Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products

Leonard J. Paas,

Vrije Universiteit, Amsterdam, The Netherlands

Jeroen K. Vermunt

Tilburg University, The Netherlands

and Tammo H. A. Bijmolt

University of Groningen, The Netherlands

Summary. The paper demonstrates application of the latent Markov model for assessing developments by individuals through stages of a process. This approach is applied by using a database on ownership of 12 financial products and various demographic variables. The latent Markov model derives latent classes, representing household product portfolios, and shows the relationship between class membership and household demographics. The analysis provides insight into switching between the latent classes, reflecting developments of individual household product portfolios, and the effects of demographics on such switches. Based on this, we formulate equations to predict future acquisitions of financial products. The model accurately predicts which product a specific household unit acquires next, for most of the products.

Keywords: Acquisition pattern analysis; Financial products; Forecasting; Hidden Markov model; Hierarchies of characteristics; Latent Markov model

1. Introduction

In economics, psychology and other social sciences, researchers investigate the order in which individuals or household units acquire the characteristics defining the stages of a process. Examples are the order in which some individuals use different types of illicit drugs (Collins and Wugalter, 1992; Graham *et al.*, 1991), the order in which children acquire various types of intellectual skills and abilities (Kingma, 1984; Verweij *et al.*, 1999) and the order in which household units acquire durable or financial products (Kamakura *et al.*, 1991; Paroush, 1965). As an example of the last, extant research results show households generally acquire savings accounts before investment trusts and after that they acquire shares (Paas, 1998). This paper investigates the order in which households acquire such financial products. The methodology discussed is relevant for other applications.

Statistical models analysing temporal patterns of consumer choice behaviour generally concern repeat purchases of consumable products, such as detergents, paper towels and ketchup

(e.g. Erdem and Sun (2001) and Uncles *et al.* (1995)). Models on more durable products, such as savings accounts, televisions and air-conditioning units, receive far less attention. Nevertheless, a technique which is known as acquisition pattern analysis (e.g. Kamakura *et al.* (1991), Paas and Molenaar (2005) and Soutar and Cornish-Ward (1997)) provides insight into the order in which households acquire such products. Behavioural scientists use this analysis to evaluate priorities that households have for products and marketing managers for purposes such as market segmentation and estimating the potential of an innovation by predicting its position in existing acquisition patterns (Gatignon and Robertson, 1985; Kamakura *et al.*, 1991). Recently, acquisition pattern analysis has received attention for predicting which product a household is most likely to acquire next (Li *et al.*, 2005; Paas and Molenaar, 2005).

In our application of acquisition pattern analysis, households have various financial objectives that cannot be fulfilled at once. Acquisitions of financial products usually imply major investments or long-term contractual obligations. Finite resources will lead to a priority structure of household financial objectives: products for more basic objectives are generally acquired before products providing utility for more advanced objectives. Furthermore, Kamakura *et al.* (1991) and Soutar and Cornish-Ward (1997) suggested the phases, through which households pass during their life cycle, will lead to a similar order in which financial priorities become salient and consequently a common order for acquiring financial products. Note that this paper studies household units as these are the principal decision-making unit in the financial product market (Guiso *et al.*, 2002; Wärneryd, 1999).

Although acquisition pattern analysis is based on plausible theoretical assumptions, relevant empirical studies have methodological shortcomings. Extant studies typically analyse cross-sectional data and derive acquisition patterns from cross-sectional differences between households (e.g. Kamakura *et al.* (1991), Paas (1998) and Soutar and Cornish-Ward (1997)). For instance, if households acquire three products in the order first A, then B and last C, then the following cross-sectional ownership patterns should occur only sparsely:

(a) ownership of product C, without ownership of products A and B and
(b) ownership of B without A.

Paas and Molenaar (2005) theoretically and empirically demonstrated that acquisition patterns, deduced from cross-sectional data, can predict future product acquisitions by individuals. Such predictions have various practical purposes, such as selecting the product to be offered to a particular household unit. However, cross-sectional data do not provide insight into divergent orders of acquisition. Divergence in this context implies that different segments of households in a population follow different orders for acquiring products. For example, one segment acquires savings accounts before investment trusts and last shares, whereas another segment starts with the acquisition of shares. This is empirically relevant, as Bijmolt *et al.* (2004) and Paas *et al.* (2005) reported the occurrence of such divergence. Besides this, it is not only interesting to know which product a household is likely to acquire next, but also the timing of such acquisitions can be salient (Kamakura *et al.*, 1991). Cross-sectional data provide little insight into such timing. Some relevant cross-sectional methods are available, such as the life calendar history method that was introduced by Freedman *et al.* (1988). Here respondents are asked to recall the order in which events or activities, such as the acquisition of various products, took place. Various approaches have been developed to offer respondents some support in this burdensome recall process. However, such methods still impose a heavy cognitive task on respondents and the results tend to be strongly biased by memory effects (Means *et al.*, 1991).

This paper applies the latent Markov model (LMM)—which is also known as the hidden Markov model and the latent transition model—for analysing acquisition patterns in the financial

product market, based on longitudinal panel data. In this application of the LMM, households may be conceived of being divided into a number of unobservable discrete categories for which product portfolios differ. To illustrate this, assume a data set containing three panel waves, collected in 1996, 1998 and 2000, with information on household ownership of products A, B and C. The usual axiom of the LMM is that latent class membership explains the association between observed variables. This concerns product ownership in our application. Moreover, each latent class represents a prototypical product portfolio. For example, if most households owning product C also own products A and B, and most households with B also own product A, the model would consist of four latent classes:

(a) a class with households owning none of the products;
(b) a class with households owning only product A;
(c) a class with households owning products A and B;
(d) a class with households owning all three products.

Besides this, the LMM describes the developments of financial product portfolios by individual households over time, using switching probabilities between latent classes over consecutive measurement occasions. Consider our previously mentioned hypothetical LMM. Here the analysis would define the probability for households in the segment owning only product A in 1996 to be in the segment owning none of the products in 1998, or, alternatively, the probability for such a household to be in the segment owning

(a) products A and B or
(b) products A, B and C.

Other changes in product portfolios are also modelled and represented by probabilities of switching from one latent class to another over consecutive measurement occasions. Further discussion and technical details on this application of the LMM are provided later in the paper.

The contribution of the paper is twofold. First, we introduce an approach for applying the LMM to predict future behaviour. For this, we modify the forward–backward algorithm (Baum *et al.*, 1970; Fearnhead and Meligkotsidou, 2004), which is an alternative to the more commonly applied expectation–maximization (EM) algorithm of Dempster *et al.* (1977). The paper concentrates on predicting household acquisitions of financial products, but the approach is applicable to other purposes, such as predicting the intellectual skill that a child will acquire next. Second, the paper contributes to theory on financial behaviour, through the presentation of new empirical findings on household financial product portfolios and developments therein.

The next section presents the LMM and Section 3 explains the application of the LMM for predicting product acquisitions of individual households. Section 4 reports an LMM analysis of a database with ownership information for 12 financial products, by 7676 households in four bi-yearly waves. Section 5 presents the accuracy with which the next acquisition is predicted. The paper concludes with implications in Section 6.

2. The latent Markov model

2.1. Model specification
Below we present the LMM with concomitant variables (Van de Pol and Langeheine, 1990; Vermunt *et al.*, 1999; Bartolucci *et al.*, 2007) and in particular how it can be applied to analyse household acquisition patterns of financial products. However, first some notation is required: $i = 1, \ldots, I$ is the index of households; $j = 1, \ldots, J$ is the index of financial products; $k = 1, \ldots, K$

is the index of covariates; $s = 1, \ldots, S$ is the index of latent classes representing household product portfolios (latent classes in an LMM are sometimes referred to as latent states, as these represent the phases of some process); $t = 0, \ldots, T$ is the index of a sequence of $T + 1$ measurement occasions; $X_{it} = s_t$ denotes that household i is a member of latent class s_t at measurement occasion t (below the symbol s or r is sometimes used instead of s_t to refer to a specific latent class); Y_{ijt} denotes whether household i owns product j at measurement occasion t ($Y_{ijt} = 1$ if subject i has j at t; otherwise, $Y_{ijt} = 0$); $\mathbf{Y}_{it}$ denotes the $1 \times J$ vector of binary variables indicating which of the J products household i owns at measurement occasion t; $\mathbf{Y}_i$ denotes the full set of products at all $T + 1$ occasions (it is a stacked row vector defined as $\{Y_{i0}, Y_{i1}, \ldots, Y_{iT}\}$); $\mathbf{Z}_{it}$ denotes the $1 \times K$ vector of values that household i has on the K covariates at t; $\mathbf{Z}_i$ denotes the full set of covariate values at all $T + 1$ measurement occasions, defined as $\{\mathbf{Z}_{i0}, \mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{iT}\}$.

The LMM specifies $P(\mathbf{Y}_i|\mathbf{Z}_i)$, i.e. the probability of having a particular combination of products at the $T + 1$ measurement occasions given a household's time-specific covariate values. Two components define this probability: a structural component models individual changes in latent states across time points (developments in household product portfolios) and a measurement component connects the latent states (the types of product portfolios) at a particular time point to the observed responses (the observed product ownerships).

The structural component has a first-order Markov model structure, i.e., when controlling for covariate values at t ($\mathbf{Z}_{it}$), X_{it} is affected only by class membership at the previous measurement occasion, $X_{i,t-1}$, but not by latent class membership at earlier occasions. The structural component involves two types of model probabilities:

(a) the initial latent state probability $P(X_{i0} = s_0|\mathbf{Z}_{i0})$ denotes the probability that household i belongs to latent class s_0 at the initial measurement occasion given its covariate values at this occasion, and

(b) the latent transition probability $P(X_{it} = s_t|X_{i,t-1} = s_{t-1}, \mathbf{Z}_{it})$ denotes the probability that a household in latent class s_{t-1} at occasion $t-1$ switches to latent state s_t at occasion t given its covariate values at t.

The measurement part of the model, which connects the latent class membership at time point t to the observed responses at t, takes the form of a standard latent class model for dichotomous response variables (Lazarsfeld and Henry, 1968; Goodman, 1974). This component can be interpreted as a latent class structure, defining a segmentation based on the observed product ownerships at each measurement occasion. The probability of having a specific combination of J products at measurement occasion t, given that household i belongs to latent class s_t, is assumed to take the form

$$P(\mathbf{Y}_{it}|X_{it} = s_t) = \prod_{j=1}^{J} (\pi_{js}^t)^{Y_{ijt}} (1 - \pi_{js}^t)^{1-Y_{ijt}} \tag{1}$$

where π_{js}^t denotes the probability of having product j conditional on membership of latent class s. The multiplication over j on the right-hand side of equation (1) indicates that the J product ownerships are treated as independent Bernoulli trials conditional on a household's latent class membership at occasion t. This is the local independence assumption, which was outlined previously as crucial to this type of model. In our application it explains the association between ownership of the products. The latent class structure is assumed to capture the observed relationships in the product ownerships. Note that the model probabilities π_{js}^t, defining the measurement part of the model, contain an index t. This index indicates that π_{js}^ts may differ across measurement occasions. Moreover, for each measurement occasion t the model

incorporates one unobserved latent variable, which can take S values called latent classes. Thus, in total the model incorporates $T + 1$ latent variables.

Combining the first-order Markov model, which connects the latent states at the various time points, and the latent class structure, connecting the latent states to the observed product ownerships, yields the LMM

$$P(\mathbf{Y}_i|\mathbf{Z}_i) = \sum_{s_0=1}^{S} \sum_{s_1=1}^{S} \ldots \sum_{s_T=1}^{S} P(X_{i0}=s_0|\mathbf{Z}_{i0}) \prod_{t=1}^{T} P(X_{it}=s_t|X_{i,t-1}=s_{t-1}, \mathbf{Z}_{it}) \prod_{t=0}^{T} P(\mathbf{Y}_{it}|X_{it}=s_t)$$
(2)

where the measurement part $P(\mathbf{Y}_{it}|X_{it}=s_t)$ has the form that is provided in equation (1). Logistic models may parameterize and restrict the model probabilities as will be considered below. The summations over the classes of each of the $T + 1$ discrete latent variables implies that we marginalize over all unobserved $T + 1$ variables to obtain the expression for $P(\mathbf{Y}_i|\mathbf{Z}_i)$, the probability of the observed product ownerships given covariate values. Furthermore, the following assumptions underlie the LMM that is defined in equations (1) and (2):

(a) households belong to only one latent class at each specific measurement occasion, but this class is unknown;
(b) a first-order Markov chain defines the latent transition structure;
(c) covariates may affect latent class membership at occasion t, X_{it}, but conditional on this latent class membership not the observed product ownerships, Y_{ijt};
(d) product ownerships at occasion t are solely affected by the latent class membership at t, not by latent class membership at other time points or by other observed product ownerships;
(e) as indicated by equation (1), ownerships of the J products at occasion t, Y_{ijt}, are mutually independent given latent class membership at t.

These simplifying assumptions are, in fact, a combination of the assumptions of a latent class model and the ones of a first-order Markov transition model (Van de Pol and Langeheine, 1990; Vermunt *et al.*, 1999; Wedel and Kamakura, 2000). Though these assumptions cannot be fully relaxed because otherwise the model would no longer be identifiable, it is possible to check and relax assumptions (b)–(e) partially. For example, we could define a second-order instead of a first-order Markov process, allow covariates to have direct effects on certain observed ownerships, allow the latent state at time point t to affect the ownership of certain products at time point $t + 1$ or allow certain product ownerships at time point t to be associated within classes.

Logistic models can be used to parameterize the probabilities for the structural components of the latent Markov model, $P(X_{i0}=s|\mathbf{Z}_{i1})$ and $P(X_{it}=s|X_{i,t-1}=r, \mathbf{Z}_{it})$, which depend on covariate values. The linear model for the logarithm of the ratio of the probability of being in latent class s relative to being in the reference class S at $t = 0$ takes the form

$$\log\left\{\frac{P(X_{i0}=s|\mathbf{Z}_{i0})}{P(X_{i0}=S|\mathbf{Z}_{i0})}\right\} = \gamma_{s0}^{0} + \sum_{k=1}^{K} \gamma_{sk}^{0} Z_{i0k}$$
(3)

for $1 \leqslant s \leqslant S - 1$, where γ_{s0}^{0} denotes an intercept and γ_{sk}^{0}, $1 \leqslant k \leqslant K$, the slope for the kth covariate. A similarly logistic model for transition probabilities is

$$\log\left\{\frac{P(X_{it}=s|X_{i,t-1}=r, \mathbf{Z}_{it})}{P(X_{it}=S|X_{i,t-1}=r, \mathbf{Z}_{it})}\right\} = \gamma_{sr0}^{t} + \sum_{k=1}^{K} \gamma_{sk}^{t} Z_{itk}$$
(4)

for $1 \leqslant s \leqslant S - 1$, $1 \leqslant r \leqslant S$ and $1 \leqslant t \leqslant T$. In equation (4) there is a separate set of intercepts, γ_{sr0}^{t}, for each origin state. In contrast, the covariate effects γ_{sk}^{t} are assumed to be constant across

origin states. This is, however, an assumption that can be relaxed and tested. A specification with varying covariate effects across origin states is equivalent to including interactions between covariates and origin state. A likelihood ratio test comparing our specification and a model with such interaction terms can be used to assess whether the simplifying assumption of constant covariate effects across origin states holds for a particular data set. The main advantage of our specification is that it is much more parsimonious, especially when the number of latent classes S is large, which applies to the models that are discussed in the empirical sections of our paper.

For each $1 \leqslant j \leqslant J$ and $0 \leqslant t \leqslant T$, the S probabilities that are relevant for product ownership may also be reparameterized through log-odds as

$$\log\left(\frac{\pi_{js}^t}{1 - \pi_{js}^t}\right) = \beta_{j0}^t + \beta_{js}^t \qquad \text{for } 1 \leqslant s \leqslant S. \tag{5}$$

Since this involves $S+1$ parameters, for identification we may, for instance, assume that $\beta_{jS}^t = 0$ so that the log-odds for latent class S are β_{j0}^t. The log-odds for the other $S-1$ classes are then expressed as differences from the log-odds for this base class. This representation also becomes useful if the parameters are then used in logistic models, e.g. to specify restrictions on the measurement models. In Rasch-type models for example (Vermunt, 2001), the β_{js}^t are equal for all items j. The current application as will be developed in Section 3.1 does not use such constraints.

2.2. *Parameter estimation*

Maximum likelihood provides estimates of the model parameters of an LMM. An EM algorithm is used for this purpose (Van de Pol and Langeheine, 1990). We use an EM algorithm called the forward–backward algorithm (Baum *et al.*, 1970; Fearnhead and Meligkotsidou, 2004), implemented in Latent GOLD version 5.0 (Vermunt and Magidson, 2007). Below follows a discussion on the forward–backward algorithm, starting with the main differences from the more conventional application of EM.

A major limitation of the standard EM estimation of Dempster *et al.* (1977) for the current LMM parameters is that the time and storage that are needed for computation increase exponentially with the number of time points (Vermunt *et al.*, 1999). The E-step of this iterative algorithm involves computation of the joint posterior latent distribution of all latent variables, which contains S^{T+1} entries. For example, with $S = 10$ and $T + 1 = 4$, this already leads to a distribution with 10000 entries. After retrieving the entries in the joint posterior distribution one collapses over the other latent variables to obtain the required univariate and bivariate marginal posterior probabilities for adjacent time points. Instead, the alternative forward–backward algorithm obtains the relevant marginal posterior probabilities directly by using a set of recursive formulae, yielding a method for which the size of the problem increases only linearly with the number of time points (McDonald and Zucchini, 1997). The original forward–backward algorithm of Baum *et al.* (1970) assumes at each measurement occasion a single response variable only. Our application has multiple responses for $J = 12$ financial products. Thus, we adjust and extend the algorithm to handle this situation. Utilizing the forward–backward algorithm instead of the conventional application of EM is one of the major differences between the LMM analyses that are employed in the current paper and those that were employed by Collins and Wugalter (1992) and Graham *et al.* (1991).

EM, implemented through the forward–backward algorithm or through other algorithms, is a general iterative procedure for maximum likelihood estimation in the presence of latent variables or other types of missing data (Dempster *et al.*, 1977). It switches between an E-step and an M-step till convergence. The E-step computes the expected value of the complete-

data log-likelihood or, more intuitively, estimates the missing data (here the unobserved class memberships). For this the algorithm employs the expected value given the current parameter values and the observed data. The M-step uses standard estimation methods to update the model parameters, such that the expected complete-data log-likelihood is maximized. Below the M-step involves using the filled-in expected values as if they were observed data in logistic regression analysis. The E- and M-step cycle till a certain convergence criterion has been reached.

For an LMM described in equations (1) and (2), the contribution of case i to the expected complete-data log-likelihood $E\{\log(L_i)\}$ is as follows:

$$
\begin{aligned}
E\{\log(L_i)\} = {}& \sum_{s_0=1}^{S} P(X_{i0}=s_0|\mathbf{Y}_i, \mathbf{Z}_i) \log\{P(X_{i0}=s_0|\mathbf{Z}_{i0})\} \\
& + \sum_{t=1}^{T} \sum_{s_{t-1}=1}^{S} \sum_{s_t=1}^{S} P(X_{i,t-1}=s_{t-1}, X_{i,t}=s_t|\mathbf{Y}_i, \mathbf{Z}_i) \log\{P(X_{it}=s_t|X_{i,t-1}=s_{t-1}, \mathbf{Z}_{it})\} \\
& + \sum_{t=0}^{T} \sum_{s_t=1}^{S} P(X_{it}=s_t|\mathbf{Y}_i, \mathbf{Z}_i) \log\{P(\mathbf{Y}_{it}|X_{it}=s)\}.
\end{aligned}
\tag{6}
$$

The M-step maximizes expression (6) summed over household observations to find updated model parameter estimates. These can then also be used to update $P(X_{i0}=s_0|\mathbf{Z}_{i0})$, $P(X_{it}=s_t|X_{i,t-1}=s_{t-1}, \mathbf{Z}_{it})$, and $P(\mathbf{Y}_{it}|X_{it}=s_t)$ through which, as is shown below, the current parameter model estimates enter the next E-step. This E-step requires updating the univariate and bivariate posterior latent class membership probabilities $P(X_{it}=s|\mathbf{Y}_i, \mathbf{Z}_i)$ and $P(X_{i,t-1}=r, X_{i,t}=s|\mathbf{Y}_i, \mathbf{Z}_i)$ for each household. All these probabilities condition on the data and also implicitly on the current model estimates. The forward–backward algorithm obtains them by a recursive scheme and once found they are used in expression (6) for the next M-step.

The probabilities $\alpha_{it}(s) = P(X_{it}=s, \mathbf{Y}_{i,t-}|\mathbf{Z}_{i,t-})$ and $\beta_{it}(s) = P(\mathbf{Y}_{i,t+}|X_{it}=s, \mathbf{Z}_{i,t+})$ are the two key components of the forward–backward algorithm. Index $t-$ refers to the information for time point t and all earlier time points, and $t+$ for all time points after t. Thus, the forward probability $\alpha_{it}(s)$ refers to having the observed set of responses, i.e. observed product ownerships, up to time point t and being in latent class s at t, conditional on covariate values and model parameters. The backward probability $\beta_{it}(s_t)$ is the probability of having the observed set of responses after time point t, conditional on being in latent class s at t, covariate values and model parameters. As shown in Appendix A the connection between these two components and the relevant posterior probabilities is as follows:

$$
P(X_{it}=s|\mathbf{Y}_i, \mathbf{Z}_i) = \frac{\alpha_{it}(s)\,\beta_{it}(s)}{P(\mathbf{Y}_i|\mathbf{Z}_i)},
\tag{7}
$$

$$
P(X_{i,t-1}=r, X_{it}=s|\mathbf{Y}_i, \mathbf{Z}_i) = \frac{\alpha_{it-1}(r)\,P(X_{it}=s|X_{i,t-1}=r, \mathbf{Z}_{it})\,P(\mathbf{Y}_{it}|X_{it}=s)\,\beta_{it}(s)}{P(\mathbf{Y}_i|\mathbf{Z}_i)}.
\tag{8}
$$

These two equations show that the relevant posteriors $P(X_{it}=s|\mathbf{Y}_i, \mathbf{Z}_i)$ and $P(X_{i,t-1}=r, X_{it}=s|\mathbf{Y}_i, \mathbf{Z}_i)$ can be obtained with $\alpha_{it}(s)$, $\beta_{it}(s)$ and the model probabilities $P(X_{it}=s|X_{i,t-1}=r, \mathbf{Z}_{it})$ and $P(\mathbf{Y}_{it}|X_{it}=s)$ from the previous M-step (Baum *et al.*, 1970). The denominator $P(\mathbf{Y}_i|\mathbf{Z}_i)$ in equations (7) and (8), as shown in Appendix A, is obtained by $\Sigma_{s=1}^{S} \alpha_{it}(s)\,\beta_{it}(s)$ for any convenient arbitrary choice of t; this includes $t=T$ when it becomes simply $\Sigma_{s=1}^{S} \alpha_{iT}(s)$.

The recursive scheme for obtaining $\alpha_{it}(s)$ and $\beta_{it}(s)$, which can be established by using basic rules of probability and the properties of the LMM, proceeds as follows:

$$
\alpha_{i0}(s) = P(X_{i0}=s|\mathbf{Z}_{i0})\,P(\mathbf{Y}_{i0}|X_{i0}=s),
\tag{9}
$$

$$\alpha_{it}(s) = \sum_{r=1}^{S} \alpha_{it-1}(r) \, P(X_{it} = s | X_{i,t-1} = r, \mathbf{Z}_{it}) \, P(\mathbf{Y}_{it} | X_{it} = s), \qquad \text{for } 1 \leqslant t \leqslant T, \qquad (10)$$

and

$$\beta_{iT}(s) = 1, \tag{11}$$

$$\beta_{it}(s) = \sum_{r=1}^{S} \beta_{it+1}(r) \, P(X_{i,t+1} = r | X_{it} = s, \mathbf{Z}_{i,t+1}) \, P(\mathbf{Y}_{i,t+1} | X_{i,t+1} = r), \qquad \text{for } T-1 \geqslant t \geqslant 0. \tag{12}$$

In the computation of $\alpha_{it}(s)$ we start with the initial time point ($t = 0$), via equation (9), and then proceed until we reach time point T. In equation (10) information on latent class membership and responses, i.e. product ownerships, at the time point t is added by the multiplication. Information on class membership at $t-1$ is removed by the summation. The computation of the $\beta_{it}(s)$ starts with the last time point T, via equation (11), and continues backwards until the initial time point ($t = 0$) has been reached. In equation (12) information on latent class membership and the product ownerships at $t+1$ is added by the multiplication. Moreover, the conditioning on the latent class membership at time point $t+1$ is replaced by the conditioning on the class membership at time point t by using the summation.

3.　Predicting product acquisitions by using the latent Markov model

3.1.　Additional model assumptions
To be useful for prediction the model that was previously discussed will need some additional restricting assumptions. Section 3.1 discusses these assumptions and Section 3.2 the prediction of product acquisitions on the basis of the assumptions. The assumptions are based on the LMM framework that was introduced by Brangula-Vlagsma *et al.* (2002). They suggested two types of change: manifest change and latent change. The former refers to dynamics in the measurement part of the model, i.e. ownership probabilities for products are not constant per latent class over time. Latent change refers to switching between latent classes by individual households, i.e. the structural part of the model. We develop a model with a time constant measurement model, which allows switching between latent classes. In other words, the model allows latent change but not manifest change. Equation (13) imposes this restriction:

$$\pi_{js}^{t} = \pi_{js} \qquad \text{for } 1 \leqslant s \leqslant S, 1 \leqslant j \leqslant J \text{ and } 0 \leqslant t \leqslant T, \tag{13}$$

Equation (13) specifies that product penetration levels should be consistent in latent classes over measurement occasions. Assume a data set containing three panel waves, collected in 1996, 1998 and 2000. Consider that in 1996, for example, 45% of the households in latent class 1 own a credit card. An absence of manifest change implies that the penetration level of the credit card in latent class 1 is also 45% in the 1998 and 2000 panel waves. Acquisitions of financial products, such as credit cards, only result in changes of latent class membership by individual households. Models without manifest change are most suitable for marketing purposes, such as segmentation, because changing segment structures would lead to a reformulation of segment-specific marketing strategies (Wedel and Kamakura, 2000). Note that imposing equation (13) implies that the time indices can be dropped from equations (1) and (2).

The other assumption, which is based on Brangula-Vlagsma *et al.* (2002), restricts the latent change transition probabilities to be constant across time. In terms of the parameters in equation (4) it specifies that

$$\gamma_{sr0}^{t} = \gamma_{sr0} \qquad \text{for } 1 \leqslant s \leqslant S-1, \, 1 \leqslant r \leqslant S \text{ and } 1 \leqslant t \leqslant T,$$
$$\gamma_{sk}^{t} = \gamma_{sk} \qquad \text{for } 1 \leqslant s \leqslant S-1, \, 1 \leqslant k \leqslant K \text{ and } 1 \leqslant t \leqslant T. \tag{14}$$

Consider that in the previously mentioned hypothetical example the overall probability of switching from latent class 1 to latent class 2 between measurement occasions 1996 and 1998 equals 10%. The restriction that is defined in equation (14) implies that this probability should also be 10% between measurement occasions 1998 and 2000. Furthermore, covariate effects are time constant. If, between 1996 and 1998, 15% of the households with a head aged between 45 and 60 years switches from latent class 1 to 2, then this 15% probability should also apply to switches between 1998 and 2000.

The current paper makes no assumptions about the form of the relationship between latent class membership and the observed product ownerships. For example, the ordinal latent class model (Croon, 1991), the Rasch model (Vermunt, 2001) or the frameworks that were presented by Proctor (1970) and Feick (1987) can impose across class monotonicity on the response probabilities. Such restrictions can test whether households tend to acquire products in a common order (Bijmolt *et al.*, 2004; Paas and Molenaar, 2005). This is the second major difference between the LMM that is introduced in the current paper and the models that were applied by Collins and Wugalter (1992) and Graham *et al.* (1991), besides the difference in the EM algorithm employed, as discussed earlier in this paper. These previously applied models are based on such cumulative restrictions. Extant studies on acquisition patterns of financial products (Bijmolt *et al.*, 2004; Paas *et al.*, 2005) suggest, however, that a common order of acquisition is questionable for the financial product market. For other applications of latent Markov modelling, relaxing the cumulative restriction leads to a more general model that can provide more complete information for the stages of a process. To illustrate this, consider that the measurement component is a table with J rows, one for each product, and S columns, one for each latent class. Cell $\{j, s\}$ presents the probability of owning product j given membership of segment s. In the more relaxed model a common order leads to a distinct type of measurement component, i.e. high proportions, e.g. at least 0.70, in the lower right-hand part of the table representing the measurement component and low proportions, e.g. below 0.30, in the upper left-hand part (Collins and Wugalter, 1992; Graham *et al.*, 1991). If households tend to follow different orders, this will be reflected by other patterns in the measurement part of the model, as will be reported in Section 4.3.

3.2. Prediction equations

For predicting future product acquisitions by individual households, we assess the probability that households own financial product j at time point $T + 1$ given all observed information that is available at T, which is denoted by $P(Y_{ij,T+1} = 1 | \mathbf{Y}_{i,T-}, \mathbf{Z}_{i,T-})$. The information at T concerns the products that are owned by and the covariate values for the households at T and all earlier occasions, which is referred to with the symbol $T-$. For the prediction of household product acquisitions, the LMM first allocates probabilities of individual households being members of latent classes at the last measurement occasion T. For predicting household latent class membership at $T + 1$ the procedure uses latent class membership at T in combination with

 (a) household covariate values and
 (b) the covariate-specific transition probabilities that are defined in the model.

The household's predicted latent class membership at $T + 1$ is used to calculate the probability of owning products at this time point. For example, if a household is likely to be found in a latent class at $T + 1$, where the probability is high for owning a credit card, this product is likely to be owned at $T + 1$. Such a finding is particularly interesting for households that do not own this product at T. They will probably acquire a credit card between T and $T + 1$.

Formulated formally, the following three steps lead to future ownership probabilities:

$$P(X_{iT} = s | \mathbf{Y}_{iT-}, \mathbf{Z}_{iT-}) = \alpha_{iT}(s) \Big/ \sum_{r=1}^{S} \alpha_{iT}(r), \tag{15}$$

$$P(X_{i,T+1} = r | \mathbf{Y}_{i,T-}, \mathbf{Z}_{i,T-}, \mathbf{Z}_{i,T+1}) = \sum_{s=1}^{S} P(X_{i,T} = s | \mathbf{Y}_{i,T-}, \mathbf{Z}_{i,T-}) P(X_{i,T+1} = r | X_{iT} = s, \mathbf{Z}_{i,T+1}), \tag{16}$$

$$P(Y_{ij,T+1} = 1 | \mathbf{Y}_{i,T-}, \mathbf{Z}_{i,T-}, \mathbf{Z}_{i,T+1}) = \sum_{s=1}^{S} P(X_{i,T+1} = s | \mathbf{Y}_{i,T-}, \mathbf{Z}_{i,T-}, \mathbf{Z}_{i,T+1})$$
$$\times P(Y_{ij,T+1} = 1 | X_{i,T+1} = s). \tag{17}$$

The first step, which is represented as equation (15), involves computing $P(X_{i,T} = s | \mathbf{Y}_{i,T-}, \mathbf{Z}_{i,T-})$, the posterior latent class membership probabilities for household i at occasion T given all observed information up to T. This involves using equation (7) for $t = T$. The second step, equation (16), calculates prior latent class membership probabilities for time $T + 1$ given the observed information up to T. This involves multiplying the posteriors from the first step by the transition probabilities $P(X_{i,T+1} = r | X_{iT} = s, \mathbf{Z}_{i,T+1})$ and marginalizing over class membership at time point T. Step 3, equation (17), provides the relevant predicted probabilities for owning the products at time point $T + 1$. Predictions are derived from the prior class membership probabilities $P(X_{i,T+1} = s | \mathbf{Y}_{i,T-}, \mathbf{Z}_{i,T-}, \mathbf{Z}_{i,T+1})$ and the latent-class-specific ownership probabilities $P(Y_{ij,T+1} = 1 | X_{i,T+1} = s)$.

Equations (16) and (17) require estimates for unknown future transition probabilities $P(X_{i,T+1} = r | X_{iT} = s, \mathbf{Z}_{i,T+1})$ and future latent-class-specific product ownership probabilities $P(Y_{ij,T+1} = 1 | X_{i,T+1} = s)$ respectively. This implies that the assumptions in equations (13) and (14) must be fulfilled. In situations in which it is not possible to predict future covariate values or in which parameters vary over time, more *ad hoc* assumptions are required for applying equations (15)–(17). For example, we could fix time-varying covariates and/or model parameters at their value at the last measurement occasion. This situation does not occur in our empirical data. Section 4.2 compares models by using the standard Bayesian information criterion BIC. Higher values when conditions (13) and (14) are relaxed imply that these assumptions are consistent with the data analysed.

4. Empirical application of the latent Markov model

4.1. Data
The Dutch division of the international market research company Growth from Knowledge provided the data that are analysed. This company conducts a bi-yearly empirical study on financial product ownership in the Netherlands. The data concern household ownership of 12 financial products in 1996, 1998, 2000 and 2002. Interviews were conducted face to face, and respondents were asked about ownership of financial products by their household and used their financial administration to verify answers, i.e. households were asked to retrieve their bank and insurance records to check which products they own. This leads to a highly accurate representation of the product portfolios of the households.

A representative sample of 7676 households participated in this study. Not all households participated in each wave of the panel, as a result of attrition or signing up with the panel after 1996. Vermunt *et al.* (1999) showed that missing data can easily be accommodated under the missingness at random assumption. This is relevant when applying the LMM to panel data with attrition. The missingness at random assumption is appropriate for the database that is analysed, i.e. the procedure replacing households that drop out of the Growth from Knowledge

Table 1. Levels of ownership of the products analysed in each panel wave

Product	Levels of ownership in the following years:			
	1996	*1998*	*2000*	*2002*
Bonds	0.04	0.04	0.03	0.03
Shares	0.08	0.10	0.11	0.11
Investment trust	0.11	0.20	0.25	0.21
Unemployment insurance	0.10	0.21	0.22	0.22
Loan	0.24	0.20	0.20	0.15
Credit card	0.29	0.30	0.35	0.41
Mortgage	0.52	0.53	0.54	0.53
Life insurance	0.59	0.60	0.59	0.55
Pension fund	0.62	0.67	0.64	0.59
House insurance	0.62	0.64	0.64	0.66
Car insurance	0.76	0.77	0.77	0.78
Savings account	0.93	0.95	0.96	0.96

panel ensures that the sample remains representative for the population with regard to various important demographic variables, such as age, income and marital status.

Table 1 presents levels of penetration of the products analysed across the four measurement occasions. This set of products is similar to those which have been analysed in extant cross-sectional studies (e.g. Bijmolt *et al.* (2004), Kamakura *et al.* (1991), Paas and Molenaar (2005) and Soutar and Cornish-Ward (1997)). There is also information on

(a) net household income,
(b) age of the head of the household and
(c) size of the household.

These demographics will be the covariates for explaining the product portfolios that households have and changes therein. Previous research shows that these variables are strongly related to household financial product portfolios (Browning and Lusardi, 1996; Soutar and Cornish-Ward, 1997).

The LMM is estimated by using the first three panel waves, using Latent GOLD 5.0. The syntax that we used is provided in Appendix B. The 2002 wave is the hold-out sample for evaluating the out-of-sample forecasting accuracy of the prediction equations.

4.2. Model selection
We first assessed the fit of models by imposing the restrictions in equations (13) and (14) and using the criterion BIC. Increasing the number of latent classes, from 1 to 10, shows that a nine-class LMM, with these restrictions, is most suitable for our data set. Models with fewer (or more) classes result in higher BIC-values. The nine-class model has 196 parameters and BIC $= 95\,193$ and is called 'the final model'. Besides statistical fit, the final model is also readily interpreted. For example, Section 4.3 reports that households owning more risky investments also tend to own the more basic financial products. This is consistent with theory on financial product portfolios (Guiso *et al.*, 2002; Wärneryd, 1999). Other model outcomes, such as the covariate effects that are discussed in Section 4.3, are also consistent with extant theory.

To assess the feasibility of restrictions (13) and (14), the relative fit of the final model is compared with two alternative models. Besides these, a third benchmark assesses the feasibility of assuming change over the period in which the data analysed were collected. The three benchmark models are based on LMM specifications that Brangula-Vlagsma *et al.* (2002) introduced and evaluated as potentially suitable. As in the final model, the first benchmark assumes a time constant measurement component and allows switching between latent classes (latent change). However, this model assumes time-varying switching probabilities. Thus, the assumption that is represented by equation (13) is imposed, but the assumption in equation (14) is relaxed. The latter is the only difference between the first benchmark model and the final model. We tested such models with 1–10 segments. All have a higher BIC than the final model. The second benchmark model has a time-varying measurement component (manifest change) and embraces time constant transition probabilities. This model relaxes the assumption in equation (13) but maintains assumption (14). Again all models with 1–10 segments have higher BIC-values than the final model. The third benchmark, the no-change model, has a time constant measurement component and assumes that households remain in the same latent class over time. Models with 1–10 segments of this type were outperformed by the final model, in terms of BIC. In sum, the restrictions in equations (13) and (14) are highly plausible and product portfolios change significantly over the measurement occasions.

4.3. Results

Table 2 presents the measurement component of the final model. For enhancing interpretation, the nine latent classes are ranked according to increasing product penetration levels across the 12 products. Products are also ordered, from the least commonly owned, i.e. bonds, to the most commonly owned product, i.e. savings accounts.

The measurement model is quite consistent with results of extant cross-sectional studies, which assume a common order for acquiring financial products (e.g. Kamakura *et al.* (1991), Paas (1998) and Soutar and Cornish-Ward (1997)). In latent classes, where the probability of

Table 2. Measurement component

Product	*Results for the following classes:*								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Bonds	0.01	0.01	0.07	0.01	0.02	0.15	0.01	0.02	0.19
Shares	0.01	0.01	0.10	0.02	0.15	0.38	0.02	0.06	0.59
Investment trust	0.04	0.08	0.13	0.08	0.11	0.57	0.10	0.18	0.75
Unemployment insurance	0.00	0.00	0.01	0.20	0.02	0.16	0.30	0.38	0.29
Loan	0.10	0.11	0.03	0.46	0.13	0.10	0.29	0.37	0.11
Credit card	0.06	0.08	0.06	0.32	0.26	0.61	0.04	0.97	0.67
Mortgage	0.00	0.00	0.03	0.00	0.99	0.03	0.99	0.99	0.99
Life insurance	0.17	0.04	0.20	0.77	0.47	0.54	0.97	0.97	0.83
Pension fund	0.28	0.37	0.29	0.85	0.39	0.57	0.94	0.93	0.82
House insurance	0.00	0.01	0.99	0.01	0.99	0.52	0.99	0.97	1.00
Car insurance	0.01	0.99	0.72	0.83	0.71	0.89	0.89	0.93	0.95
Savings account	0.81	0.88	0.95	0.93	0.96	0.95	0.95	0.98	1.00
Average number of products	1.49	2.58	3.58	4.19	5.20	5.47	6.49	7.75	8.19

Table 3. Transition matrix

Latent class at t	Results for the following latent classes at $t+1$:								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	0.95	0.03	0.00	0.01	0.01	0.00	0.00	0.00	0.00
2	0.05	0.93	0.00	0.00	0.01	0.00	0.01	0.00	0.00
3	0.01	0.04	0.93	0.00	0.01	0.00	0.01	0.00	0.00
4	0.00	0.00	0.00	0.94	0.03	0.00	0.03	0.00	0.00
5	0.01	0.01	0.01	0.00	0.95	0.01	0.00	0.01	0.01
6	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.01	0.02
7	0.00	0.00	0.00	0.01	0.00	0.00	0.87	0.11	0.01
8	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.94	0.03
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

owning a less commonly owned product is high, households also tend to own products with higher overall penetrations (as reported in Table 1). This pattern suggests a common order of acquisition (Bijmolt *et al.*, 2004; Paas and Molenaar, 2005). However, the results suggest some divergence from the common order. For example, in class 6 the credit card has a higher penetration than the mortgage product, whereas in the whole sample the latter product has a higher level of penetration. This suggests that most households acquire a mortgage before owning a credit card, whereas some acquire a credit card before a mortgage. Similar situations occur because of the relatively low probabilities of ownership for house insurance in class 4. Besides this, loans and unemployment insurance cut across the entire acquisition pattern, because the rate of penetration is far from monotonically increasing from class 1 to class 9.

Table 3, the latent transition matrix, presents the probabilities of switching classes. A large percentage of households remained in the same latent class between consecutive measurement occasions, as indicated by the high proportions in the cells on the diagonal. However, 14% of the 7676 households switched latent class in the period 1996–2000. This percentage is consistent with the expected gradual development of a household's financial product portfolios (Browning and Lusardi, 1996; Guiso *et al.*, 2002).

Table 3 also provides a more precise interpretation of the dynamics of the acquisition pattern. In particular, Table 3 shows how often certain switches occur. For example, the most common switch is from latent class 7 to 8 (11% between consecutive measurement occasions). Extant cross-sectional studies do not provide such information. Also, we find that households in latent class 4 relatively often switch into latent class 7. The average household in latent class 7 owns 6.49 products, whereas the average number is only 4.19 in class 4. This switch implies that a relatively rapid accumulation of financial products may occur, as switching between these classes implies the acquisition of multiple products between consecutive measurement occasions. Again, extant cross-sectional studies do not report such tendencies.

All three covariates significantly affect initial latent class membership, in 1996 (income, Wald coefficient 709.34, 24 degrees of freedom, $p < 0.01$; age, Wald coefficient 405.06, 24 degrees of freedom, $p < 0.01$; household size, Wald coefficient 288.40, 16 degrees of freedom, $p < 0.01$). For brevity we discuss only the general tendencies. First, high income households are relatively often allocated to latent classes with relatively high ownership probabilities for most products. For age of the head of the household a similar tendency occurs; only now intermediate age takes on the role of the high income category. These effects are consistent with extant

theory (Browning and Lusardi, 1996; Guiso *et al.*, 2002; Wärneryd, 1999). Furthermore, larger households are relatively often found in latent classes where overall product ownership probabilities are relatively low. Perhaps, having more children leads to higher expenditures and, therefore, fewer assets will be left for financial products. The three covariates also significantly affect switching probabilities (income, Wald coefficient 45.95, 24 degrees of freedom, $p < 0.004$; age, Wald coefficient 66.35, 24 degrees of freedom, $p < 0.0001$; household assets, Wald coefficient 70.81, 16 degrees of freedom, $p < 0.0001$). Thus, the covariates can improve the predictions of which households in a specific latent class are more likely to switch to another latent class. This is important for the accuracy of forecasting of the LMM. The covariate effects on switching probabilities are similar to the effects on initial latent class membership, i.e. where values on covariates imply a larger probability of belonging to an initial latent class s the model tends to show that these covariate values also imply a greater probability of switching into s.

5. Forecasting accuracy of the latent Markov model

The prediction equations (15)–(17) are applied to the LMM that was derived from the 1996–2000 data. The data from the 2002 measurement occasion are used as the hold-out sample. This specific analysis employs only data for the 2239 households that participated in the survey in both 2000 and 2002. The extent to which households predicted as owning product j in 2002 are distinguished from households not owning this product in 2002 quantifies the accuracy of forecasting of the prediction equations. This distinction is made between households that do not own product j in 2000.

The Gini coefficient, a measure of concentration, is used for assessing the accuracy of forecasting of the prediction equations. Kamakura *et al.* (2003) previously applied the Gini coefficient to evaluate the accuracy of forecasting of models assessing product ownership at different financial firms. We use the definition which is commonly accepted (Sen, 1997):

$$\text{Gini}_j = 1 + \frac{1}{n_j} - \frac{2}{n_j^2 \mu_j} \sum_{i=1}^{n} r_{ij} y_{ij} \tag{18}$$

where n_j represents the number of households not owning product j in 2000, μ_j is the percentage of these n_j households that own product j in 2002 and r_{ij} is the rank of household i with regard to the predicted probability of owning product j in 2002. Values of r_{ij} are higher as the probability for owning product j is lower. Furthermore, $y_{ij} = 1$ when household i owns product j in 2002; otherwise $y_{ij} = 0$. Values of Gini_j range from 0 to 1 where 0 implies that predictions are no better than random and value 1 implies perfect forecasting.

The out-of-sample forecasting accuracy and the value of Gini_j can be presented graphically. For example, the x-axis in Fig. 1 represents the cumulative percentage of households not owning mortgages in 2000. These households are ordered on the basis of the predicted probability that they will own a mortgage in 2002. Households with large predicted probabilities are found closer to the origin. Further from this point are those households with smaller predicted probabilities. The y-axis displays the cumulative percentage of households actually owning a mortgage in 2002. Consider the 10% of respondents without a mortgage in 2000, and with the highest predicted probability of owning this product in 2002. Fig. 1 shows that 37% of all the respondents that *do not* own a mortgage in 2000 but *do* own this product in 2002 are among this 10% group. This is considerably better than random predictions that are represented by the diagonal line. More generally, for Fig. 1 $\text{Gini}_j = 0.47$. This implies that 47% of the surface above the straight diagonal line, representing forecasting accuracy under random prediction, is found between the
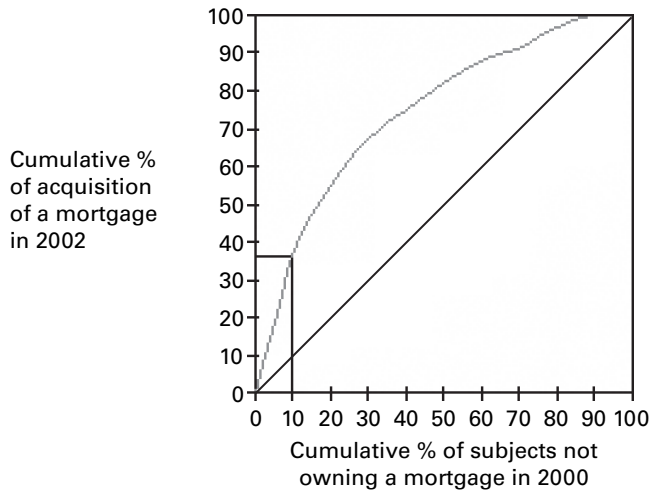
Fig. 1. Power curve (the term 'power curve' refers to the diagonal line representing the accuracy of forecasting of the model-based predictions; the term in this context should not be confused with its more standard use in inference)

straight diagonal line and the power curve. As more of the surface above the straight diagonal line lies under the power curve, the accuracy of forecasting is higher and hence the value of $Gini_j$ increases.

Table 4 reports values of $Gini_j$ for all 12 LMM-based predictions. The following seven products have power curves that are similar to the curve in Fig. 1: bonds, shares, investment trusts, unemployment insurances, loans, credit cards and house insurance. The accuracy of forecasting is somewhat lower for savings accounts. This product has a very high level of penetration at all measurement occasions; at least 93% of the households own this product at each occasion (see Table 1). Perhaps the LMM-based predictions are less effective for predicting acquisitions of such commonly owned products. Also, predictions are less accurate for life insurance, and pension funds and least effective for car insurance. It is possible that car insurance is redundant for many households in the Netherlands, owing to lease cars that are supplied by employers.

Table 4. Gini coefficient values

Product	Gini coefficient
Bonds	0.50
Shares	0.31
Investment trust	0.26
Unemployment insurance	0.36
Life insurance	0.17
Pension fund	0.19
Loan	0.49
Credit card	0.62
Mortgage	0.47
House insurance	0.40
Car insurance	−0.23
Savings account	0.11

Such conditions of labour could be unrelated to household behaviour in the financial services market, and ownership of this product may, therefore, be inadequately modelled in combination with ownership of other financial products. Similar labour conditions exist for life insurance and pension funds. Perhaps future research should employ covariates reflecting the conditions of labour of household members.

6. Discussion

This paper presents the LMM as a suitable technique for investigating acquisition patterns based on longitudinal data. We formulated equations supporting application of the LMM for assessing the out-of-sample probability that specific individuals or household units would acquire the overt characteristics defining the latent stages of a process. This is a novel utilization of the LMM. We employed the out-of-sample predictions to assess predictive validity of the LMM and to forecast household acquisitions of financial products. The paper has both theoretical and methodological contributions.

From the theoretical perspective, the LMM provided substantive insight into the orders in which households acquire financial products. Consistent with previous studies, a common order of acquisition occurs for many products. The results presented also provide new insights. In particular, some products do not fully fit into a single order of acquisition. For example, some households acquire a mortgage before owning a credit card, whereas others acquire a credit card before they own a mortgage. Also, the loan product and unemployment insurance in our data set are not captured in the acquisition pattern. This contradicts extant theory, assuming that a utility structure of financial objectives leads to a common order of product acquisitions which all households tend to follow (Kamakura *et al.*, 1991; Paas, 1998; Soutar and Cornish-Ward, 1997). Concerning the dynamics of the acquisition pattern, little switching occurs between latent classes. This is consistent with the expectation that households only gradually develop their financial product portfolios (Guiso *et al.*, 2002; Wärneryd, 1999). Besides this, covariate effects in the model reported are also highly plausible, owing to consistency with the life cycle model (Browning and Lusardi, 1996). New in the current paper is the provision of insight into the timing of acquisition of products by the LMM. Unexpected is the finding that switches between latent classes may involve the acquisition of multiple products in the relatively short period of 2 years between the measurement occasions. This behaviour was not considered in previous studies into acquisition patterns of financial products.

The methodological contribution is that we introduced prediction equations (15)–(17) for forecasting out-of-sample behaviour. Moreover, the empirical results reported demonstrate that the prediction equations (15)–(17) are effective for forecasting household acquisitions for most products in our database. This suggests high predictive validity of the LMM reported and that this model is suitable for predicting future household acquisition behaviour in the financial products market. For consumer behaviour and economics these findings imply that the LMM can predict the entrance into product categories, by individual consumers or household units, in other markets where consumers use products over longer periods, e.g. the durable product market or the market for telecommunication products. More generally, the paper demonstrates that the LMM can support out-of-sample prediction. This is important for various prediction purposes involving individual progression through stages defining a process, such as the use of different illicit drugs or criminal behaviour, to which the LMM has previously been applied (Collins and Wugalter, 1992; Graham *et al.*, 1991; Bartolucci *et al.*, 2007). Nevertheless, further research should provide a more rigorous explanation of the inaccurate predictions for one of the

12 products in our data set (car insurance) and the differences between accuracy of forecasting between the other 11 products.

Appendix A: The connection between the key components of the forward–backward algorithm and the posterior probabilities

The univariate posterior probability may be expressed as

$$P(X_{it} = s | \mathbf{Y}_i, \mathbf{Z}_i) = \frac{P(X_{it} = s, \mathbf{Y}_i | \mathbf{Z}_i)}{P(\mathbf{Y}_i | \mathbf{Z}_i)} = \frac{P(X_{it} = s, \mathbf{Y}_{i,t-}, \mathbf{Y}_{i,t+} | \mathbf{Z}_{i,t-}, \mathbf{Z}_{i,t+})}{P(\mathbf{Y}_i | \mathbf{Z}_i)}. \tag{19}$$

Here the observed data have been separated into disjoint sets; occurring at time t or before indexed by $t-$, and for time points after it indexed by $t+$. Using basic rules of probability the numerator on the right-hand side of expression (19) can be expressed as

$$P(X_{it} = s, \mathbf{Y}_{i,t-} | \mathbf{Z}_{i,t-}, \mathbf{Z}_{i,t+}) \, P_i(\mathbf{Y}_{i,t+} | X_{it} = s, \mathbf{Y}_{i,t-}, \mathbf{Z}_{i,t-}, \mathbf{Z}_{i,t+}). \tag{20}$$

The conditioning on $\mathbf{Z}_{i,t+}$ in the first term can be removed since the situation at t is independent of covariate values after it. Also because of the Markov process, the latent class sequence after time t and therefore $\mathbf{Y}_{i,t+}$ are independent of the history before time t. Thus expression (20) becomes $P(X_{it} = s, \mathbf{Y}_{i,t-} | \mathbf{Z}_{i,t-}) \, P_i(\mathbf{Y}_{i,t+} | X_{it} = s, \mathbf{Z}_{i,t+})$ and, on substitution in equation (19),

$$\begin{aligned} P(X_{it} = s | \mathbf{Y}_i, \mathbf{Z}_i) &= \frac{P(X_{it} = s, \mathbf{Y}_{i,t-} | \mathbf{Z}_{i,t-}) \, P(\mathbf{Y}_{i,t+} | X_{it} = s, \mathbf{Z}_{i,t+})}{P(\mathbf{Y}_i | \mathbf{Z}_i)} \\ &= \frac{\alpha_{it}(s) \, \beta_{it}(s)}{P(\mathbf{Y}_i | \mathbf{Z}_i)} \end{aligned} \tag{21}$$

which is expression (7) in the text on defining $\alpha_{it}(s)$ and $\beta_{it}(s)$.

Using similar ideas and the properties of independence of $\mathbf{Y}_{i,t+}$ of prior history conditional on the state at time t the posterior bivariate probability can initially be expressed as

$$\begin{aligned} P(X_{i,t-1} = r, X_{it} = s | \mathbf{Y}_i, \mathbf{Z}_i) &= \frac{P(X_{i,t-1} = r, X_{it} = s, \mathbf{Y}_i | \mathbf{Z}_i)}{P(\mathbf{Y}_i | \mathbf{Z}_i)} \\ &= \frac{P(X_{i,t-1} = r, X_{it} = s, \mathbf{Y}_{it-} | \mathbf{Z}_{it-}) \, P(\mathbf{Y}_{it+} | X_{it} = s, \mathbf{Z}_{it+})}{P(\mathbf{Y}_i | \mathbf{Z}_i)} \\ &= \frac{P(X_{i,t-1} = r, X_{it} = s, \mathbf{Y}_{it-} | \mathbf{Z}_{it-}) \, \beta_{it}(s)}{P(\mathbf{Y}_i | \mathbf{Z}_i)}. \end{aligned} \tag{22}$$

Now $\mathbf{Y}_{it-}$ can be separated into disjoint sets $\mathbf{Y}_{it}$ at time t and $\mathbf{Y}_{i,(t-1)-}$ for time point $t-1$ and all earlier time points, as can $\mathbf{Z}_{it-}$ in a similar way. The first probability in the numerator above by using basic rules can then be written

$$P(X_{i,t-1} = r, \mathbf{Y}_{i,(t-1)-} | \mathbf{Z}_{it}, \mathbf{Z}_{i,(t-1)-}) \, P(X_{it} = s, \mathbf{Y}_{it} | X_{i,t-1} = r, \mathbf{Y}_{i,(t-1)-}, \mathbf{Z}_{it}, \mathbf{Z}_{i,(t-1)-}). \tag{23}$$

The first probability in expression (23) is $\alpha_{it-1}(s)$ since $\mathbf{Z}_{it}$ as a future covariate can be dropped from the conditioning. The second probability in expression (23) reduces to $P(X_{it} = s, \mathbf{Y}_{it} | X_{it-1} = r, \mathbf{Z}_{it})$ owing

to the historical properties of the Markov process and on further decomposition becomes $P(X_{it} = s | X_{i,t-1} = r, \mathbf{Z}_{it}) \, P(Y_{it} | X_{it} = s)$. On substitution of these results into expression (23) and then back into equation (22)

$$P_i(X_{it-1} = r, X_{it} = s | \mathbf{Y}_i, \mathbf{Z}_i) = \frac{\alpha_{it-1}(s) \, P(X_{it} = s | X_{i,t-1} = r, Z_{it}) \, P(Y_{it} | X_{it} = s) \, \beta_{it}(s)}{P(\mathbf{Y}_i | \mathbf{Z}_i)},$$

which is equation (8) in the text.

The denominator $P(\mathbf{Y}_i | \mathbf{Z}_i)$ which is required for both univariate and bivariate posteriors may be found in terms of $\alpha_{it}(s)$ and $\beta_{it}(s)$ as follows. For arbitrarily chosen t $P(X_{it} = s, \mathbf{Y}_i | \mathbf{Z}_i) = \alpha_{it}(s) \, \beta_{it}(s)$ as shown above. Marginalizing to $\mathbf{Y}_i$ by summing over the latent classes yields

$$P_i(\mathbf{Y}_i | \mathbf{Z}_i) = \sum_{s=1}^{S} \alpha_{it}(s) \, \beta_{it}(s).$$

The simplest form of this is for $t = T$ when $P_i(\mathbf{Y}_i | \mathbf{Z}_i) = \sum_{s=1}^{S} \alpha_{iT}(s)$, i.e. when $\beta_{iT}(s) = 1$.

Appendix B: Syntax for Latent GOLD 5.0

We used Latent GOLD version 5.0 to obtain the results that are reported in the text. This is the syntax file for estimating the LMM of interest:

```
options
   missing all;
   coding last;
variables
   caseid id;
   dependent Y1, Y2, Y3, Y4, Y5, Y6, Y7, Y8, Y9, Y10, Y11, Y12;
   independent age nominal, income nominal, hsize nominal;
   latent X nominal markov 9;
equations
   X[=0] <- 1 + age + income + hsize;
   X <- 1 | X[-1] + age + income + hsize;
   Y1 <- 1 + X;
   Y2 <- 1 + X;
   Y3 <- 1 + X;
   Y4 <- 1 + X;
   Y5 <- 1 + X;
   Y6 <- 1 + X;
   Y7 <- 1 + X;
   Y8 <- 1 + X;
   Y9 <- 1 + X;
   Y10 <- 1 + X;
   Y11 <- 1 + X;
   Y12 <- 1 + X;
```

As options we indicate records with missing values that should be retained in the analysis and that dummy coding applies to the last category as reference category. The variables section defines the dependent, independent and latent variables as well as the `case id` connecting the multiple records of a case (the data file should be a person–period file). The identifier `markov` indicates that the nominal latent variable X is a dynamic latent variable. The equations section defines the logit equations that were described in the text: one equation for the initial state probability, one equation for the transition probabilities and one for each of the 12 response variables (observed ownerships). Provided that X is the label for the latent variable, the initial state at $t = 0$ is denoted `X[=0]`, the state at a particular time point t by X and the state at time point $t - 1$ by `X[-1]`. The right-hand side terms in the equations are self-explanatory, except for the '`|X[-1]`' indicating that the logit parameters concerned (the constants in the equation for X) vary across origin states (across level of `X[-1]`).

References

Bartolucci, F., Pennoni, F. and Francis, B. (2007) A latent Markov model for detecting patterns of criminal activity. *J. R. Statist. Soc.* A, **170**, 115–132.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41**, 164–171.

Bijmolt, T. H. A., Paas, L. J. and Vermunt, J. K. (2004) Country and consumer segmentation: multi-level latent class analysis of financial product ownership. *Int. J. Res. Markt.*, **21**, 323–340.

Brangula-Vlagsma, K., Pieters, R. G. M. and Wedel, M. (2002) The dynamics of value segments: modeling framework and empirical illustration. *Int. J. Res. Markt.*, **19**, 267–285.

Browning, M. and Lusardi, A. (1996) Household savings: micro theories and micro facts. *J. Econ. Lit.*, **34**, 1797–1855.

Collins, L. M. and Wugalter, S. E. (1992) Latent class models for stage-sequential dynamic latent variables. *Multiv. Behav. Res.*, **27**, 131–157.

Croon, M. A. (1991) Investigating Mokken scalability of dichotomous items means of ordinal latent class analysis. *Br. J. Math. Statist. Psychol.*, **44**, 315–332.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Erdem, T. and Sun, B. (2001) Testing for choice dynamics in panel data. *J. Bus. Econ. Statist.*, **19**, 142–152.

Fearnhead, P. and Meligkotsidou, L. (2004) Exact filtering for partially observed continuous time models. *J. R. Statist. Soc.* B, **66**, 771–789.

Feick, L. F. (1987) Latent class models for analysis of behavioural hierarchies. *J. Markt. Res.*, **24**, 174–186.

Freedman, D., Thorton, A., Camburn, D., Alwin, D. and Young-DeMarco, L. (1988) The life-history calendar: a technique for collecting retrospective data. In *Sociological Methodology 1988* (ed. C. C. Clogg), pp. 37–68. Oxford: Basil Blackwell.

Gatignon, H. and Robertson, T. S. (1985) A propositional inventory for new diffusion research. *J. Consum. Res.*, **11**, 849–867.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

Graham, J. W., Wugalter, S. E., Chung, N. K. and Hansen, W. B. (1991) Modeling transitions in latent stage-sequential processes: a substance use prevention example. *J. Consult. Clin. Psychol.*, **59**, 48–57.

Guiso, L., Haliossos, M. and Jappelli, T. (2002) *Household Portfolios*. Cambridge: MIT Press.

Kamakura, W. A., Ramaswami, S. N. and Srivastava, R. K. (1991) Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *Int. J. Res. Markt.*, **8**, 329–349.

Kamakura, W. A., Wedel, M., De Rosa, F. and Mazzon, J. A. (2003) Cross-selling through database marketing: a mixed data factor analyser for data augmentation and prediction. *Int. J. Res. Markt.*, **20**, 45–65.

Kingma, J. (1984) A comparison of four methods of scaling for the acquisition of early number concept. *J. Gen. Psychol.*, **110**, 23–45.

Lazarsfeld, P. F. and Henry, N. W. (1968) *Latent Structure Analysis*. Boston: Houghton Mifflin.

Li, S., Sun, B. and Wilcox, R. T. (2005) Cross-selling sequentially ordered products: an application to consumer banking services. *J. Markt. Res.*, **42**, 233–239.

McDonald, I. L. and Zucchini, W. (1997) *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman and Hall.

Means, B., Swan, G. E., Jobe, J. B. and Esposito, J. L. (1991) An alternative approach to obtaining personal history data. In *Measurement Errors in Surveys* (eds P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz and S. Sudman), pp. 127–144. New York: Wiley.

Paas, L. J. (1998) Mokken scaling characteristic sets and acquisition patterns of durable- and financial products. *J. Econ. Psychol.*, **19**, 353–376.

Paas, L. J., Kuijlen, A. A. A. and Poiesz, Th. B. C. (2005) Acquisition pattern analysis for relationship marketing: a conceptual and methodological redefinition. *Serv. Indust. J.*, **25**, 661–673.

Paas, L. J. and Molenaar, I. W. (2005) Analysis of acquisition patterns: a theoretical and empirical evaluation of alternative methods. *Int. J. Res. Markt.*, **22**, 87–100.

Paroush, J. (1965) The order of acquisition of consumer durables. *Econometrica*, **33**, 225–235.

Proctor, C. H. (1970) A probabilistic formulation and statistical analysis of Guttman scaling. *Psychometrika*, **35**, 73–78.

Sen, A. (1997) *On Economic Inequality*. Oxford: Clarendon.

Soutar, G. N. and Cornish-Ward, S. (1997) Ownership patterns for durable goods and financial assets: a Rasch analysis. *Appl. Econ.*, **29**, 903–911.

Uncles, M., Ehrenberg, A. and Hammond, K. (1995) Patterns of buyer behaviour: regularities, models, and extensions. *Markt. Sci.*, **14**, G71–G78.

Van de Pol, F. and Langeheine, R. (1990) Mixed Markov latent class models. In *Sociological Methodology 1990* (ed. C. C. Clogg), pp. 213–247. Oxford: Basil Blackwell.

Vermunt, J. K. (2001) The use of latent class models for defining and testing non-parametric and parametric item response theory models. *Appl. Psychol. Measmnt*, **25**, 283–294.

Vermunt, J. K., Langeheine R. and Böckenholt, U. (1999) Latent class Markov models with time-constant and time-varying covariates. *J. Educ. Behav. Statist.*, **24**, 178–205.

Vermunt, J. K. and Magidson, J. (2007) *LG—Syntax User's Guide: Manual for Latent Gold 4.5 and Latent Gold Choice 4.5 Syntax Module*. Belmont: Statistical Innovations. To be published.

Verweij, A. C., Sijtsma, K. and Koops, W. (1999) An ordinal scale for transitive reasoning by means of a deductive strategy. *Int. J. Behav. Devlpmnt*, **23**, 241–264.

Wärneryd, K.-E. (1999) *The Psychology of Saving: a Study of Economic Psychology*. Northampton: Elgar.

Wedel, M. and Kamakura, W. A. (2000) *Market Segmentation: Conceptual and Methodological Foundations*, 2nd edn. Amsterdam: Kluwer.