2;2} = N$$
$$n_{h;0} + n_{h;1} + n_{h;2} + n_{h;3} = N \text{ for } h = 3; 4; \ldots; n$$

This is the information we need in order to evaluate the profitability and creditworthiness for the given group of borrowers because it enables a comparison of the expected value of the contractual payments with the actual one.

If the contractual instalments are $r = 1$ and if x is the internal periodic rate evaluated, for example, on a monthly basis, the discounted value of the payments, equal to the totally granted amount, is

$$A := N \cdot \frac{1 - (1 + x)^{-n}}{x}$$

while the discounted value of the actual payments are

$$B := \sum_{h=1}^{n} \left(N - n_{h;3} + n_{h-1;1} - n_{h;1} + n_{h-1;2} - n_{h;2} \right) \cdot (1 + x)^{-h}$$

The ratio B/A will represent, *ex post*, the creditworthiness index of the portfolio of contracts, as described in Quirini and Vannucci (2010).

The problem is how to infer the model parameters given the sequence of $(n_{h,0}, n_{h,1}, n_{h,2}, n_{h,3})$ for $h = 0, 1, 2, \ldots, n$. In the HMM theory (Rabiner, 1989; Zucchini and MacDonald, 2009), there are standard methodologies, primarily the Baum–Welch algorithm and the Viterbi algorithm, for the estimation of the model parameters. These inference methods are based on the fact that at each step a realization, coherent with the probability distribution associated to the hidden state, of a single random variable is known, but now the information is not of this type.

When estimation of a given d-dimensional time series generated by a stochastic process, having a large number of parameters, is attempted, the application of standard methods (maximum likelihood estimation, methods of moments) can become too cumbersome.

In our application of HMM, such procedures can be used if a single borrower is observed in time; but here we deal with the case in which, at each step and according to the hidden state, there are N random transition jumps in the four classes and such dynamics depend on the state each borrower has at the previous step. It appears as a quite difficult situation, not managed by the known actual software at disposal for HMM.

Therefore, we propose a method, naive and ingenious at the same time, based on simulation techniques, for point estimation. First, a random set of models, say M, has been chosen and let $(p_0, Q, P^{(B)}, P^{(G)})$ be the generic element of

this set. Then, each chosen model has been simulated ‘ times, finding a set of results similar to the observed data; that is

$$\left(n_{h;0}^{\delta i P}; n_{h;1}^{\delta i P}; n_{h;2}^{\delta i P}; n_{h;3}^{\delta i P}\right)$$

for $h = 0, 1, 2, \ldots, n$, where $i = 1; 2; \ldots; ‘$.

Among the randomly generated models the one for which the following quantity is minimized is selected:

$$\sum_{i=1}^{} \sum_{h=0}^{} \sum_{j=0}^{} \left(n_{h;j}^{\delta i P} - n_{h;j}\right)^2$$

This quantity is the sum of the ‘quadratic distances’ among the ‘ simulated paths and the actual one.

Here the ‘empirical’ data are the ones generated in one run of the 400 simulations of the model numerically described in this paragraph. Thus, we have considered as $(n_{h,0}, n_{h,1}, n_{h,2}, n_{h,3})$ for $h = 0, 1, 2, \ldots, 24$ the data of Table 4.

It is noteworthy that this has been a simulation with frequencies close to the corresponding probabilities (see Table 3); hence, such a data set could be seen as being a lucky result from the simulation.

Then, 500 models have been chosen randomly in M, and among such models the one that minimizes the above objective function is selected: that is, the sum of ‘ = 20 ‘quadratic distances’. The estimated values of the parameters (with three decimals) of the best model among the 500 sampled are:

$$d^{b} = 0{:}672;\ f^{b} = 0{:}030;\ g^{b} = 0{:}065;$$
$$a^{b}_{0;B} = 0{:}041;\ a^{b}_{1;B} = 0{:}030;$$
$$a^{b}_{2;B} = 0{:}084;\ b^{b}_{1;B} = 0{:}025;\ b_{2;B} = 0{:}084;$$
$$b^{b}_{2;B} = 0{:}168;\ a_{0;G} = 0{:}032;$$
$$a^{b}_{1;G} = 0{:}024;\ a^{b}_{2;G} = 0{:}074;\ b^{b}_{1;G} = 0{:}001;$$
$$b^{b}_{2;G} = 0{:}069;\ c^{b}_{2;G} = 0{:}047$$

The sample means, $\bar{n}_{h;0}; \bar{n}_{h;1}; \bar{n}_{h;2}; \bar{n}_{h;3}$, obtained by the 20 simulations of the selected best model are listed in Table 5.

The sample standard deviations, $\bar{s}(n_{h;0}); \bar{s}(n_{h;1}); \bar{s}(n_{h;2}); \bar{s}(n_{h;3})$, obtained by the 20 simulations of the selected best model are listed in Table 6.

Some considerations on the selected model look right.

First of all, it is possible to estimate a ‘poor’ model (in terms of explanatory power): if the estimated $p^{\delta B P}$ and $p^{\delta G P}$ are very near, then the empirical data are referred to periods in which the credit market conditions are very stable and the results of the payment sequences mainly depend only on the creditworthiness of the borrowers. When this situation happens, the estimates z^{b} and Q^{b}_{0} regardless of their values, assume a spurious character and they lose any meaning, because any other value for their elements cannot modify the distribution of the random

Table 4 $(n_{h,0}, n_{h,1}, n_{h,2}, n_{h,3})$ for $h=0, 1, 2, \ldots, 24$ for $N=1000$ borrowers

h	$n_{h,0}$	$n_{h,1}$	$n_{h,2}$	$n_{h,3}$
0	1000	0	0	0
1	961	39	0	0
2	929	69	2	0
3	901	94	5	0
4	873	116	11	0
5	849	135	16	0
6	811	169	20	0
7	786	186	27	1
8	758	207	33	2
9	739	221	38	2
10	717	237	44	2
11	685	257	53	5
12	660	274	61	5
13	643	286	65	6
14	625	296	69	10
15	616	297	75	12
16	602	305	79	14
17	596	307	81	16
18	587	312	82	19
19	565	327	89	19
20	549	335	91	25
21	532	343	93	32
22	528	340	99	33
23	519	341	104	36
24	511	343	109	37

Table 5 Sample means for the selected model

h	$n_{h,0}$	$n_{h,1}$	$n_{h,2}$	$n_{h,3}$
0	1000.0	0.0	0.0	0.0
1	962.0	38.9	0.0	0.0
2	926.0	72.6	1.4	0.0
3	892.2	104.4	3.3	0.1
4	859.2	135.0	5.5	0.3
5	831.0	159.8	8.4	0.8
6	803.6	184.4	10.4	1.6
7	774.9	209.6	13.2	2.3
8	750.5	231.4	14.8	3.3
9	730.5	247.8	17.7	4.0
10	713.1	262.8	18.9	5.2
11	696.9	276.3	20.0	6.8
12	677.1	292.4	21.7	8.8
13	660.3	304.7	24.0	11.0
14	646.6	315.0	25.9	12.5
15	632.0	325.1	28.3	14.6
16	616.7	336.9	29.1	17.3
17	602.3	349.0	29.3	19.4
18	589.9	358.3	29.7	22.1
19	576.4	368.3	31.0	24.3
20	567.2	373.8	32.2	26.8
21	558.8	377.4	34.0	29.8
22	550.4	383.4	33.5	32.7
23	539.9	388.9	35.2	36.0
24	531.3	393.7	35.8	39.3

Table 6 Sample standard deviations for the selected model

h	$s(n_{h,0})$	$s(n_{h,1})$	$s(n_{h,2})$	$s(n_{h,3})$
0	0.0	0.0	0.0	0.0
1	8.9	8.9	0.0	0.0
2	10.4	10.7	1.4	0.0
3	12.9	12.5	2.2	0.5
4	14.8	14.1	3.0	0.7
5	17.1	15.5	3.4	0.7
6	16.4	14.8	3.7	1.2
7	17.1	16.4	4.7	1.4
8	16.6	15.7	5.2	1.8
9	15.7	14.9	5.8	2.1
10	15.9	15.8	3.8	2.4
11	15.2	14.2	4.4	3.2
12	16.3	15.2	4.9	3.2
13	17.3	16.3	3.7	3.6
14	17.9	17.1	4.3	3.6
15	19.4	18.9	4.3	3.8
16	20.7	19.1	4.8	4.2
17	20.2	17.9	5.6	4.8
18	18.2	16.0	5.1	4.4
19	19.2	16.9	6.1	5.2
20	18.3	16.8	5.8	4.8
21	18.7	16.1	6.7	5.3
22	20.2	17.7	6.9	5.3
23	23.0	19.2	7.4	5.1
24	24.2	18.6	6.9	5.3

instalments in arrears related to successive time periods. In other words, if $P^{(B)}$ and $P^{(G)}$ have very similar values, we can avoid modelling by means of HMM and a more simple approach based on a single Markov chain may be sufficient.

There are two trivial methodological suggestions in order to improve the efficacy of the outlined procedure: first, increasing the number of selected models and, second, bounding the choice of values for some parameters. The second suggestion could be useful especially with respect to the matrices $P^{(B)}$ and $P^{(G)}$, the elements of which can vary in a limited range, near some known experiencied values.

It is natural to imagine that the algorithm could be iterated, to explore whether any better model could be found, 'near' the selected one at a certain stage, reducing the amplitude and modifying the position of the intervals at which to proceed in search of new better models. For instance, in the above illustrated application, the comparison of the values in the columns of Table 4 and Table 5 suggests that slight changes, essentially decreasing the values of the parameters 'b' and 'c' with respect to those estimated, could improve the fitting of the model.

A final remark. The outlined inference has been focussed on a cohort of borrowers, but it must consider a sequence of cohorts. A hard problem to solve is related to the coherence of the estimations for all the sequences of cohorts, which share, month after month, the same credit market conditions. To find a solution that can be operated over a long period of time, a possible approach could be, for instance, to average the parameter values for the different cohorts. This is a crucial point, because only if a certain persistence of the model is assumed over the time period will it be possible to read the future from the past and the present.

References

Cyert R, Davidson J and Thompson G (1962). Estimation of the allowance for doubtful accounts by Markov chains. *Management Science* 8(3): 3–19.

Giampieri G, Davis M and Crowder M (2005). Analysis of default data using hidden Markov models. *Quantitative Finance* 5(1): 27–34.

Kallberg JG and Saunders A (1983). Markov chain approaches to the analysis of payment behaviour of retail credit customers. *Financial Management* 12(2): 5–14.

Malik ML and Thomas L (2009). Modelling Credit Risk in Portfolios of Consumers Loans: Transition Matrix Model for Consumer Credit Ratings, Credit Scoring and Credit Control XI. Business School, University of Edinburgh.

Quirini L and Vannucci L (2010). A new index of creditworthiness for retail credit products. *Journal of Operation Research Society* 61(3): 455–461.

Quirini L and Vannucci L (2011). Monitoring Creditworthiness: Markov Chains and Replication Portfolios, Credit Scoring and Credit Control XII. Business School, University of Edinburgh.

Quirini L and Vannucci L (2012). Sul controllo dinamico dell'affidabilita` creditizia mediante catene di Markov. *Studi e Note di Economia, Anno XVII* 2012(1): 133–148.

Rabiner LR (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. In: *Proceedings of the IEEE* 77(2): 257–286.

Thomas LC, Edelman DB and Crook JN (2002). *Credit Scoring and its Applications*. SIAM Monographs on Mathematical Modeling and Computation, Philadelphia, USA.

Zucchini W and MacDonald IL (2009). *Hidden Markov Models for Time Series*. Boca Raton, FL, CRC Press.

Received January 2012;
accepted December 2012 after two revisions