Chapter XX

Hidden Markov Models in Marketing

Oded Netzer, Columbia University

Peter Ebbes, HEC Paris

Tammo Bijmolt, University of Groningen

Please cite as:

Netzer, O., P. Ebbes, and T. Bijmolt (2016), "Hidden Markov Models in Marketing," forthcoming as Chapter XX in Advanced Methods in Modeling Markets, Eds. P.S.H. Leeflang, J.E. Wieringa, T.H.A. Bijmolt, and K. H. Pauwels, Springer International Series in Quantitative Marketing.

XX.1 Introduction: Capturing Dynamics

Hidden Markov models (HMMs) have been used to model how a sequence of observations is governed by transitions among a set of latent states. HMMs were first introduced by Baum and co-authors in late 1960s and early 1970 (Baum and Petrie 1966; Baum et al. 1970), but only started gaining momentum a couple decades later. HMMs have been applied in various domains such as speech or word recognition (Rabiner 1989), image recognition (Yamato, Ohya and Ishii 1992), economics (Hamilton 1989, 2008), finance (Mamon and Elliott 2007), genetics (Eddy 1998), earth studies (Hughes and Guttorp 1994), and organization studies (Wang and Chan 2011). Over the last decade the number of applications of HMMs in marketing has grown substantially (see section XX.4).

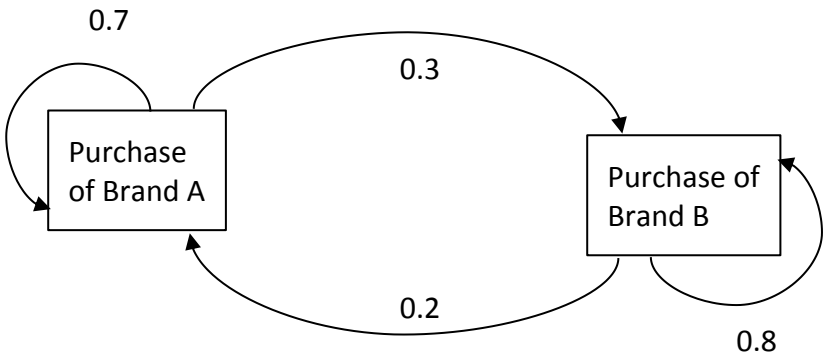
In the context of marketing HMMs are often used to model a time series of customer or firm behavior such as customer choices or firm sales. These observations evolve over time following a latent Markovian process. That is, the firm or customer transition over time (in a Markovian manner) among a set of latent states and given each one of the states the customer or firm (probabilistically) behaves in a particular fashion. The observations provide a noisy measure of the underlying state. The main objective in utilizing a HMM is often to capture the dynamics in customer behavior over time. For simplicity we will describe the HMM in this chapter in the context of capturing dynamics in customer behavior and how firm actions may influence these behaviors. We note that HMMs in marketing are not limited to modeling behavior of customers, and have been applied in B2B contexts where the unit of analysis is a firm (see details in Section XX.4).

Markovian models (see e.g. Leeflang et al. 2015, section 7.2) have been used in marketing to capture dynamics in customer behavior since the mid-1960s (e.g., Ehrenberg 1965). In these models the customer's choice at time t is assumed to be a function of the customer's choice at time $t-1$, and according to a typical Markov model, depends only on the customer's choice at time $t-1$ and not the customer's choice in earlier time periods. This type of Markovian relationship between current customer choices and previous choices has been often referred to in marketing and economics as state dependence (e.g., Keane 1997; Chintagunta 1998; Seetharaman 2004, Dubé, Hitsch and Rossi 2010).

To illustrate the notion of state dependence, consider a customer choice between Brand A and Brand B. State dependence suggests that the customer may have a different utility for Brand A depending on whether brand A or brand B was previously chosen. For

example, positive state dependence suggests that the customer's utility from choosing Brand A will be higher if the customer purchased Brand A (rather than Brand B) in the previous time period. A related construct called variety seeking would predict the opposite effect, such that following a purchase of Brand A the utility the customer obtains from choosing Brand A again will be lower than its utility had the customer purchased Brand B instead. For example, consider the Markov process of purchase probabilities of brands A and B in Figure XX.1.

Figure XX.1: A Markov model of brand choice

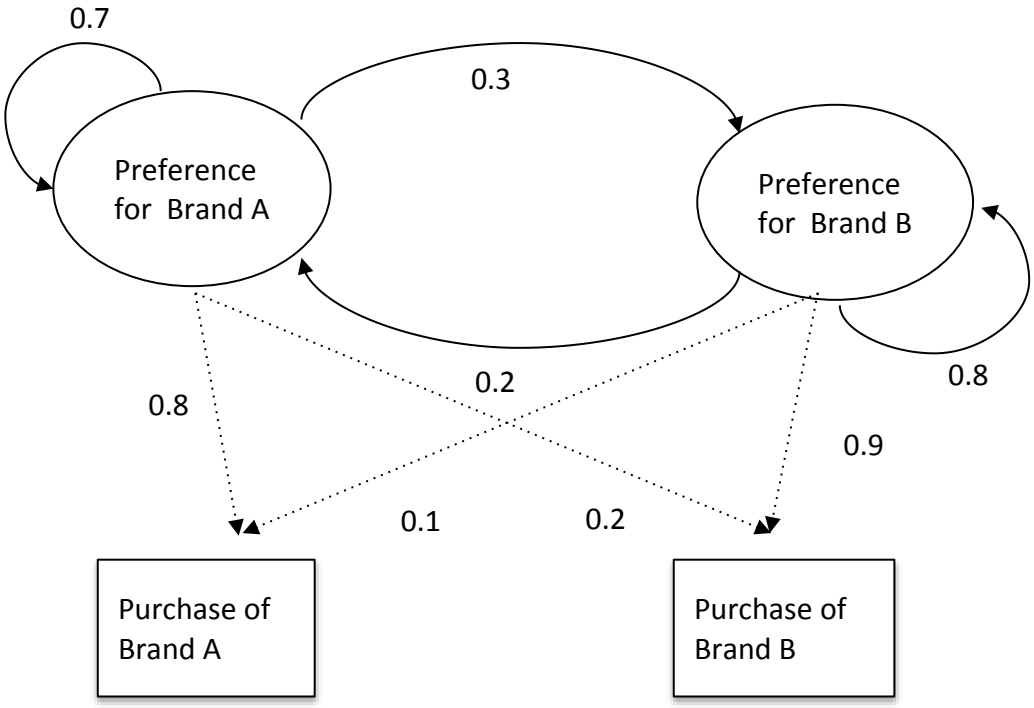


Based on Figure XX.1, the probability of buying Brand A given that Brand A was previously chosen is 0.7, i.e. $P(A_t|A_{t-1}) = 0.7$, and the probability of buying Brand B given that Brand A was previously chosen is 0.3, i.e. $P(B_t|A_{t-1}) = 0.3$. Similarly, the probability of buying Brand B given that Brand B was previously chosen is 0.8, i.e. $P(B_t|B_{t-1}) = 0.8$, and the probability of buying Brand A given that Brand B was previously chosen is 0.2, i.e. $P(B_t|B_{t-1}) = 0.2$. Thus, this example demonstrates positive state dependence as staying with the same brand from one period to the next is

considerably more likely than switching to the other brand. The model of customer behavior as depicted in Figure XX.1 is fairly simplistic. This model assumes that the customer purchase in the current period dependent *only* on the customer's purchase in the previous period. This is a result of two modeling assumptions: 1) that the state of the world is defined purely based on the customer's observed purchase in the previous period, and 2) the Markovian assumption that only the last purchase and not the purchases before the last matter. The first assumption could be relaxed by adding observed variables that may affect the customer behavior such as advertising or price as covariates in the model. The second assumption could be relaxed by defining the state by a longer history of purchases such as a running average of past purchases or a weighted sum of past purchases (e.g., Guadagni and Little 1983).

An additional limitation of the model depicted in Figure XX.1, or its extensions described above, is that these models assume that the customer state can be fully characterized by the observed behavior. However, the customer decision of which product to purchase is often governed by an underlying latent state of preference for different brands. While the customer may switch brands at times due to stock out or a visitor from out of town without changing her intrinsic preferences, the underlying preferences are likely to be stickier and better reflecting the long-term customer behavior. HMMs offer a solution to this difficulty by proposing a model of *latent* customer preference and the transitions among them. In the context of the example described in Figure XX.1, one could model the customer behavior using a HMM as shown in Figure XX.2.

Figure XX.2: A hidden Markov model of brand choice



In the model described in Figure XX.2, the two states represent the customer latent preference states for brand A and brand B. Unlike Figure XX.1, the states in Figure XX.2 are unobserved. Given the customer's latent preference state the customer probabilistically choses the brands (the observed purchases). For example, a customer who has a higher preference for Brand B does not choose Brand B with probability 1 but rather with probability 0.9, for example, because the customer may occasionally have a visitor from out of town that prefers Brand A. Similarly, when the customer is in the preference for Brand A state she has a probability of 0.8 to choose Brand A and a probability of 0.2 to choose Brand B. Thus, in a HMM the observed behavior (purchases)

serve as a noisy measure of the customer's true state. Additionally, the customer may change her preference for the brands over time. Such preference evolution will follow a Markovian process. In the example in Figure XX.2, the customer has an 80% chance of staying in the preference for brand B state from one period to another and a 20% of transition to the Brand A preference state. We call these the transition probabilities. In the example in Figure XX.2, the transition probabilities for a customer in the preference for brand A state are [0.7 0.3] and in the preference for brand B state [0.2 0.8]. Because the states in Figure XX.2 are fairly sticky, once a customer transitions to a different preference state she is likely to stay there for a while.

One may wonder how the latent preference states can be identified from a sequence of observed purchases. If the researcher observes a sequence of purchases that involves mainly purchases of Brand A (though the customer may occasionally purchase Brand B), the researcher will infer that the customer belongs to the Brand A preference state. If at some point the customer starts buying more frequently Brand B, the researcher may infer that the customer has transitioned to the Brand B preference state. If the researcher also observes some marketing actions along with the observed purchases, she can relate these to the transition probabilities for the underlying preference states in order to understand their effect on shifting consumers' preferences.

Thus, the HMM in Figure XX.2, and HMMs in general, have two main components: 1) a stochastic state dependent distribution – given a state the observations are stochastically determined, and 2) a state Markovian evolution – the system can transition from one state to another according to a set of transition probabilities.

Note that if the customer choses Brand A (B) with probability 1 when she is in preference state A (B), the HMM in Figure XX.2 collapses to the observed Markov process in Figure XX.1. Thus, the distinction between a HMM and an observed Markov process model is that in a HMM the states are stochastically determined by the sequence of observations, whereas in a Markov model the observations deterministically determine the states.

An alternative way of thinking about a HMM of customer purchase behavior, is to think about a HMM as an approach to incorporate time dynamics in customer preferences and responses to marketing actions. Consider for example a customer that has the following utility function, as is commonly described in marketing and economics choice models:

$$u_{itj} = X'_{itj}\beta_{it} + \varepsilon_{itj}, \qquad \text{(XX.1)}$$

for $i = 1,2,\dots,N, J = 1,2,\dots,J$, and $t = 1,2,\dots,T$. In this model u_{itj} is customer i's utility for product j at time t, X_{ijt} is a $P \times 1$ vector of time-varying, customer-specific, covariates relevant for product *j* and customer *i*, such as price and advertising, β_{it} is a $P \times 1$ vector of customer-specific and time varying response parameters, and ϵ_{itj} is an error term, capturing unobserved shocks. In the model described in Equation XX.1 the vector β_{it} varies across customers and time, thus capturing full heterogeneity and dynamics in customer preferences and customer responses to the covariates in X_{ijt}.

Estimating such model without putting any structure on the heterogeneity distribution across customers or across time (or both), is largely impractical for most empirical applications in marketing, because we often observe at most one observation per customer per time period. Two main approaches have been suggested in the literature

to capture unobserved, cross-customer, heterogeneity in β_{it}, but without capturing dynamics (i.e. $\beta_{i1} = \beta_{i2} = \cdots = \beta_{iT}$). The first approach is a latent class or finite mixture approach (see Chapter YY on Mixture Models) in which, instead of estimating a preference vector (β_i) for each individual, the researcher estimates a smaller set of vectors $\tilde{\beta}_s$, where $s = 1,2 \ldots, S$, and $S \ll N$. Here, the S latent classes are sometimes interpreted as segments (e.g. Wedel and Kamakura, 2000). Another approach is the random effects approach in which a multivariate distributional structure is assumed to describe the heterogeneity in β_i in the population of customers (e.g., $\beta_i \sim N(\mu_\beta, \Sigma_\beta)$) (see Chapter YY on Bayesian Models). Here, each customer is assumed to be unique in its preferences (i.e. form its own segment of size 1), but the preferences are drawn from a population distribution.

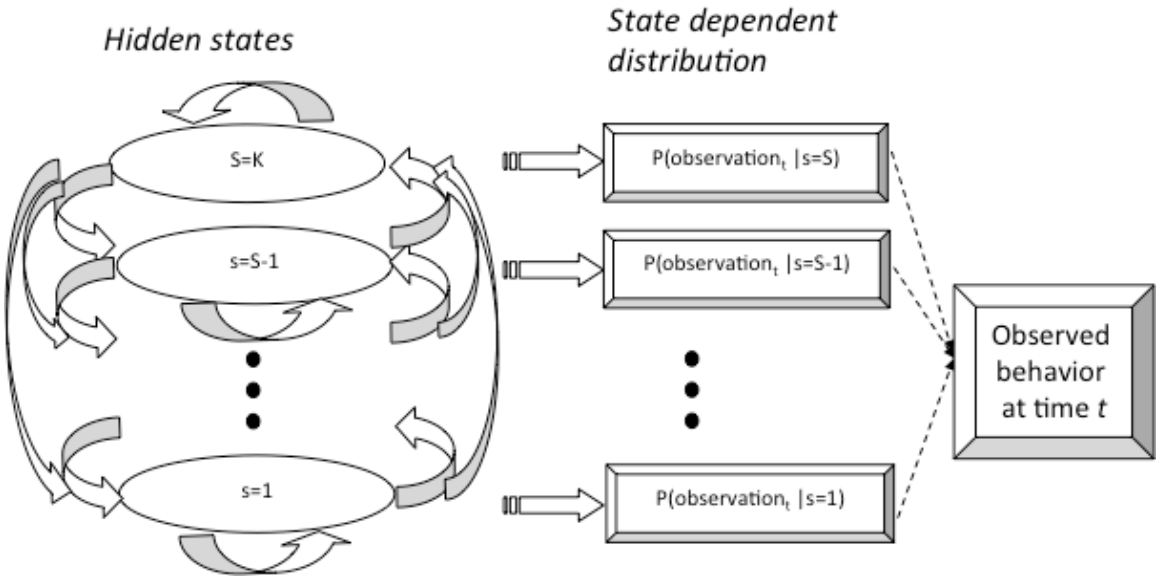
In a similar manner one can define a distribution for how the vector of parameters β_{it} varies over time. On the one hand, HMMs can be thought of as the dynamic analogue to the latent class or finite mixture approach. On the other hand, dynamic linear models (DLMs) based on the Kalman filter approach (see Chapter YY on State Space Models) can be seen as the dynamic analogue to the random-effects approach.

Now that we have introduced the basic intuition behind the HMM and its relationship to other models in marketing, we detail the components of the HMM, the modeling considerations that one needs to take into account when building a HMM, and we highlight the importance of accounting for cross-customer heterogeneity when estimating a HMM.

XX.2 Building a HMM

A HMM describes the customer's transition among a finite set of latent states over time and the stochastic process that converts the customer's state of the world to the observed behavior. Figure XX.3 extends Figure XX.2 to a more general HMM of customer behavior.

Figure XX.3: An illustration of a general hidden Markov model



As can be seen in Figure XX.3, the customer can transition over time among the K hidden states. As discussed before, the states follow a Markovian process. However, because the researcher generally does not observe the customer's latent state, we must convert the set of latent states at time t to the set of observed behaviors using a state dependent distribution. Although not explicitly shown here, covariates can affect both the customer's likelihood of transitioning among states as well as the customer observed behaviors given a state (e.g., Netzer, Lattin and Srinivasan 2008).

It is important to note that in the context of modeling customer behavior we often assume that the customer observes all of the components in Figure XX.3. That is, the customer knows her latent state, knows the likely behavior given a state and of course

observes her actions given a state. The researcher on the other hand, observes only a sequence of observations. Hence, the hidden states, the transitions among them, the distribution of customer behavior given a state, and even the number of states (K), are parameters to be inferred or estimated from the available data.

XX.2.1 The Basic Components of a HMM[1]

We consider typical marketing data where we observe a time series of observations (e.g. choices), say $\{Y_{i1}, Y_{i2}, \dots, Y_{iT}\}$, for a set of customers ($i = 1, \dots, N$). Y_{it} may be a discrete or continuous variable, and may be univariate or multivariate. In a HMM, we assume that the probability distribution of Y_{it} depends on the realization of an unobserved, i.e. latent or hidden, discrete stochastic process S_{it}, with a finite state space $\{1, \dots, K\}$. Hence, while we observe Y_{it} directly, we can only observe S_{it} indirectly through its stochastic outcome or noisy measure Y_{it}.

In the HMM the state membership S_{it} is assumed to satisfy the Markov property such that $P(S_{it+1}|S_{it}, S_{it-1}, \dots, S_{i1}) = P(S_{it+1}|S_{it})$. That is, the state customer i is at in time period $t + 1$ only depends on what state she is at in time period t. While higher order HMMs are possible, i.e. where the conditioning extends beyond the most recent time period, the first order assumption is often made for convenience and is often sufficient to capture the dynamics in the data. It should be noted that even though the state transitions are assumed to follow a first-order Markov process, the sequence of

[1] This and the following sections build on the excellent book of Zucchini and MacDonald (2009). We adapt and extend their framework to a context typical for marketing where we have panel data. Zucchini and MacDonald (2009) mostly consider applications of HMMs for a single time series.

observations can follow any order of autocorrelation, depending on the values of the state transition probabilities.

The basic HMM for customer i transitioning among K states over T time periods can be written as (see section 2.2 for an intuitive derivation of the following equation for an example with three time periods and two states):

$$P(Y_{i1}, Y_{i2}, \dots, Y_{iT}) = \sum_{s_1, s_2, \dots, s_T} P(S_{i1} = s_1) \prod_{\tau=2}^{T} P(S_{i\tau} = s_\tau | S_{i\tau-1} = s_{\tau-1}) \prod_{\upsilon=1}^{T} P(Y_{i\upsilon} | S_{i\upsilon} = s_\upsilon) \quad \text{(XX.2)}$$

Hence, a standard HMM as presented in Equation (XX.2) consists of three main components, each of which we will discuss in more detail below:

- *The initial state distribution* $P(S_{i1} = s_1),\ s_1 = 1,2, \dots, K$, which may be represented by a $1 \times K$ row vector π.
- *The transition probabilities* $P(S_{it+1} = s_{t+1} | S_{it} = s_t)$ for $s_{t+1}, s_t = 1,2, \dots, K$, which may be represented by a $K \times K$ transition matrix Q .
- *The state-dependent distributions* of observed activity $P(Y_{it} | S_{it} = s_t)$, $s_t = 1,2, \dots, K$, which may be represented by a $K \times K$ matrix $M_{i\tau}$, that has the elements $P(Y_{it} | S_{it} = s_t)$ on the diagonal and zeros on the off-diagonal.

We refer the interested reader to Zucchini and MacDonald (2009) for further details of the modeling aspects of the HMM. Specification of these three components will be discussed next.

<u>The initial state distribution</u>

The initial state distribution describes the state membership at the beginning of the time series. Here, the researcher needs to choose how to specify the vector of initial state probabilities $\pi = \{\pi_1, \pi_2, \dots, \pi_K\}$, where π_k is the probability of the customer being in

state k at the first time period ($\pi_k = P(S_{i1} = s_1), s_1 = 1,2, \dots, K$). One possibility is to assume a-priori, based on theoretical grounds, that all customers start at one particular state. For example, in the context of a prescription of a new pharmaceutical drug, Montoya, Netzer, and Jedidi (2010) assume that all physicians, prior to the introduction of the drug, start at the lowest state of prescription behavior (i.e. $\pi = \{1, 0, \dots, 0\}$). However, this requires strong prior knowledge regarding the initial process.

Another option is to assume that the process started from its stationary distribution. In this case we would estimate π from solving the systems of equations $\pi = \pi Q$, where Q is the K × K transition probabilities matrix. This constraint is reasonable if the customer has had a long history of transactions with the firm prior to the start of the observation period (Netzer et al. 2008). A necessary condition to be able to calculate such stationary distribution is that the transition matrix is ergodic or irreducible. That is, it is possible to eventually get from every state to every other state with positive probability.

Finally, in the most general form, one could estimate π directly using a vector of $K - 1$ parameters. While this approach is most flexible, its primary drawback is that it increases the risk of local maxima, particularly when estimating the HMM using a maximum likelihood or an expectation maximization (EM) approach (Zucchini and MacDonald 2009).

The transition matrix Q

The conditional probabilities $P(S_{it+1}|S_{it})$ are called the transition probabilities and can be represented by a $K \times K$ transition matrix of conditional probabilities, Q. Each row in Q contains the conditional probabilities that the customer would be in any of the K latent

states in the next time period, given the customer's current state. Thus, each element of the matrix Q needs to be in between 0 and 1, and the row-sum of each row in Q needs to equal 1. One can represent the transition matrix Q as in Figure XX.4.

Figure XX.4: A schematic representation of the transition probability matrix Q of a HMM

State at t \ State at t+1	1	2	…	K
1	q_{11}	q_{12}	…	qq_{1K}
2	q_{21}	q_{22}	…	qq_{2K}
⋮	⋮	⋮	⋱	⋮
K	q_{K1}	q_{K2}	…	q_{KK}

In the transition matrix Q depicted in Figure XX.4, q_{11} is the conditional probability $P(S_{it+1} = 1|S_{it} = 1)$, Similarly, q_{12} is the conditional probability $P(S_{it+1} = 2|S_{it} = 1)$, and, in general, $q_{s_t s_{t+1}} = P(S_{it+1} = s_{t+1}|S_{it} = s_t)$ for $s_{t+1}, s_t = 1,2, \dots, K$.

In most applications outside of marketing the states are considered to be "states of the world" and therefore the transition matrix is not dependent on time. In such as case the HMM is a homogenous HMM with $Q_t = Q$ for $t = 1,2, \dots, T$, and Q can be represented as in Figure XX.4. In marketing, however, the states are often states of customer behavior, which could be affected by firm's actions. In such cases the transition matrix Q may depend on time and/or on time varying covariates, in which case we would write Q_t instead of Q. The resulting HMM is referred to as a non-homogeneous HMM

(e.g., Netzer et al. 2008). Furthermore, if the transition matrix depends on how long the customer has been in the state, then the HMM is referred to as a semi-HMM (e.g., Montgomery et al. 2004).

As the number of states increases, the number of transition parameters grows at a rate of approximately the square of the increase in the number states. Therefore, it is sometimes beneficial to impose restrictions on the transition matrix. For example, one could impose that transitions are allowed only among adjacent states. In such case, only q_{jj}, q_{jj-1}, and q_{jj+1} (for $j = 2, 3, \ldots K-1$) along with q_{11}, q_{12}, q_{KK-1}, and q_{KK} are estimated, and the other transition matrix elements are set to 0. Alternatively, restrictions on Q could arise from the desire to capture a particular customer behavior. For example, customer churn could be captured by an absorbing state. In order to create a HMM with an absorbing state, one would restrict in the transition matrix Q all probabilities in the row of the absorbing state to zero except the probability on the diagonal, which is set equal to one.

The state dependent distributions of observed data Y_{it} in time period t

In a HMM, given the customer's state S_{it}, the observed behavior Y_{it} is a noisy measure and a probabilistic outcome of the state. If the customer's latent state S_{it} is known, the probability distribution of Y_{it}, $P(Y_{it}|S_{it})$, only depends on the current state. Thus, the temporal dependencies across observations are only driven by the customer's state membership over time and conditional on the customer's state the conditional probabilities $P(Y_{it}|S_{it})$ are independent over time.

The state dependent distribution is probably the most flexible component of the HMM as it can be fully adapted to capture the distribution of the observed outcome Y_{it}. For example, if the observed behavior is a binary outcome one can use a binary logit or binary probit distribution (e.g., Netzer et al. 2008), for multinomial choice one can use a multinomial logit or multinomial probit (e.g., Schweidel, Bradlow and Fader 2011), for count data one can use a Poisson distribution (e.g., Ascarza and Hardie 2013), and for continuous Y_{it} one can use a normal distribution (e.g., Ebbes, Grewal and DeSarbo 2010). In cases in which multiple outcomes are observed given a state one can use any combination of the above (e.g., Ebbes et al. 2010; Zhang, Netzer and Ansari 2014, Ebbes and Netzer 2016). For example, Ebbes and Netzer (2016) consider a combination of different user behaviors on LinkedIn, consisting of activities that are discrete which are modeled as a binary logit model (e.g. the user updated her profile page or not), and activities that are continuous which they model as a tobit-regression model (e.g. how many pages did the user visit).

The state dependent distributions are often specified as a generalized linear model, with or without covariates, where the (regression) parameters are state dependent. For instance, if we have just one dependent variable which indicates a binary choice, and we have P time-varying covariates given by the $P \times 1$ vector X_{it} (including an intercept), then $P(Y_{it}|S_{it} = s_t)$ could be modeled as a binary logit model, given by

$$m_{its} = P(Y_{it}|S_{it} = s_t, X_{it}) = \frac{\exp(X_{it}'\beta_{s_t})}{1 + \exp(X_{it}'\beta_{s_t})} \qquad \text{(XX.3)}$$

The state dependent distributions differ across states according to K vectors of regression coefficients β_{s_t}, one vector for each state $s_t = 1,2, \dots, K$. We can define a

matrix M_{it} that collects the state dependent probabilities of consumer i in time t as a $K \times K$ diagonal matrix:

$$M_{it} = \begin{bmatrix} m_{it1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & m_{itK} \end{bmatrix}.$$

If the state dependent distributions differ across states not only in terms of the value of the distribution parameters but also in the distributional functional form, then the model is sometimes called a hidden Markov mixture of experts. For example, in the context of behavioral games, Ansari, Montoya and Netzer (2012) build a HMM in which one of states represents reinforcement learning and the other state represents belief learning.

Hence, the state dependent distributions in the HMM are rather modular, and depending on the behavior modeled, one can consider almost any general distribution or a mix of distributions, to capture the nature of the observed dependent variable(s).

XX.2.2 The HMM Likelihood Function

In this section we put together the three components of the HMM, namely, the initial state distribution, the transition matrix, and the state dependent distribution to form the HMM likelihood function of observing the sequence of data. To build the intuition for the likelihood function (and Equation (XX.2)), we start with a simple example, where we have two states ($K = 2$) and three time periods ($T = 3$). For customer i, we therefore observe Y_{i1}, Y_{i2}, and Y_{i3} and this customer is in (latent) states S_{i1}, S_{i2}, and S_{i3}, in periods 1,

2 and 3, respectively. The joint probability of data and latent states is given by $P(Y_{i1}, Y_{i2}, Y_{i3}, S_{i1}, S_{i2}, S_{i3})$, and can be written as follows[2]:

$$\begin{aligned} P(Y_{i1}, Y_{i2}, Y_{i3}, S_{i1}, S_{i2}, S_{i3}) &= P(Y_{i3}, S_{i3}, Y_{i2}, S_{i2}, Y_{i1}, S_{i1}) \\ &= P(Y_{i3}|S_{i3}, Y_{i2}, S_{i2}, Y_{i1}, S_{i1}) \times P(S_{i3}|Y_{i2}, S_{i2}, Y_{i1}, S_{i1}) \times P(Y_{i2}|S_{i2}, Y_{i1}, S_{i1}) \\ &\times P(S_{i2}|Y_{i1}, S_{i1}) \times P(Y_{i1}|S_{i1})P(S_{i1}) \end{aligned}$$

Here is where the Markov property together with the fact that the state dependent distributions are conditionally independent help simplifying the previous product of conditional probabilities:

I. $P(Y_{i3}|S_{i3}, Y_{i2}, S_{i2}, Y_{i1}, S_{i1}) = P(Y_{i3}|S_{i3})$ – the distribution of Y_{13} only depends on the current state S_{i3} and not on pervious states nor previous observations;

II. $P(S_{i3}|Y_{i2}, S_{i2}, Y_{i1}, S_{i1}) = P(S_{i3}|S_{i2})$ – the state membership in $t = 3$ *only* depends on the customer's previous state S_{i2} (the Markov property).

III. $P(Y_{i2}|S_{i2}, Y_{i1}, S_{i1}) = P(Y_{i2}|S_{i2})$ – following the same rational as I;

IV. And, $P(S_{i2}|Y_{i1}, S_{i1}) = P(S_{i2}|S_{i1})$ – following the same rational as II.

Hence, the likelihood of observing the set of observations and states can be more succinctly written as:

$$\begin{aligned} &P(Y_{i1}, Y_{i2}, Y_{i3}, S_{i1}, S_{i2}, S_{i3}) \\ &= P(S_{i1})P(Y_{i1}|S_{i1})P(S_{i2}|S_{i1})P(Y_{i2}|S_{i2})P(S_{i3}|S_{i2})\ P(Y_{i3}|S_{i3}) \end{aligned} \quad \text{(XX.4)}$$

However, in practice, we do not observe the customer states. That is, we observe the customer's activity (Y_{i1}, Y_{i2}, and Y_{i3}) but not the customer's state in each time period (S_{i1}, S_{i2}, and S_{i3}). Thus, to obtain the likelihood for the observed data, we need to

[2] Here we use a general product rule to calculate the probability of the joint distribution using conditional probabilities. Under the general product rule the joint distribution of four 'events' (B_1, B_2, B_3, B_4) can be written as the product of conditional distributions as follows: $P(B_1, B_2, B_3, B_4) = P(B_1|B_2, B_3, B_4)P(B_2|B_3, B_4)P(B_3|B_4)P(B_4)$.

'integrate out' the latent states, across all state paths that the customer could take over time:

$$P(Y_{i1}, Y_{i2}, Y_{i3}) = \sum_{s_1=1}^{2} \sum_{s_2=1}^{2} \sum_{s_3=1}^{2} P(Y_{i1}, Y_{i2}, Y_{i3}, S_{i1} = s_1, S_{i2} = s_2, S_{i3} = s_3) =$$

$$\sum_{s_1,s_2,s_3} P(S_{i1} = s_1) \times P(Y_{i1}|S_{i1} = s_1) \times P(S_{i2} = s_2|S_{i1} = s_1) \times P(Y_{i2}|S_{i2} = s_2) \times P(S_{i3} = s_3|S_{i2} = s_2)$$

$$\times P(Y_{i3}|S_{i3} = s_3) =$$

$$\sum_{s_1,s_2,s_3} P(S_{i1} = s_1) \times P(S_{i2} = s_2|S_{i1} = s_1) \times P(S_{i3} = s_3|S_{i2} = s_2) \times P(Y_{i1}|S_{i1} = s_1) \times P(Y_{i2}|S_{i2} = s_2)$$

$$\times P(Y_{i3}|S_{i3} = s_3) =$$

$$\sum_{s_1,s_2,s_3} P(S_{i1} = s_1) \prod_{\tau=2}^{3} P(S_{i\tau} = s_\tau|S_{i\tau-1} = s_{\tau-1}) \prod_{v=1}^{3} P(Y_{iv}|S_{iv} = s_v),$$

where $s_\tau = 1$ or 2 for $\tau = 1,2,3$. One limitation with the likelihood function as presented here, is that the summation over all possible states' paths that the customer could take, involves K^T terms in the summation, which can create computational burden when the number of time periods and state increase (see also Equation (XX.2)). Zucchini and Macdonald (2009, p. 37) show that the HMM likelihood function can be written in a more convenient matrix form instead. Extending the simple example to a more general case with K states and T time periods, and using matrix notation, we can write the HMM likelihood function for customer i as:

$$L_{iT} = P(Y_{i1}, Y_{i2}, \dots, Y_{iT}) = \pi M_{i1} Q M_{i2} \dots Q M_{iT} \iota \qquad \text{(XX.5)}$$

where π, M_{it}, and Q are defined as above and ι is a $K \times 1$ vector of ones. The likelihood function (XX.5) also provides the intuition for the HMM process. The process starts with customer i belonging to a particular latent state k, which follows the initial state distribution π. Given her state in period 1 the customer behaves in particular manner, as described by the probabilities M_{i1}. Next, the customer may transition from her state at

time period 1 to her state at time period 2 (as described by the transition probabilities Q). Subsequently, given the state the customer transitioned to in period 2 (which may be her current state), M_{i2} captures customer behavior in period 2, followed by another state transition between period 2 and period 3 according to the probabilities in the transition matrix Q. This process repeats itself until we reach the final behavior of customer i in time period T.

The likelihood for the complete sample of customers $i = 1,2, \dots, N$ is given by the following product: $L_T = \prod_{i=1}^{N} L_{iT}$. In Section XX.3, we discuss several approaches to estimate the HMM parameters after observing the data.

The forward and backward probabilities

For the purpose of state recovery, prediction, and estimation, it is useful to split the likelihood function in (XX.5) into forward and backward components.

Let the $1 \times K$ row vector α_{it} be defined as follows: $\alpha_{it} = \pi M_{i1} \prod_{s=2}^{t} Q M_{is}$. Thus, we can rewrite the likelihood function up to time T as $L_{iT} = \alpha_{iT}\iota$, which can be obtained recursively as $\alpha_{it} = \alpha_{it-1} Q M_{it}$ $(t \geq 2)$ with, for $t = 1$, $\alpha_{i1} = \pi M_{i1}$. The row vector α_{it} is called the vector of *forward* probabilities. Furthermore, it can be shown (e.g. Zucchini and MacDonald, 2009, p. 60) that the j-th element of α_{it}, say $\alpha_{it}(j)$, is the joint probability $P(Y_{i1}, Y_{i2}, \dots, Y_{it}, S_{it} = j)$.

Similarly, one can define a $1 \times K$ vector of *backward* probabilities β_{it}. This vector captures the last $T - t$ terms of the HMM likelihood recursion, that is $\beta'_{it} = (\prod_{s=t+1}^{T} Q M_{is})\iota$, for $t = 1,2, \dots, T$, with $\beta'_{iT} = \iota$. It can be shown (e.g. Zucchini and MacDonald p. 61) that the j-th element of this vector, say $\beta_{it}(j)$, is the conditional

probability $P(Y_{it+1}, Y_{it+2}, \dots, Y_{iT} | S_{it} = j)$. This is, the probability of observing $Y_{it+1}, Y_{it+2}, \dots, Y_{iT}$ given that customer i is in state j in time period t.

In fact, the forward and backward probabilities can be combined to give the joint probability $P(Y_{i1}, Y_{i2}, \dots, Y_{iT}, S_{it} = j)$ as the product of the two, i.e. $\alpha_{it}(j)\beta_{it}(j)$. Then, $P(Y_{i1}, Y_{i2}, \dots, Y_{iT}) = \sum_{j=1}^{K} P(Y_{i1}, Y_{i2}, \dots, Y_{iT}, S_{it} = j) = \sum_{j=1}^{K} \alpha_{it}(j)\beta_{it}(j) = \alpha_{it}\beta_{it}'$. Hence, another way to compute the likelihood L_{iT} is through any of the $t = 1,2, \dots, T$ combinations $\alpha_{it}\beta_{it}'$. The likelihood function given in (XX.5) is a special case of the product of the forward-backward probabilities for $t = T$, where we only need the forward probabilities (as $\beta_{iT}' = \iota$).

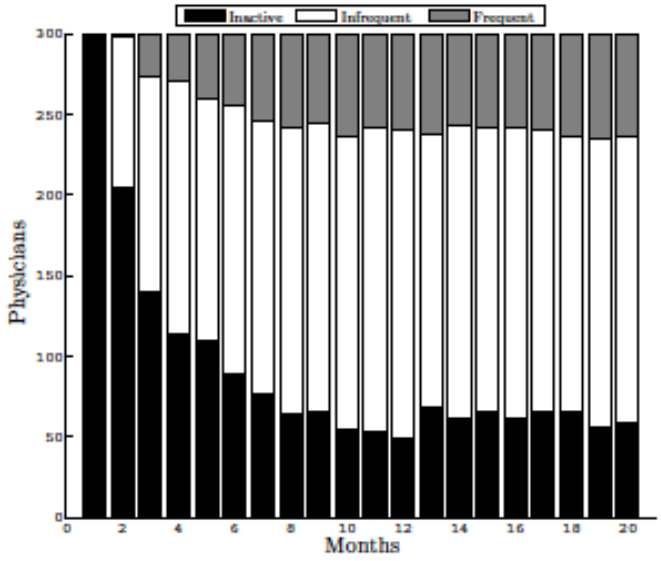
When the time series is long the calculation of the forward and backward probabilities can suffer from underflow. Zucchini and MacDonald (2009, Section 3.2) discuss appropriate scaling of these probabilities to avoid underflow.

XX.2.3 HMM State Recovery and Prediction

In some cases HMMs are primarily used as a predictive model with the objective of predicting customer behavior (Y_{it}) in future time period $t = T + h$, $h = 1,2, \dots$., H. One advantage of using HMMs for that purpose is that it is easy to predict a few periods ahead. For example, Paas, Vermunt and Bijmolt (2007) present a HMM for household ownership of financial products and use the HMM to predict future acquisitions of such products. In other cases the primary objective of the HMM is to recover the customer's state (S_{it}) at each time period. For example, Ebbes and Netzer (2016) use a HMM and observations on users' activity on LinkedIn with the primary objective of inferring which users are in a state of a job search. State recovery can also be used to capture how the

firm's customer base has evolved over time. Figure XX.5 depicts such an example from Montoya et al. (2010). Following the introduction of a new drug, and marketing efforts by the firm, the physicians' base has transitioned from the inactive prescription state prior to the introduction of the drug to a majority of the physicians in an infrequent prescription state, and approximately 20% of the physicians in a frequent prescription state. Note that it took the physicians' base approximately eight months post the introduction of the drug to stabilize on the prescription state membership.

Figure XX.5: An example of customer state membership evolution from Montoya et al. (2010).



Both predictions and state recovery are closely related to the HMM likelihood function and forward/backward probabilities described in Section XX.2.2.

Recovering state membership

Two approaches have been suggested for recovering the state membership distribution: filtering and smoothing. Filtering utilizes only the information known up to time *t* to recover the individual's state at time *t*, while smoothing utilizes the full information available in the data to predict the customer state at any point in time during the observed data period. The smoothing approach is quite common in fields such as speech recognition where one wants to infer the meaning of a particular word both by words that appeared prior to the focal word and words that appeared after the focal word. In most marketing applications, the researcher is more interested to infer a customer state only based on the history of the observed behavior and not based on future behavior and hence the filtering approach is more common.

The smoothing state membership probabilities can be computed using the Bayes formula:

$$P(S_{it} = j|Y_{i1}, Y_{i2}, \dots, Y_{iT}) = \frac{P(Y_{i1}, Y_{i2}, \dots, Y_{iT}, S_{it} = j)}{P(Y_{i1}, Y_{i2}, \dots, Y_{iT})},$$

which can be further simplified using the forward and backward probabilities discussed in the previous section as follows:

$$P(S_{it} = j|Y_{i1}, Y_{i2}, \dots, Y_{iT}) = \frac{\alpha_{it}(j)\beta_{it}(j)}{L_{iT}}. \quad \text{(XX.6)}$$

Similarly, the filtering probabilities can be written as:

$$P(S_{it} = j|Y_{i1}, Y_{i2}, \dots, Y_{it}) = \frac{P(Y_{i1}, Y_{i2}, \dots, Y_{it}, S_{it} = j)}{P(Y_{i1}, Y_{i2}, \dots, Y_{it})},$$

which can be computed using the forward probabilities as:

$$P(S_{it} = j|Y_{i1}, Y_{i2}, \dots, Y_{it}) = \frac{\alpha_{it}(j)}{L_{it}}. \quad \text{(XX.7)}$$

Another approach related to recovering state membership attempts to recover the most probable sequence of states given the full information in the data. Unlike smoothing, which predicts state membership at one point in time, this approach decodes the best hidden state path given a sequence of observations. In principle the most likely path could be discovered by running the forward algorithm for each possible sequence of states, and then find the path which corresponds to the highest probability. Clearly, this would easily become impossible given the potentially large number of state sequences. Instead, for this task one could use the Viterbi algorithm which is a recursive algorithm (leveraging the forward and backward probabilities algorithm) akin to dynamic programing algorithms (e.g. Viterbi 1967; Jurafsky and Martin 2008). If the main purpose of the analysis is to recover and interpret the sequence of state membership, it is recommended to test the accuracy of the Viterbi algorithm using simulation (see e.g., Zucchini and MacDonald (2009) pp. 84-86). For marketing applications, one could potentially compute one such sequence for each customer. For instance, in the context of the preference example for Brands A and B discussed above, the Viterbi algorithm would allow a manager to infer the most probably sequence of preference states that a customer took during the observation window.

Predicting future activity

In some applications of HMMs, the researcher is interested in predicting future values of the observed variable Y_{it}. Hence, we want to compute the probability of observing customer i's activity in the time period $T + h$, $h > 0$, given the activity we have observed until time T. This probability is derived from Bayes theorem, i.e.

$$P(Y_{iT+h}|Y_{i1}, Y_{i2}, \dots, Y_{iT}) = \frac{P(Y_{i1}, Y_{i2}, \dots, Y_{iT}, Y_{iT+h})}{P(Y_{i1}, Y_{i2}, \dots, Y_{iT})}.$$

The denominator of the above equation is simply L_{iT} given in (XX.5). The numerator can be computed by multiplying the customer's forward probabilities by *h* transition matrices and by the customer's state dependent distribution in period $T + h$ (see also Zucchini and MacDonald, 2009, pp.33 and 37). That is,

$$P(Y_{i1}, Y_{i2}, \dots, Y_{iT}, Y_{iT+h}) = \pi M_{i1} Q M_{i2} \dots Q M_{iT} Q Q \dots Q M_{iT+h} \iota = \alpha_{iT} Q^h m_{iT+h} \iota$$

The predicted customer behavior in period $T + h$ can be written as:

$$P(Y_{iT+h} | Y_{i1}, Y_{i2}, \dots, Y_{iT}) = \frac{\pi M_{i1} Q M_{i2} \dots Q M_{iT} Q^h M_{iT+h} \iota}{\pi M_{i1} Q M_{i2} \dots Q M_{iT} \iota} = \frac{\alpha_{iT} Q^h M_{iT+h} \iota}{\alpha_{iT} \iota}. \qquad \text{(XX.8)}$$

This expression can be obtained as a by-product in likelihood estimation.

XX.2.4 Accounting for Cross-customer Heterogeneity

One of the aspects that differentiates HMMs in marketing from HMMs in other fields is that in other fields there is usually one single long sequence of observations (e.g. pixels in an image or the GDP in the united states over the past few decades) and the HMM is estimated for a single time series using the entire sequence of observations. In marketing, on the other hand, we often have a panel data structure in which we observe multiple sequences of observations, one for each customer, allowing for different customers to possibly have heterogeneous preferences and behaviors.

The idea that customers are different in terms of preferences and behavior has a long history in marketing. Modeling consumer heterogeneity has been the central focus of many econometric marketing applications (see chapter XX). Voluminous research has demonstrated the bias that may arise from not accounting for heterogeneity across customers. Moreover, in many marketing applications the researcher is interested in

targeting individual consumers based on their individual preferences, which would require a detailed understanding of customer heterogeneity.

Accounting for cross-customer heterogeneity is even more important in the context of dynamic models, such as the HMM. Heckman (1981) demonstrates that estimating a random utility homogenous choice model based on a heterogeneous sample may lead to a strong spurious state dependence, even when the actual choices were independent over time. Similarly, a model that accounts for heterogeneity but ignores state dependence may overestimate the degree of heterogeneity (Keane 1997). In the context of HMMs, not accounting for cross-customer heterogeneity forces the states to capture both heterogeneity (similar to latent class model) and dynamics. To see this, imagine a group of customers who are static in their preferences (a HMM estimated for these customers should lead to an identity transition matrix as customers do not switch states over time) and a second group of customers who have a 60% chance of staying in their previous preference state and 40% likelihood of transitioning to another preference state at each time period. A homogenous HMM estimated using data from both groups of customers would lead to a single transition matrix, that is an "average" of an identity matrix and a matrix reflecting the switching behavior of the second group. Consequently, the estimated transition matrix would suggest that *all* customers are dynamic in their preferences (including the first group with static preferences), whereas at the same time, the estimate transition matrix overstates the stickiness of the states (i.e. the likelihood of staying in the same state from one period to another) for the second, dynamic group of customers.

With long enough time series per customer, in principle, one could estimate a separate HMM for each customer. This would give a unique set of estimated parameters for each customer. That is, each customer would have its own estimated initial state probabilities (π_i), transition probability matrix (Q_i), and parameters of the state-dependent activity distribution (M_{it}). Because typically the number of observations per customer are insufficient to estimate a unique HMM for each customer, data can be pooled across customers by including customer-level heterogeneity in the HMM. For instance, random-effect parameters can be included in a HMM and estimated using either a hierarchal Bayes MCMC estimation or a simulated maximum likelihood approach (Train 2009). Alternatively, the heterogeneity across customers can be captured using a latent class approach (Kamakura and Russell 1989). One could also account for observed heterogeneity by including covariates, such as demographics in the model. However, because observed heterogeneity covariates often capture only limited degree of heterogeneity, we recommend controlling for unobserved heterogeneity as well.

In the most general case one could allow of cross-customer heterogeneity in each of the three HMM components: initial state distributions (π_i), transition matrix (Q_i), and state-dependent distribution (M_{it}). Capturing heterogeneity in π_i and Q_i but not in M_{it} allows different customers to have different level of stickiness to the states but assume that, given a state, all customers have the same preference structure, exhibit similar behavior, or respond in a similar manner to marketing actions. The attractiveness of such an approach is that the interpretation of the states becomes easier, because the states now mean the same thing for all customers. On the other hand, allowing for heterogeneity in the state-dependent distribution (M_{it}), implies that what a "high state" is for one

customer may be very different from what a "high state" is for another customer. However, a limitation of not accounting for heterogeneity in the state-dependent distribution is that if the behavior given a state is highly heterogeneous and has a wide support (e.g., food expenditure which can vary substantially among customers given their household size and income), not accounting for heterogeneity in the state-dependent distribution could lead to confusion between heterogeneity and dynamics, as some states will capture heterogeneity in addition to dynamics.

Finally, to the best of our knowledge, all HMMs in marketing (and in other fields) assumed the same number states for all customers (even if heterogeneity in the model parameters is allowed). Failure to account for heterogeneity in the number of states leads to a mis-specified model for customers for whom the number of states does not match their dynamic behavior. Recent work by Padilla, Montoya and Netzer (2016) attempts to relax this assumption and allows for heterogeneity in the number of states across customers.

In sum, when one estimates a HMM for a heterogeneous set of consumers, we encourage researchers to carefully account for unobserved heterogeneity in order to disentangle heterogeneity from dynamics. It is almost always advisable to allow for heterogeneity in the transition matrix (Q_i) and the initial state distribution (π_i) and wherever possible or needed also in the state-dependent distributions (M_{it}).

XX.2.5 Non-homogenous HMMs[3]: Time-varying Covariates in the Transition Matrix

[3] In the context of HMMs the convention is to call a non-homogenous HMM a HMM with a time variant transition matrix (Q_t). This is not be confused with a heterogeneous HMM, in which the transition matrix,

In most non-marketing applications the states of the HMM are exogenous states of the world. Accordingly, the transition matrix in these applications is rarely a function of covariates. However, in marketing, because the states of the HMM are often customer behavior states, the firm may believe that it can affect customers' transitions among states. Therefore, marketing applications of HMMs often allow the transition matrix to be a function of covariates such as marketing actions. Indeed, until the diffusion of HMMs to marketing, HMMs rarely incorporated time-varying covariates in the transition matrix (see Hughes and Guttorp 1994 for an exception). Early work on HMMs in marketing (e.g., Paas et al. 2007; Netzer et al. 2008; Montoya et al 2010) proposed non-homogenous HMMs in which the transitions among the state were a function of customer activities or marketing actions. In these cases the transition probabilities in Q_i are both customer and time specific, which can be modeled by standard (or ordered) logit models. For instance, $P(S_{it} = s_t | S_{it-1} = s_{t-1}) = f(Z_{it})$ where $f(.)$ is the logit function and Z_{it} is a vector of covariates that are specific to customer i and time period t, and $s_t, s_{t-1} = 1,2, \dots, K$. Now, the elements of the transition matrix are a function of time and customer, and we write Q_{it}.

As discussed earlier, one could also add covariates in state dependent distributions. These covariates would affect the customer behavior, conditional on the customer's state. The choice of which covariates should go in the transition matrix and which should go in the state-dependent distribution is a researcher decision. In general, covariates that are included in the transition matrix should be covariates that are postulated to have a long-term effect on the customer's behavior. The rational is that

and possibly other model parameters can vary across consumers (Q_i) and from a non-stationary HMM in which the state transition are a function of time itself.

these covariates create a regime shift in the customer behavior by transitioning the customer to a different and often sticky state of customer behavior. Covariates that are included in the state-dependent distribution, by definition affect the customer behavior only in the current time period, conditional on the customer state, and therefore have a short-term effect. In the context of pharmaceutical drugs prescriptions by physicians, Montoya et al. (2010) demonstrate that including detailing and sampling to physicians covariates in both in the transition probabilities and the state dependent distribution can capture both the short- and long-term effects of these marketing activities.

XX.2.6 Selecting the Number of States and Model Selection

The first order of business in estimating a HMM is to select the number of hidden states (K). The number of states could either be estimated from the data or defined based on theoretical grounds. If the researcher has a strong theoretical basis with respect to the number and the interpretation of each of the states, then the researchers could determine the number of states a-priori. For example, Ansari et al. (2012) choose a-priori two states which correspond to reinforcement and belief learning over repeated rounds of behavioral games.

A more common approach is to use model selection procedures to choose the number of states based on the fit of the model to the data. The approach involves estimating a range of models with increasing number of states K until the point at which adding an additional state does not further improve or leads to a worse model selection criterion value. Increasing the number of hidden states adds flexibility and parameters to the model and, hence, will always improve model fit as measured by the likelihood.

However, as the number of model parameters increases, the key issue is whether the improvement in model fit is large enough relative to the increase in the number of parameters. Accordingly, one often uses panelized model selection fit measures, such as information criteria, which balance model fit and model parsimony.

Information criteria add a penalty to the model fit (-2×loglikelihood) on the basis of the number of parameters g. A typical and fully specified HMM with no covariates has $K-1$ parameters in π, $K \times (K-1)$ parameters in Q, and K parameters in M, leading to $g = K \times (K+1)$ parameters. The Akaike Information criterion (AIC) equals: $-2 \times loglikelihood + 2 \times g$, the Bayesian Information Criterion (BIC) equals $-2 \times loglikelihood + g \times ln(n)$, and the Consistent Akaike Information criterion (CAIC) equals: $-2 \times loglikelihood + g \times (ln(n)+1)$, where n denotes the sample size (which for the case of panel data equals to $N \times T$, where $i = 1,2,\dots,N$ is the number of customers and $t = 1,2,\dots,T$ is the number of time periods per customer). The choice among alternative model specifications can be made by selecting the model with the minimum value of a specific information criterion.

For reasonable sample sizes, the penalty per additional parameter is typically much larger for BIC and CAIC than for AIC. Accordingly, the AIC tends to favor models with many, oftentimes too many, states. Accordingly, the BIC is commonly the preferred criterion to determine the number of states (Bartolucci, Farcomeni and Pennoni, 2014).

When one estimates the model based on Bayesian estimation procedures, typical Bayesian model selection criteria such as the Log Marginal Density and the Bayes Factor are often used. These criteria could be calculated from the output of the MCMC procedure (see Chib, 1995, 2001 for details). It has been shown that the BIC measure in

classical estimation asymptotically approximates the Log Marginal Density (Congdon 2002 p. 473). Alternatively, because the Log Marginal Density and the Bayes Factor, sometimes recommend non-parsimonious models, researchers have used a modified Deviance Information Criterion (Celeux et al. 2006), cross validation approaches, and posterior predictive checks for model selection. Another advantage of these model selection criteria is that they do not require the calculation of g (the number of parameters), as for e.g. AIC or BIC, which is often cumbersome in particular if the researcher accounts for cross-customer heterogeneity through random coefficients.

Several studies have proposed model selection criteria that are specific for HMM estimation. Bacci, Pandolfi, and Pennoni (2014) propose a classification-based or entropy-based criterion, which examines the posterior probabilities of state membership of each of the customers. The idea behind these measures is that if the states of the HMM are well-separated, the posterior probabilities of state membership are close to one, resulting in an entropy that is close to zero. They find that most decision criteria tend to work reasonably well, and their performance improves if the sample size or the number of time periods increases. They find that when the number of states is large, BIC, and the classification-based criteria tend to underestimate the correct number of states.

Smith, Naik and Tsai (2006) build on the Kullback–Leibler (KL) divergence criterion and propose a Markov Switching Criterion (MSC), which is specifically suited for states selection in Markov and latent Markov models. Using simulations, they find that the MSC preforms well in term of retaining the correct number of states and unlike measures such as the AIC avoids overstating the true number of states. We encourage future research to explore the use of the reversible jump algorithm

(e.g., Green, 1995, Ebbes, Liechty and Grewal, 2015) to simultaneously estimate the HMM with varying number of states and select the best fitting model.

Similar model selection criteria to the ones described in this section can be used to select among different model specifications other than selecting the number of latent states K, such as whether and which covariates to include in the transition probabilities or in initial state distribution.

XX.3 Estimating a HMM

As discussed above, a HMM has three main components leading to three sets of parameters to be estimated: (1) the initial state probabilities π_i, (2) the transition probability matrix Q_i, and (3) parameters of the state dependent distributions M_{it}. In this section we retain the subscript i for π and Q implicitly assuming that we would like to control for customer-specific heterogeneity, either by estimating a separate HMM for each customer, or by estimating one HMM by pooling across customers while including customer-specific unobserved and/or observed heterogeneity through covariates.

Three main approaches have been proposed to estimate the model parameters of a HMM: (1) maximum likelihood estimation by the Expectations Maximization (EM) algorithm, (2) Maximum likelihood estimation by directly optimizing the likelihood function, and (3) Bayesian estimation. We will briefly discuss each approach in turn, focusing on the essentials, and provide references for further details of the implementations. We note that several software packages are available to estimate basic HMMs (e.g., R-HMM in CRAN, Latent GOLD),

XX.3.1 The Expectation Maximization (EM) Algorithm

A popular way to estimate a HMM is through the EM algorithm, also known as the Baum-Welch forward-backward algorithm (Baum et al. 1970; Baum 1972; Dempster et al. 1977; Welch 2003). The main idea behind the EM algorithm is to treat the state memberships, which are unobserved, as missing data. The algorithm then iteratively finds the parameters that maximize the likelihood function by an E step and an M step. The E step is designed to obtain the conditional expectations of the missing data (here, the state memberships). Then, in the M step, the complete data log likelihood is maximized. The complete data now comprises the observed data and the conditional expectations of the missing data. Generally, the complete data log-likelihood function can be easily maximized, often much more straightforwardly then the (log) likelihood function of only the observed data.

To derive the EM algorithm for HMMs, we start with the complete data likelihood function. Extending the three-time periods and two states example used in Section XX.2.2 to motivate the construction of the likelihood function, we can write the complete data log-likelihood function of observing the customer states and the customer behavior at each time period t, $t = 1,2, \dots, T$, as:[4]

$$\log P(y_1, y_2, \dots, y_T, s_1, s_2, \dots, s_T) = \log(\pi_{s_1}) + \sum_{t=2}^{T} \log(q_{s_{t-1}s_t}) + \sum_{t=1}^{T} \log(m_{ts_t}), \quad \text{(XX.9)}$$

[4] For ease of exposition we drop in the description of the EM algorithm the subscript i for customer. Estimating a HMM with heterogeneous parameters across customers using the EM algorithm is challenging, as it would involve integrating out (in the M step) the unobserved heterogeneity.

where, $y_1, y_2, \dots, y_T, s_1, s_2, \dots, s_T$ are the realizations of the customer's state and activities at each time period. π_{s_1} is the s_1-th element of the initial state distribution vector π, which corresponds to the state the customer is at in time period 1. Similarly, $q_{s_{t-1}s_t}$ is the element from the transition matrix Q that corresponds to the customer probability of transitioning from her state at time $t-1$ to her state at time t, and m_{ts_t} is the s_t-diagonal element from the matrix M_t that corresponds to the customer's state dependent distribution given the customer state at time t (s_t).

To implement the EM algorithm, it would be more convenient to represent the state assignments by the $K \times 1$ dummy vector $v_t = (v_{t1}, v_{t2}, \dots, v_{tK})$ where $v_{tj} = 1$ if $s_t = j$, and 0 otherwise, and the $K \times K$ dummy matrix W_t, with elements $w_{tij} = 1$ if $s_{t-1} = i$ and $s_t = j$, and 0 otherwise. We can now rewrite the complete data log likelihood in (XX.9) as:

$$\log P(y_1, y_2, \dots y_T, s_1, s_2, \dots, s_T) = v_1'\tilde{\pi} + \sum_{t=2}^{T} \iota'\left(W_t \circ \tilde{Q}\right)\iota + \sum_{t=1}^{T} v_t'\tilde{m}_t, \qquad \text{(XX.10)}$$

where, $\tilde{\pi}$ is a $K \times 1$ vector such that, $\tilde{\pi} = \log(\pi)$, $\tilde{Q}$ is a $K \times K$ matrix defined by $\log Q$, $\circ$ denotes the Hadamard matrix product, and $\tilde{m}_t$ is a $K \times 1$ vector with the log of the diagonal elements of M_t.

If one observes panel data structure with multiple observations per person, the total sample complete data log likelihood (SCDLL), ignoring unobserved heterogeneity across customers, would be the sum of (XX.10) across all customers $i = 1,2, \dots, N$, i.e.

$$SCDLL = \sum_{i=1}^{N} \log P(y_{i1}, y_{i2}, \dots, y_{iT}, s_{i1}, s_{i2}, \dots, s_{iT}).$$

From the previous expression in Equation (XX.10) it can be seen that the complete data log likelihood has three additive terms: a term involving the initial states, a term involving the transitions, and a term involving the state dependent distributions. Therefore, maximizing this function boils down to maximizing each of these terms separately. For the first two terms involving the initial state and transition probabilities, it is possible to obtain closed-form expressions. For the last term, closed-form expressions exist for many common specifications of the state-dependent distribution (e.g., a normal distribution), otherwise numerical maximization will be necessary.

In the E step, the quantities v_t and W_t are 'estimated' by their conditional expectations, given the observed data and the current parameter estimates, using the forward and backward probabilities, see e.g. Zucchini and MacDonald (2009; p. 65):

$$\hat{v}_t(j) = P(S_t = j|y_1, y_2, \dots, y_T) = \alpha_t(j)\beta_t(j)/L_T$$

and

$$\widehat{W}_t(j,k) = P(S_{t-1} = j, S_t = k|y_1, y_2, \dots, y_T) = \alpha_{t-1}(j)Q_i(j,k)P(y_t|k)\beta_t(k)/L_T$$

for $j, k = 1, \dots, K$.

The intuition behind the estimates of $\hat{v}_t(j)$ and $\widehat{W}_t(j,k)$ is that $\hat{v}_t$ is the likelihood that the customer visits each state and $\widehat{W}_t(j,k)$ is the customer's likelihood of transitioning from state j to state k.

Then, in the M step of the EM algorithm the complete data log likelihood is maximized after replacing v_t and W_t by their updated quantities $\hat{v}_t$ and $\widehat{W}_t$, which gives a set of updated parameter estimates. The E and M steps are repeated sequentially until the change in the estimated parameter values or the likelihood function does not further improve beyond some threshold value.

The EM algorithm can suffer from local maxima. Additionally, the calculation of the forward and backward probabilities can suffer from under flow. We briefly discuss these challenges in the next subsection (further details are provided in Section 3.2 in Zucchini and MacDonald, 2009).

XX.3.2 Directly Maximizing the Likelihood Function

In Section XX.2 we derived the likelihood function for the general HMM for a sample of N customers and T time periods. The sample likelihood is given by $L_T = \prod_{i=1}^{N} L_{iT}$ and can be computed recursively using the forward probabilities α_{it}. Rather than using the EM algorithm discussed in the previous section, the likelihood function can be maximized directly using numerical optimization routines in order to estimate the HMM parameters. The main obstacles are under and overflow challenges in computing the likelihood, constraining the probabilities such that they sum up to one and are all non-negative, and the risk of local maxima. Similar to the forward and backward probabilities discussed earlier, the customer log likelihood (e.g. Equation (XX.5)) is comprised of multiplications of probabilities over time and states, leading to a risk of underflow. For details of the likelihood scaling, we refer the reader to Zucchini and MacDonald (2009, Section 3.2).

Both the initial state probabilities and the transition matrix parameters are probabilities. Thus, each one of these parameters needs to be between 0 and 1, and the vector of initial probabilities and each row of the transition matrix needs to sum to one. This can be achieved by running a constraint optimization, or by optimizing the likelihood not in the actual parameters (e.g. $\pi_1, \pi_2, \dots, \pi_K$) but in transformed parameters

(say) $\lambda_1, \lambda_2, \ldots, \lambda_{K-1}$. We need one parameter less than the number of states K, as the sum of the probabilities is 1. Now, the actual parameters are parameterized as:

$$\pi_j = \frac{\exp(\lambda_j)}{1+\sum_{j=1}^{K-1}\exp(\lambda_j)},$$

for $j = 1,2, \ldots, K-1$ and $\pi_K = 1 - \sum_{j=1}^{K-1}\pi_j$ (implicitly we set $\lambda_K = 0$). Similarly, this reparametrization may be done for each row in the transition probability matrix Q.

As with many numerical optimization problems, the likelihood function of the HMM is often multimodal and therefore the optimization procedure can get stuck in a local maxima instead of the desirable global maximum point. The problem of a multimodal likelihood function and the risk of local maxima is higher when one estimates the initial state probabilities rather than fixing these a-priori, or assuming these to be at the stationary distribution (see Section XX.2.1). Unfortunately, there is no simple approach to guarantee a global maximum, when applying numerical optimization for multimodal distributions. We advise researchers to use theory and judgment in selecting the initial starting values and explore a wide range of starting values, and if different starting value result in different maxima, select the maximum likelihood solution that leads to the highest value of log-likelihood.

One of the limitations of classical likelihood optimization (either through the EM algorithm or by directly optimizing the likelihood function) is that it is not obvious how to incorporate in these methods random-effect parameters to capture cross-customer unobserved heterogeneity. One could still use these methods to capture unobserved heterogeneity using the latent class approach and/or include covariates to control for observed heterogeneity. The two Bayesian estimation approaches we discuss next allow

for a more natural incorporation and estimation of cross-customer unobserved heterogeneity.

XX.3.3 Bayesian Estimation

HMMs may also be estimated in a Bayesian framework. For details on Bayesian statistics we refer the reader to chapter XXX in this handbook. We take a pragmatic point of view whether one should consider estimating a HMM in a Bayesian framework or in a classical statistics framework using the EM algorithm or a direct maximum likelihood approach. While the EM algorithm and the maximum likelihood approach are often easier to implement and require considerable less computational time, the Bayesian approach is less susceptible to a local maxima problem. More importantly, if one wishes to estimate cross-customer heterogeneity, which we highly recommend when estimating HMMs across multiple customers (see Section XX.2.4), the Bayesian approach seems the natural way to estimate the model parameters.

We discuss two Bayesian approaches to estimate a HMM in a Bayesian framework, both of which are based on Markov Chain Monte Carlo (MCMC) estimation: (1) a direct approach using the complete likelihood through a Metropolis-Hastings step, and (2) a data-augmentation approach by treating the unobserved states as missing data. In both approaches one needs to address label switching, which we briefly discuss at the end of this section.

XX.3.3.1 Sampling the posterior distribution using the Metropolis-Hastings (MH) algorithm

This approach uses a standard hierarchical Bayes estimation procedure, where we distinguish between two main sets of parameters: random effect parameters (say θ_i) that are particular to each customer i (Section XX.2.4) and parameters (say ψ) that are common to all customers. Heterogeneity is introduced in the model by assuming a priori a distribution for the random effects parameters (e.g. $\theta_i \sim MVN(\bar{\theta}, \Omega)$). As is common in most marketing applications, the Bayesian model specification is completed by assuming standard diffuse conjugate priors for all model parameters. Then, the MCMC algorithm is operationalized by sequentially drawing from a set of full conditional posterior distributions. Because the full conditional distribution constructed from the HMM likelihood function (e.g. Equation (XX.5) with cross-customer heterogeneity) combined with the priors does not have a closed form, an acceptance-rejection MH step is needed to estimate the parameters ψ and possibly θ_i. At each iteration of the MCMC algorithm, we would draw a candidate value for the parameters from a proposal distribution, which then is accepted with a certain probability. If it is accepted, then the likelihood function is updated with new parameters. If it is not accepted, then the current value for the parameters is retained. After running the algorithm for a long time, we end up with a sequential sample from the posterior distribution of the parameters of the HMM. Because the sequentially generated draws from the posterior distribution can be highly correlated, we found the adaptive MH algorithm as described in Atchadé and Rosenthal (2005) to be quite useful in reducing the autocorrelation and achieving convergence faster. The Atchadé and Rosenthal (2005) algorithm automatically adjusts the tuning parameter (the variance of the proposal density) of the MH algorithm. We refer to reader to chapter XXX of this handbook for more general information about the MH algorithm. Further

details on the estimation of HMMs using the MH approach can also be found in Appendix A of Netzer et al. (2008).

XX.3.3.2 Sampling the posterior distribution using Gibbs sampling and data augmentation

Similar to the EM algorithm to maximize the likelihood function (Subsection XX.3.1), this second Bayesian approach uses data augmentation by treating the unobserved states as missing data. In each step of the MCMC algorithm, one draws the state the customer is at, at each time period, given the current set of parameters estimates and observed data. Then, by principle of MCMC estimation, conditional on the customer's state, we need only to sample the parameters of the distribution of the state-dependent behavior, which is often rather straightforward to do using standard Gibbs sampling. This approach was proposed by Scott (2002) and was implemented in marketing by several papers (e.g. Ebbes et al., 2010, Ascarza and Hardie, 2013).

One of the limitations of the data augmentation approach is that the sampler can "get stuck" in a sticky state, where the customer is continuously being drawn to be in the same state. Früthwirth-Schnatter (2006; Sec. 11.5) provides a very useful summary of various algorithms to mitigate such issues. While these algorithms are more challenging to implement than the MH approach discussed in the previous subsection, it is found that they mix more rapidly, generally implying faster convergence (Scott 2002). Further details on the estimation of HMMs using the Gibbs sampling approach can be found in Scott (2002).

Lastly, similar to classical likelihood optimization for HMMs, Bayesian approaches for HMMs are also susceptible to under/over flow in computing the likelihood function. Fortunately, the same scaling solution referred to above for likelihood estimation can be applied to Bayesian estimation. Furthermore, while less problematic than for numerical likelihood estimation, starting values can also play a role in MCMC estimation, particularly with respect to time to convergence. One option is to run several MCMC chains starting from different randomly chosen starting values. Another approach is to choose "smart" starting values by starting the MCMC algorithm around the robust maximum likelihood estimate obtained for the simpler HMM model ignoring cross-customer heterogeneity.

XX.3.3.3 Label switching

Label switching refers to the invariance of the likelihood function of the HMM to a relabeling of the components of the model. This is also an important concern for finite mixture models. While for maximum likelihood estimation or EM algorithm label switching only means that the interpretation of the state ordering may vary from one run to another, it is very important to properly address this issue in Bayesian MCMC estimation of HMMs. The reason is that over the course of the sampling in the MCMC draws, the labeling of the unobserved states can shift leading to mixing posterior parameter draws from multiple states. Label switching is particularly problematic when the HMM states are not well separated, as in such situations the sampler is more likely to jump between states. We refer the reader to Frühwirth-Schnatter (2006; Section 3.5.5) for a detailed discussion and an illustration.

Label switching can often be detected by investigating the iteration plots of the MCMC sampler. If label switching occurs at some point in the iteration sequence, two sets of parameters will switch their value. One approach to deal with label switching when running an MCMC algorithm is to force a unique labeling by imposing constraints on the parameter space. For instance, the means of a normally distributed variable across the K states may be ordered such that $\mu_1 < \mu_2 < \cdots < \mu_K$. This could be implemented in the likelihood function by the re-parametrization $\mu_k = \varphi_1 + \sum_{k'=2}^{k} \exp(\varphi_{k'})$ for $k = 2, \ldots, K$, and $\mu_1 = \varphi_1$. Several researchers have criticized this approach (e.g. Celeux 1998), because the choice of the constraints can shape the posterior distribution of the parameters. A second approach is to run an unconstrained MCMC algorithm and apply post-processing where the unique labels are recovered through choosing an ordering of state-specific parameters or through clustering (Celeux 1998; Frühwirth-Schnatter 2001; Richardson and Green 1997). It is advisable to post-process the MCMC run according to different choices of the labels to investigate the consequences on the final solution and interpretation of the state-specific parameters.

XX.4. Applications of HMMs in Marketing

Probably the first HMM-like model in marketing was the model of Poulson (1990), in which customers were allowed to change their membership in latent classes over time. HMMs in marketing have been primarily used to model how customers (and sometimes firms) transition among a set of latent states over time. However, it is only in the mid and late 2000s that these model started to diffuse to the marketing literature (see Table XX.1 for a non-comprehensive summary of HMMs in marketing). In the context of customers,

the latent states could represent attention (Liechty et al. 2003, Wedel et al. 2008, van der Lans, Pieters and Wedel 2008ab; Shi, Wedel and Pieters 2013), the relationship between the customer and the firm (Netzer et al. 2008; Ascarza and Hardie 2013; Romero, van der Lans and Wierenga 2013; Ma, Sun and Kekre 2015), customers' value system (Brangule-Vlagsma, Pieters and Wedel 2002), channel migration (Mark, Lemon and Vandenbosch 2014), internet browsing behavior and search (Montgomery et al. 2004; Stuttgen, Boatwright and Monroe 2012), consumers' choice among portfolio of products (Paas, Vermunt and Bijmolt 2007; Schweidel et al. 2011), customer satisfaction (Ho, Park and Zhou 2006), store loyalty and promotion sensitivity (Shi and Zhang 2014), purchase cycles states (Park and Gupta 2011), latent behavioral learning strategies (Ansari, Montoya and Netzer 2012), bidding strategies (Shachat and Wei 2012), and households lifecycle stages (Du and Kamakura, 2006). HMMs have also been used to capture how marketing actions could affect the transition among states (Netzer et al. 2008; Montoya et al. 2010; Li, Sun and Montgomery 2011; Kumar et al. 2011; Luo and Kumar 2013; Zhang et al. 2014).

In the most general sense the latent attrition models (e.g., Fader, Hardie and Shang 2010; Schweidel and Knox 2013) can thought of as a special case of a HMM with two states, where attrition is an absorbing state. This model has been extended to allow for an always share model (Ma and Buschken 2011). Schwartz, Bradlow and Fader (2014) explore the relationship between the HMM and several of its constraint versions such as the latent attrition model.

Article	Capturing Unobserved Heterogeneity	Estimation	Non Homogenous HMM (Covariates in Q)	# of states*	State dependent distribution
Poulsen (1990)	No	EM algorithm	No	2	Multinomial choice
Brangule-Vlagsma, Pieters and Wedel (2002)	No	Maximum likelihood	Yes	6	Rank-order logit
Liechty, Pieters and Wedel (2003)	No	Reversible Jump MCMC	No	2 (theory)	First-order continuous time Markov chain
Montgomery et al. (2004)	Yes, Q, M and π	Reversible Jump MCMC	No	2	Multinomial Probit
Du and Kamakura, (2006)	No	EM algorithm	No	13	Multivariate (Bernoulli/Normal)
Paas, Vermunt and Bijmolt (2007)	No	EM algorithm	Yes	9	Multivariate (Bernoullis)
Moon, Kamakura and Ledolter (2007)	Yes, only in M	MCMC State Augmentation	No	2 (theory)	Linear regression (Normals)
Netzer, Lattin and Srinivasan (2008)	Yes, in Q and π	MCMC Direct Likelihood	Yes	3	Binary logit
Wedel, Pieters and Liechty (2008)	No	Reversible Jump MCMC	No	2 (theory)	First-order continuous time Markov chain
van der Lans, Pieters and Wedel (2008a)	Yes, only in Q	MCMC State Augmentation	No	2 (theory)	Categorical - Square-root link function
van der Lans, Pieters and Wedel (2008b)	Yes, only in M	MCMC State Augmentation	No	2 (theory)	Spatial point process
Montoya, Netzer and Jedidi (2010)	Yes, in Q and M	MCMC Direct Likelihood	Yes	3	Binomial
Ebbes, Grewal and Desarbo (2010)	No	MCMC State Augmentation	No	3	Multivariate Normal
Schweidel, Bradlow and Fader (2011)	Yes, in Q and π	MCMC Direct Likelihood	Yes	4	Multivariate (Markov chain / Multinomial Logit)
Park and Gupta (2011)	Yes, only in M	Simulated Maximum likelihood	Yes	2 (theory)	Multinomial Logit
Li, Sun and Montgomery (2011)	Yes, Q, M and π	MCMC	Yes	3	Multivariate Probit
Kumar et al. (2011)	Yes, only in M	Maximum likelihood	Yes	3	Multivariate To
Lemmens, Croux and Stremersch (2012)	Yes, only in M	EM algorithm	Yes	3	Linear regression (Normal)
Stuttgen, Boatwright and Monroe (2012)	Yes, in Q and M	MCMC State Augmentation	Yes	2 (theory)	Multivariate (Markov chain / Multinomial)
Ansari, Montoya and Netzer (2012)	Yes, Q, M and π	MCMC Direct Likelihood	Yes	2 (theory)	Multinomial Logit
Shachat and Wei (2012)	No	EM algorithm	No	3 (theory)	Normal
Ascarza and Hardie (2013)	Yes, in Q and M	MCMC State Augmentation	No	3	Poisson
Romero, van der Lans and Wierenga (2013)	Yes, in M and π	EM algorithm	No	7 and 9	Multivariate (Truncated NBD / Gamma-Gamma)
Shi, Wedel and Pieters (2013)	No	MCMC State Augmentation	No	3	2 Layers of Hidden States
Luo and Kumar (2013)	Yes, in Q and M	MCMC State Augmentation	Yes	3	Multivariate Tobit model
Mark et al. (2013)	No	EM algorithm	No	4	Hurdle Poisson
Mark, Lemon and Vandenbosch (2014)	No	Maximum likelihood	No	3	Poisson
Shi and Zhang (2014)	Yes, only in Q	MCMC Direct Likelihood	Yes	3	Type-II Tobit model
Zhang et, Netzer and Ansari (2014)	Yes, Q, M and π	MCMC Direct Likelihood	Yes	2	Multivariate (Log-logistic / Log-normal / Binary logit)
Schwartz, Bradlow and Fader (2014)	Yes, Q, M and π	MCMC State Augmentation	No	2 (theory)	Bernoulli
Ma, Sun and Kekre (2015)	Yes, Q, M and π	MCMC State Augmentation	Yes	3	Multinomial Logit
Zhang, Watson and Palmatier (2016)	Yes, only in π	MCMC Direct Likelihood	Yes	4	Normal

Table XX.1 Non-comprehensive list of marketing papers using HMMs

* "theory" means the number of states were selected based on theoretical grounds rather than based on model fit.

One common theme across all of the above applications of HMMs in marketing is that in all cases the customer behavior was governed by an underlying state that is unobserved to the researcher, while it is possibly for the customer to change to a different state over time. Such states were often the state of customer attention to marketing information, the customer's strategy of making choices, her lifecycle stage, or her loyalty, trust, satisfaction level, or, more generally, her relationship status with the firm. Research has often investigated the customers' transitions among these states and how the context of the decision and the firm's action affects customer's transitions to states that are more favorable to the firm or lead to higher welfare.

In some marketing applications the unit of analysis was not the consumer. Luo and Kumar (2013), Zhang et al. (2014) and Zhang, Watson and Palmatier (2016) have all used HMMs to investigate the relationship between buyers and sellers in the context of B2B relationships. Ebbes et al. (2010) looked at how firms' (banks') competitive landscape changed over time. Moon, Kamakura and Ledolter (2007) used a HMM to uncover firms' latent competitive promotions. Lemmens, Croux and Stremersch (2012) looked at evolving segments of countries in the context of new product growth.

Several aspects make the application of HMMs in marketing different from applications in other fields. First, HMMs in marketing often leverage the latent structure as a means to capture the data generating process of the customer's behavior, and use this in order to understand and predict the outcome of the customer behavior, whereas in many of the HMM applications outside of marketing the objective is mostly to recover the underlying state (e.g., words in speech recognition). An exception in marketing is

Ebbes and Netzer (2016) who use HMMs to identify the latent job seeking state using social media data.

Second, as discussed in Section XX.2.4, because most HMM applications in marketing involve multiple time series for different consumers, capturing heterogeneity is very important. Indeed, as can be seen in Table XX.1, most marketing applications have captured unobserved heterogeneity using a random-effect or latent class approach. Finally, one of the main reasons to apply a HMM in marketing is to investigate what customer or firm behavior can create a regime shift (i.e. a transition among states) in the customer behavior. Accordingly, many non-homogeneous HMMs that incorporate time-varying covariates in the transition matrix are much more common in marketing relative to other fields. For example, Montoya et al. (2010), have looked at how detailing and sampling can affect physicians' drug prescription and found that detailing can help transition physicians from a low prescription state to a higher one and sampling was mainly useful in keeping physicians in the prescription state. In the context of B2B buyer-seller relationships Luo and Kumar (2013) find that direct mail and phone calls can help transitioning a buyer from a lower to a higher state of purchase behavior. One of the main benefits of using HMMs in marketing is to disentangle the short-term and long-term effects of marketing activities through the incorporation of these variable in the transition probabilities and in the state-dependent distributions of observed data.

From the above discussion it is clear that the body of literature that utilizes HMMs to capture marketing dynamics is sizeable and growing. We expect to see many more application of these useful models, to model latent and dynamic customer behavior. For example, as behavioral researchers in marketing increase their use of repeated

observational experiments and secondary data, HMMs can be used to capture the dynamics of customer behavior in areas such arousal, fatigue, or goal pursuit.

XX.5. An Illustrative Application of HMM

To illustrate several of the considerations involved in building and estimating a HMM in marketing, we describe a typical marketing application of HMMs involving the customer relationship management (CRM) between a business-to-business (B2B) company and its industrial clients. For this illustration we use simulated data.

XX.5.1 Description of the Data

We consider a B2B firm that has CRM data for $N = 1{,}080$ customers. Based on the sales to these customers and their total category expenditures, the firm computed Share-Of-Wallet (SOW) at the customer-level for 20 consecutive months (time periods), i.e. $T = 20$. In addition, the customer database contains time-varying marketing mix variables such as price, sales representative visits, and a direct mail. Table XX.2 shows the structure of the database (the first 22 observations). Such panel data structures are commonly used in HMM applications in marketing. The variable CustomerID is used to identify all observations that belong to the same customer (index i), and the variable Period will be used to identify the consecutive time periods (index t in Section XX.2).

Observation	CustomerID	Period	SOW	Price	SalesVisit	DirectMail
1	1	1	63.60	4.00	1	0
2	1	2	45.10	4.48	1	1
3	1	3	43.56	4.36	1	0
4	1	4	36.93	4.34	0	0
5	1	5	19.37	5.50	1	0
6	1	6	60.62	4.29	0	0

7	1	7	71.45	3.86	0	0
8	1	8	53.95	5.87	1	0
9	1	9	42.99	4.55	1	1
10	1	10	41.71	5.82	1	1
11	1	11	28.77	2.17	0	0
12	1	12	18.30	4.91	1	0
13	1	13	23.06	4.56	0	0
14	1	14	22.77	5.76	1	0
15	1	15	15.11	5.06	1	0
16	1	16	24.96	2.53	1	0
17	1	17	30.17	4.66	1	1
18	1	18	14.43	3.76	1	0
19	1	19	28.46	5.55	1	1
20	1	20	18.48	5.65	1	0
21	2	1	34.39	5.99	0	1
22	2	2	37.49	6.48	0	1

Table XX.2. Data of the first 22 observations

XX.5.2 The basic model setup

The firm is interested in inferring the states of loyalty (SOW) between the firm and its clients. Importantly, the firm would like to know to which loyalty state each customer belongs to during each time period, and how the firm could potentially increase the SOW using its marketing mix. In this example, we will use a HMM for that purpose.

In the HMM, SOW is the observed variable Y_{it}. The last three columns in Table XX.2 present the covariates X_{it} that can affect the SOW of each client at each time period. In other words, the covariates X_{it} have a short-term effect on customer behavior. Thus, the SOW of a customer in a specific time period depends on the price level, whether or not the customer was visited by a sales representative, and whether or not the customer received a direct mailing:

$$SOW_{it} = \beta_{0s} + \beta_{1s}Price_{it} + \beta_{2s}SalesVisits_{it} + \beta_{3s}DirectMail_{it} + \epsilon_{it}. \quad \text{(XX.11)}$$

In the model in (XX.11), we allow for multiple states of SOW with different effects of marketing actions in each state. Therefore, the regression parameters ($\beta_{0s}, \beta_{1s}, \beta_{2s}$, and β_{3s}) in (XX.11) are state-specific and have a subscript s. We simulated the customer data consisting of three hidden states (i.e. $K = 3$ and $s = 1,2,3$), with initial state probabilities (π) of 0.43, 0.40, and 0.17, and the following values for the regression parameters for each of the three states:

1. Low SOW ($\beta_{01} = 30$), marketing effects (price: $\beta_{11} = -2.5$; sales visit: $\beta_{21} = 0$; DM: $\beta_{31} = 2.5$);
2. Medium SOW ($\beta_{02} = 60$), marketing effects (price: $\beta_{12} = -1.5$; sales visit: $\beta_{22} = 0$; DM: $\beta_{32} = 1$);
3. High SOW ($\beta_{03} = 85$), marketing effects (price: $\beta_{13} = -0.5$; sales visit: $\beta_{23} = 0$; DM: $\beta_{33} = 0$).

We take ϵ_{it} to be i.i.d. following a normal distribution. Therefore, each diagonal element of the 3×3 state-dependent distribution matrix (M_{it}) is a univariate normal distribution with mean $\beta_{0s} + \beta_{1s}Price_{it} + \beta_{2s}SalesVisits_{it} + \beta_{3s}DirectMail_{it}$ and standard deviation σ_s, for $s = 1,2,3$. Additionally, the model includes the transition matrix (Q), which we will further discuss below in the results section.

We estimate two versions of the HMM:

1) A basic HMM with a homogeneous transition matrix (no covariate effects) but with effect of covariates on the state-dependent distribution of SOW;
2) A non-homogenous (effects of a covariate (sales visits) in the transition matrix) and heterogeneous (cross-customer heterogeneity using a Latent Class approach) HMM.

XX.5.3 Estimation of HMMs using Latent Gold

We use the software program Latent Gold 5.1 (Vermunt and Magidson, 2015), distributed by www.statisticalinnovations.com, to estimate the two HMMs. One of the advantages of Latent Gold 5.1 for HMM estimation is that it includes a module for estimating basic HMMs, which can be accessed either through the menu (model option Markov) or through the syntax. To set up a HMM in Latent Gold, the observed variable(s) Y_{it} (here: SOW) has to be selected as Indicator. Next, the state-dependent distribution that corresponds to this variable has to be selected. Latent Gold allows for the following options: continuous, count, ordinal and nominal, which indirectly specifies the underlying distribution of Y_{it} (see Section XX.2.1). In this application, SOW is a continuous variable, which will be modeled as a Normal distribution by Latent Gold. In addition, Latent Gold allows the user to include covariates X_{it}, which can have an impact on the initial state, the transition probabilities and/or the state-dependent distribution. As discussed in section XX.2.5, when covariates are included in the transition probabilities, they are postulated to create a regime shift in the customer behavior and have a long-term impact, whereas covariates that are included in the state-dependent distribution only affect the customer behavior in the current time period, and therefore have a short-term impact. As mentioned previously, in this application, we first include the marketing mix covariates in the state-dependent distribution of SOW only, assuming they only have a short-term impact on customer behavior. Later on we extend that model and include some of these variables in the transition probability matrix as well, to investigate their effect on long-term customer behavior. Several of the other model specifications described in this chapter can be estimated as well. For detailed information on how to specify various

HMMs in Latent Gold, we refer to the manual (Vermunt and Magidson, 2015). In Latent Gold, parameter estimates are obtained by means of the EM algorithm (see Section XX.3.1). For estimation of HMMs using Bayesian approaches, or for specific, more advanced, specification of HMMs, we recommend using other statistical programing tools such as R or Matlab.

XX.5.4 Results of Alternative Specifications of the HMMs

Results for a basic HMM. First, we estimate a basic HMM with a homogeneous transition matrix and covariates in the state dependent distribution. Model estimates are obtained for 1 to 6 hidden states, requiring the estimation of, respectively, 5 to 65 parameters (including estimates for the initial state probabilities, the transition probabilities matrix, the parameters of the Normal distribution of SOW for each state, given the covariates price, sales visits and direct mail). We compare various information criteria across these solutions to determine the most appropriate number of states K, see Table XX.3. Minimum values of BIC and CAIC are obtained for an HMM with 3 hidden states, which corresponds to the true number of states in this simulated data example. AIC is minimized for 6 hidden states, reflecting the common finding that AIC tends to overestimate the number of states (see Section XX.2.6). Therefore, we choose the HMM with three hidden states and we present the detailed estimation results for this model in Table XX.4.

Number of States	Number of Parameters	AIC	BIC	CAIC
1	5	200151.28	200176.20	200181.20

2	13	188460.35	188525.15	188538.15
3	23	179899.53	**180014.18**	**180037.18**
4	35	179898.79	180073.25	180108.25
5	49	179786.49	180030.74	180079.74
6	65	**179733.66**	180057.67	180122.67

Table XX.3. Information criteria for the HMMs with state-dependent covariate effects on SOW*

*Figures in bold are the minimum values for AIC/BIC/CAIC

Initial state distribution			
State (t=0)	1	2	3
Probability	0.44 (.02)	0.40 (.02)	0.16 (.01)
Transition probability matrix			
	State (t)		
State (t-1)	1	2	3
1	0.80 (.01)	0.14 (.01)	0.06 (.01)
2	0.09 (.01)	0.80 (.01)	0.11 (.01)
3	0.07 (.01)	0.10 (.01)	0.83 (.01)
State-dependent distribution parameters of the observed variable (SOW)			
	1	2	3
Intercept (b_0)	35.84 (.62)	59.17 (.61)	84.74 (.59)
Price (b_1)	-2.65 (.12)	-1.43 (.11)	-0.55 (.11)
Sales visit (b_2)	-0.25 (.23)	0.65 (.23)	0.49 (.24)
Direct Mail (b_3)	2.55 (.25)	1.36 (.25)	-0.04 (.25)

Table XX.4. Estimation results (parameter estimates and standard errors) of the HMM with state-dependent covariate effects on SOW, with three states ($K = 3$).

Most estimated parameter values closely match their true values underlying the data generation procedure. The estimates for the intercept range from 35.84 for customers in State 1 to 84.74 for customers in State 3. The three states also differ substantially in terms of their response to marketing actions: customers in State 1 (the low SOW state as

indicated by the intercept) are the most sensitive to price and direct mail. Price has a negative effect on SOW in all states (all p-values < .01), with the largest effect in State 1. Furthermore, the effect of direct mail on SOW is significant and positive for States 1 and 2 (p-values <.01) and not significant for State 3 (p=.88).

Interestingly, the estimated effect of sales visits is relatively small, though significantly positive in States 2 and 3 (p<.01 and p=.04, respectively). We note that the *true* effect of sales visits on SOW is 0 in each state, i.e. there is *no* short-term effect of sales visits on SOW, and this bias in the estimated effect of sales visit on SOW is due to a model misspecification. As we will demonstrate below, sales visits have a significant positive effect on the transition among states, stimulating customers to switch to a state with higher SOW. In other words, sales visits have a long-term effect but no short-term effect on customer behavior. Hence, the long-term effect of sales visits on SOW is wrongfully captured by the short-term effect of sales visits on SOW in the state-dependent distribution in this particular HMM, leading to a potentially misinterpretation by the manager of the usefulness of sales visits on short-term behavior.

Examining the estimates for the initial state distribution and transition probability matrix, we see that customers are most likely to start in the low and medium SOW states (States 1 and 2). Subsequently, the customers have a fairly high probability of staying in the same state from one time period to the next, as the estimated diagonal elements of the transition probability matrix are fairly high (80% or higher), suggesting that the states are "quite sticky" across customers. While the estimated parameters for the initial state distribution are fairly close to their true values, we will demonstrate below that the

stickiness in the transition probabilities is overestimated by this simple HMM, because cross-customer heterogeneity is not appropriately accounted for.

Results for a non-homogenous, heterogeneous HMM. We will now use the same simulated data to estimate a non-homogenous (covariates in the transition probability matrix) and heterogeneous (latent class approach to capture cross-customer heterogeneity) HMM with Latent Gold. This can be done in Latent Gold by specifying the number of latent classes in the Advanced Menu option, or by defining a latent variable "Class" and including it in the "equation" lines in a Latent Gold syntax file. We allow for cross-customer heterogeneity in the model parameters only in the transition probability matrix. In addition, we include the sales visits covariate in the transition probabilities such that sales visits have a potential long-term effect on customer behavior. Including sales visits as a covariate to the transition probability matrix adds $K \times (K-1)$ additional parameters to the model, where K is the number of states in the HMM. Allowing for the transition probability matrix to be heterogeneous through (say) S latent classes, multiplies the number of transition matrix parameters by S, because we now have one transition matrix for each of the S latent class segments.

In addition to determining the number of states, we now also need to determine the number of latent class segments. For brevity, we only estimate HMMs with 2 and 3 latent classes and with 1 to 6 latent states, and compare the relative fit of these 12 model specifications.[5] The minimum CAIC rule suggests the model with 3 hidden states and 2 latent classes as most appropriate (Table XX.5), which corresponds to the true number of latent states and latent classes with which the data was generated. In addition, the CAIC

[5] In general one may wish to vary the number of latent classes from 1 to a larger number than 3.

values are lower than those of the previous homogenous HMM with three latent states (see Table XX.3), which indicates that accommodating cross-customer heterogeneity in the transition matrix, by means of a latent class structure and by including the covariate sales visits in the transition probabilities, is warranted. Therefore, we present the detailed estimation results of the HMM with 2 latent classes and 3 hidden states in Table XX.6.

Number of States	Number of Latent Classes	
	2	3
1	200189.19	200197.17
2	188363.87	188401.11
3	**179771.97**	179860.98
4	179957.21	180136.37
5	180182.21	180448.84
6	180488.81	180965.82

Table XX.5 Information criterion CAIC for the HMMs with heterogeneity and covariates in the transition matrix*

*Figure in bold is the minimum value for CAIC

The estimates for the parameters of the state-dependent distributions of the HMM with heterogeneity and covariates in the transition matrix, are quite similar to those obtained from the simple homogenous HMM (Table XX.4 and XX.6). Importantly, all estimated parameter values now closely match the true values underlying the data generation procedure, including the null-effect of sales visits. More specifically, the estimated values for the intercepts are very close to the true values (35, 60 and 85), and similarly for the estimated price effects (all p-values < .01; true values -2.5, -1.5, and -

0.5). The estimated effects of direct mail are significant and positive for States 1 and 2 (p-values <.01; true values 2.5 and 1.0), and not significant for State 3 (p=.76; true value 0.0). In other words, looking at short-term customer behavior, the customers in the low SOW state (state 1) are most sensitive to price and respond positively to direct mail. On the other hand, the customers in the high SOW state, are least sensitive to price, and direct mail is not an effective marketing instrument for these type of customers to increase their SOW.

Lastly, considering the sales visits covariate, the estimated direct effects of sales visits on SOW are very small and not significant anymore for each of the three states (true values 0.0 for each state; p-values .09, .64 and .50, respectively). Hence, sales visits is not an effective marketing instrument to influence short-term customer behavior. This highlights the importance of specifying the correct heterogeneity in HMMs. Apparently, the sales visit covariate picked up spurious correlation in the basic homogenous HMM (Table XX.4). As such, the manager would incorrectly conclude that sales visits have a short-term, positive, effect on behavior. In fact, as we will see next, sales visits have a long-term effect on behavior by moving customers to a higher SOW state, i.e. inducing a (positive) regime shift among customers.

Before we discuss the long-term effect of sales visits on behavior, we will first discuss the results for the latent class approach that was used to capture unobserved cross-customer heterogeneity. As mentioned before, the best model to capture cross-customer heterogeneity (according to the model selection criteria) is a HMM with two latent classes. As indicated in Table XX.6, the two latent classes are estimated to represent 72 and 28 percent of the customers, respectively. As we only included

heterogeneity in the transition probability matrix, we would get two estimated transition probability matrices, one for each latent class. Because we also included the sales visit covariate in the transition probabilities, where sales visit is dummy variable, we get two estimated transition probability matrices, one for sales visit and one for no sales visit, for each latent class. These four matrices are also given in Table XX.6.

Importantly, the four estimated transition probability matrices are quite different between the two classes and depending on whether a sales visit was made in a particular time period. Two points are worth noting about the estimated transition probability matrices. First, we see that, for customers in the first segment (latent class 1), sales visits have a substantial (and significant) impact on transitioning customers between the middle and the high SOW states (i.e. between States 2 and 3) and on keeping them in the high SOW state (State 3). For example, following a sales visit, an average Segment 1 customer in the highest SOW state (State 3) has a 90% chance of staying in that state in the next period, but only a 63% chance of staying in that state in the next period without a sales visit. Second, for customers in the second segment (latent class 2), sales visits are mostly effective as an acquisition tool, transitioning them from the low (State 1) to the middle (State 2) SOW state, whereas high SOW customers (state 3) are not much affected by sales visits in the long-run.

The importance of accounting for cross-customer heterogeneity in a HMM, through an unobserved heterogeneity approach (e.g. latent classes), but also through observed covariates (e.g. sales visits), is clearly shown in this example. If we compare the estimation results for the transition probability matrix of the non-homogenous, heterogeneous HMM in Table XX.6 and the basic homogenous HMM in table XX.4, we

see that the stickiness of customers to their state is considerably overestimated by the basic homogenous HMM. A manager may now wrongfully conclude that little can be done to move customers up to a more favorable (i.e. higher) SOW state. In fact, the results of the non-homogenous, heterogeneous HMM show that sales visits can be an effective marketing tool, to either reduce the chance that customers move from a high SOW state to a lower SOW state (latent class 1), or to move customers up from a low SOW state to a higher SOW state (latent class 2). In other words, sales visits have the potential to make customers more "sticky" in staying in a higher SOW state, and by using sales visits the firm has the chance to favorably (for the firm) influence customers' long term behavior towards higher SOW. Such insights would not have been possible using the basic homogenous HMM that ignores cross-customer heterogeneity.

Initial state distribution						
State (t=0)	1	2	3			
Probability	.44 (.02)	.40 (.02)	.16 (.01)			
Transition probability matrix						
		Latent Class 1 size: .72			Latent Class 2 size .28	
		State (t)				
No Sales Visit	State (t-1)	1	2	3	1	2	3
	1	.80 (.01)	.16 (.01)	.04 (.01)	.74 (.01)	.17 (.01)	.09 (.01)
	2	.16 (.01)	.75 (.01)	.09 (.01)	.08 (.02)	.82 (.03)	.10 (.02)
	3	.15 (.02)	.22 (.02)	.63 (.03)	.04 (.01)	.06 (.02)	.90 (.03)
Sales visit	1	.84 (.01)	.11 (.01)	.06 (.01)	.59 (.05)	.25 (.04)	.16 (.03)
	2	.05 (.01)	.82 (.01)	.14 (.01)	.09 (.01)	.80 (.02)	.11 (.01)
	3	.06 (.01)	.04 (.01)	.90 (.01)	.04 (.01)	.09 (.01)	.87 (.01)

State-dependent distribution parameters of the observed variable (SOW)			
State	1	2	3
Intercept (b_0)	36.00 (.62)	59.68 (.61)	85.06 (.59)
Price (b_1)	-2.66 (.12)	-1.47 (.11)	-0.56 (.11)
Sales visit (b_2)	-0.40 (.24)	0.16 (.23)	0.11 (.24)

Direct Mail (b_3)	2.56 (.25)	1.33 (.25)	-0.08 (.25)

Table XX.6. Estimation results (parameter estimates and standard errors) of the HMM with heterogeneity and covariates in the transition matrix, with three states and two latent classes.

To sum, this section illustrates an application of HMMs to a typical marketing problem. It demonstrates the considerations involved in specifying the HMM and structuring the data. We also discuss how to estimate the model and how to choose the number of latent states. Importantly, we highlight the relevance of accounting for cross-customer heterogeneity. Our illustration demonstrates the type of insights that can be generated from interpreting the model's parameter estimates, and, in particular, the effect of marketing actions on the transitions among states and on the state dependent behavior. In this simulation example, the estimation procedure through Latent Gold was able to correctly recover the model parameters of an HMM with cross-customer heterogeneity.

XX.6. Conclusions

In this chapter we have provided an overview of HMMs with a particular focus on the unique aspects of HMMs applied to marketing problems. HMMs are a flexible class of models that can be used to model dynamics in a sequence of observations. While HMMs have been developed and applied in many fields other than marketing, their application and implementation in marketing requires further development. In particular, the availability of "panel data" in marketing implies that we have multiple time series (one for each customer), which requires special attention as the basic HMMs have traditionally been developed for applications where there is only one (often very long) time series (see e.g., Zucchini and MacDonald (2009) for such HMM applications in various fields).

Particularly, addressing customer-specific heterogeneity is a primary concern when applying HMMs to a marketing problem with possibly heterogeneous agents, as we know from extant literature in marketing (see also Chapter XXX). Not properly accounting for such heterogeneity can lead to misleading insights regarding the dynamics underlying the behavioral process. Indeed as we report in Table XX.1 almost all HMM applications in marketing have accounted for heterogeneity in one form or another.

Another important difference in HMM applications in marketing relative to other fields is the notion that firms can (and would like to) nudge customers' behavior in a way that would be profitable to the firm. In a HMM application context, this often means that the firm would like to move the customer from one state to another (e.g. from a low loyalty state to a high loyalty state). Or, in another application context, would the firm like to prevent the customer from drifting down from an active state to a passive or churn state. Such research questions can be addressed by extending the basic HMMs and including marketing activity into the model. Specifically, this can be done by developing non-homogenous HMMs that relate the probabilities in the transition matrix to marketing actions. Such HMMs can capture a long-term or a regime shift effect of marketing actions on customer behavior. Indeed, non-homogenous HMMs are quite common in marketing but are fairly rare outside of marketing.

We have discussed in this chapter many recent papers and applications in marketing, as summarized in Table XX.1. From our discussion it becomes clear that most of the applications of HMMs in marketing are fairly recent (within the last decade) and the use of these models is growing rapidly (we are aware of many working papers applying HMMs that were not included in this chapter).

In our experience there are aspects of HMMs in marketing that are worthwhile further research. First, building on the notion of heterogeneity, it is possible that customers are different not only in terms of the way they transition among states or how they behave given a state, but also in the number of states they transition among. In other words, instead of developing a HMM with K states, one could consider a HMM with K_i states, i.e. a different number of states for each customer. Such a model would greatly complicate the model selection problem, as we would now need to select the number of states at the customer level (see Padilla et al. 2016 for recent work in this area). Reversible jump algorithms (e.g. Ebbes et al. 2015) can be a useful avenue to address these issues. Similarly, a fruitful avenue of research could explore the topic of state generation. A customer could be moving among a set of states for a while and due to an exogenous to the customer event (e.g. introduction of a new product) or an endogenous to the customer event (e.g. getting married), she may start visiting a state she has never visited before. Modeling such state generation could help to better understand the evolution of customers over time.

A second fruitful area of research in applying HMMs in marketing may be in the context of data fusion. Because HMMs model a latent state that evolves over time, one can use the latent state for data fusion by merging together different sets of information at different time intervals, using their common latent state. For instance, Ebbes and Netzer (2016) merge survey data collected at specific time intervals with continuously observed customer activity data.

Third, in most HMM applications in marketing the interpretation of the states have been empirically inferred from the estimation results. However, psychological and

consumer behavior research in marketing often has a strong a-priori theory regarding what could in fact be the meaning of underlying states. Research in this area typically conducts experiments with a particular manipulation aimed at transitioning an individual from one behavioral state to another (e.g., affective states). Therefore, we encourage future researchers to use HMMs in the context of behavioral experiments with repeated observations to identify the latent behavioral states and the transitions among them, as a function of the experimental design.

One natural question to ask is when to use a discrete version of dynamics such as HMMs and when to use a more continuous version of dynamics such as state-space dynamics as in Kalman filter-type approaches (See Chapter XX). There are several notable advantages of HMMs over their continuous counterparts. First, HMMs are particularly useful in capturing dynamics when the underlying dynamics are reflective of regime shift dynamics, as opposed to a gradual dynamics. On the other hand, when the underlying dynamic process is more gradual we recommend using the continuous state-space approaches. Second, HMMs capture dynamics in a semi-parametric manner and are therefore more flexible than most of the continuous state-space approaches, which often rely on specific distributional and parametric assumptions (e.g., the change from one period to another is drawn from a normal distribution). Third, from an interpretation point of view, relative to continuous dynamic models, applications of HMMs in marketing are attractive because they are easily interpretable and often lead to easy to communicate managerial insights (similar to segmentation studies) such as "marketing action X shifts customers from a low state of activity to a high state of activity." Finally, if one estimates a HMM on a truly continuous dynamics process, the HMM would approximate the

continuous dynamic process well by letting the number of states K grow. This of course comes at a cost, as for such cases the HMM is less parsimonious than a continuous dynamics state-space model. Therefore, if the number of states recommend by the model selection criteria becomes large, we recommend the researcher to explore also continuous dynamics state-space models. Future research could investigate the similarities and differences between HMMs and other state-space models in marketing problems.

References

Ansari, A., Montoya, R. and Netzer, O., 2012. Dynamic learning in behavioral games: A hidden Markov mixture of experts approach. *Quantitative Marketing and Economics*, *10*(4), pp.475-503.

Ascarza, E. and Hardie, B.G., 2013. A joint model of usage and churn in contractual settings. *Marketing Science*, *32*(4), pp.570-590.

Atchadé, Y.F. and Rosenthal, J.S., 2005. On adaptive markov chain monte carlo algorithms. *Bernoulli*, *11*(5), pp.815-828.

Bacci, S., Pandolfi, S., and Pennoni, F. (2014), "A comparison of some criteria for states selection in the latent Markov model for longitudinal data," *Advances in Data Analysis and Classification*, 8, 125-145.

Bartolucci, F., Farcomeni, A. and Pennoni, F., 2014. Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *Test*, *23*(3), pp.433-465.

Baum, L.E., 1972. An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, *3*, pp.1-8.

Baum, Leonard E., and Ted Petrie. "Statistical inference for probabilistic functions of finite state Markov chains." *The annals of mathematical statistics* 37, no. 6 (1966): 1554-1563.

Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains." *The annals of mathematical statistics* 41, no. 1 (1970): 164-171.

Brangule-Vlagsma, K., Pieters, R.G. and Wedel, M., 2002. The dynamics of value segments: modeling framework and empirical illustration. *International Journal of Research in Marketing*, *19*(3), pp.267-285.

Celeux, G., 1998. Bayesian inference for mixture: The label switching problem. In *Compstat* (pp. 227-232). Physica-Verlag HD.

Celeux, G., F. Forbes, C. P. Robert, and D. M. Titterington, (2006), "Deviance Information Criteria for Missing Data Models", *Bayesian Analysis*, 1 (4), 651-674

Chib, S., 1995, Marginal likelihood from the gibbs output, *Journal of the American Statistical Association*, 90, 1313-21.

Chib, S., 2001, Markov Chain Monte Carlo methods: computation and inference, in: Handbook of Econometrics, Volume 5, editors Heckman, J.J. and E. Leamer, Elsevier

Chintagunta, P.K., 1998. Inertia and variety seeking in a model of brand-purchase timing. *Marketing Science*, *17*(3), pp.253-270.

Congdon, Peter. "Bayesian statistical modelling." (2002): 643.

Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pp.1-38.

Du, R.Y. and Kamakura, W.A., 2006. Household life cycles and lifestyles in the United States. *Journal of Marketing Research*, *43*(1), pp.121-132.

Dubé, J.P., Hitsch, G.J. and Rossi, P.E., 2010. State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics*, *41*(3), pp.417-445.

Ebbes, P., Grewal, R. and DeSarbo, W.S., 2010. Modeling strategic group dynamics: A hidden Markov approach. *QME, 8*(2), pp.241-274.
Ebbes, P., Liechty, J.C. and Grewal, R., 2015. Attribute-level heterogeneity. *Management Science*, *61*(4), pp.885-897.
Ebbes, P and O. Netzer, 2016, Using social media data to identify and target job seekers, *working paper*.
Eddy, S.R., 1998. Profile hidden Markov models. *Bioinformatics*, *14*(9), pp.755-763.
Ehrenberg, A.S., 1965. An appraisal of Markov brand-switching models. *Journal of Marketing Research*, pp.347-362.
Fader, P.S., Hardie, B.G. and Shang, J., 2010. Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science*, *29*(6), pp.1086-1108.
Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, *96*(453), pp.194-209.
Frühwirth-Schnatter, S., 2006. *Finite mixture and Markov switching models*. Springer Science & Business Media.
Green, P. J., (1995), Reversible jump markov chain monte carlo computation and Bayesian model determination, Biometrika, 82 (4), 711-732
Guadagni, P.M. and Little, J.D., 1983. A logit model of brand choice calibrated on scanner data. *Marketing science*, *2*(3), pp.203-238.
Hamilton, J.D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pp.357-384.
Hamilton, J.D., 2008. "Regime switching models." *The New Palgrave Dictionary of Economics*. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan. The New Palgrave Dictionary of Economics Online. Palgrave Macmillan. 08 September 2008
Heckman, J.J., 1981. Heterogeneity and state dependence. In *Studies in labor markets* (pp. 91-140). University of Chicago Press.
Ho, T.H., Park, Y.H. and Zhou, Y.P., 2006. Incorporating satisfaction into customer value analysis: Optimal investment in lifetime value. *Marketing Science, 25*(3), pp.260-277.
Hughes, James P., and Peter Guttorp. "A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena."*Water Resources Research* 30.5 (1994): 1535-1546.
Jurafsky, D. and J. H. Martin, 2008, Speech and Language Processing, Prentice Hall, 2nd edition
Kamakura, W.A. and Russell, G., 1989. A probabilistic choice model for market segmentation and elasticity structure. *Journal of marketing research*, *26*, pp.379-390.
Keane, M.P., 1997. Modeling heterogeneity and state dependence in consumer choice behavior. *Journal of Business & Economic Statistics*, *15*(3), pp.310-327.
Kumar, V., Sriram, S., Luo, A. and Chintagunta, P.K., 2011. Assessing the effect of marketing investments in a business marketing context. *Marketing Science*, *30*(5), pp.924-940.

Leeflang, Peter S. H., Jaap E. Wieringa, Tammo H. A. Bijmolt, and Koen H. Pauwels. 2015. Modeling Markets: Analyzing Marketing Phenomena and Improving Marketing Decision Making. New York: Springer.

Lemmens, A., Croux, C. and Stremersch, S., 2012. Dynamics in the international market segmentation of new product growth. *International Journal of Research in Marketing*, *29*(1), pp.81-92.

Li, S., Sun, B. and Montgomery, A.L., 2011. Cross-selling the right product to the right customer at the right time. *Journal of Marketing Research*, *48*(4), pp.683-700.

Liechty, J., Pieters, R. and Wedel, M., 2003. Global and local covert visual attention: Evidence from a Bayesian hidden Markov model. *Psychometrika*, *68*(4), pp.519-541.

Luo, A. and Kumar, V., 2013. Recovering hidden buyer-seller relationship states to measure the return on marketing investment in business-to-business markets. *Journal of Marketing Research*, *50*(1), pp.143-160.

Ma, L., Sun, B. and Kekre, S., 2015. The Squeaky Wheel Gets the Grease—An Empirical Analysis of Customer Voice and Firm Intervention on Twitter. *Marketing Science*, *34*(5), pp.627-645.

Ma, S. and Büschken, J., 2011. Counting your customers from an "always a share" perspective. *Marketing Letters*, *22*(3), pp.243-257.

Mamon, R.S. and Elliott, R.J. eds., 2007. *Hidden markov models in finance* (Vol. 104). Springer Science & Business Media.

Mark, T., Lemon, K.N. and Vandenbosch, M., 2014. Customer Migration Patterns: Evidence from a North American Retailer. *Journal of Marketing Theory and Practice*, *22*(3), pp.251-270.

Montgomery, A.L., Li, S., Srinivasan, K. and Liechty, J.C., 2004. Modeling online browsing and path analysis using clickstream data. *Marketing Science*, *23*(4), pp.579-595.

Montoya, R., Netzer, O. and Jedidi, K., 2010. Dynamic allocation of pharmaceutical detailing and sampling for long-term profitability. *Marketing Science*, *29*(5), pp.909-924.

Moon, S., Kamakura, W.A. and Ledolter, J., 2007. Estimating promotion response when competitive promotions are unobservable. *Journal of Marketing Research*, *44*(3), pp.503-515.

Netzer, O., Lattin, J.M. and Srinivasan, V., 2008. A hidden Markov model of customer relationship dynamics. *Marketing Science*, *27*(2), pp.185-204.

Nylund, K.L., Asparouhov, T., and Muthen, B.O. (2007), "Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study," *Structural Equation Modeling*, 14 (4), 535-569.

Paas, L.J., Vermunt, J.K. and Bijmolt, T.H., 2007. Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(4), pp.955-974.

Padilla, N., Montoya, R. and Netzer O. 2016, Heterogeneity in HMMs: Allowing for heterogeneity in the number of states. Working paper, Columbia University.

Park, S. and Gupta, S., 2011. A regime-switching model of cyclical category buying. *Marketing Science*, *30*(3), pp.469-480.
Poulson, C. S. (1990). Mixed Markov and latent Markov modelling applied to brand choice behavior. International Journal of Research in Marketing, 7, 5–19.
Rabiner, L.R., Lee, C.H., Juang, B.H. and Wilpon, J.G., 1989, May. HMM clustering for connected word recognition. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on* (pp. 405-408). IEEE.
Richardson, S. and Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, *59*(4), pp.731-792.
Romero, J., Van der Lans, R. and Wierenga, B., 2013. A partially hidden Markov model of customer dynamics for clv measurement. *Journal of Interactive Marketing*, *27*(3), pp.185-208.
Schwartz, E.M., Bradlow, E.T. and Fader, P.S., 2014. Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. *Marketing Science*, *33*(2), pp.188-205.
Schweidel, D.A. and Knox, G., 2013. Incorporating direct marketing activity into latent attrition models. *Marketing Science*, *32*(3), pp.471-487.
Schweidel, D.A., Bradlow, E.T. and Fader, P.S., 2011. Portfolio dynamics for customers of a multiservice provider. *Management Science*, *57*(3), pp.471-486.
Scott, S. L. (2002). Bayesian methods for hidden Markov models: recursive computing in the 21st century. *Journal of the American Statistical Association*, 97, 337–351.
Seetharaman, P.B., 2004. Modeling multiple sources of state dependence in random utility models: A distributed lag approach. *Marketing Science*, *23*(2), pp.263-271.
Shachat, J. and Wei, L., 2012. Procuring commodities: first-price sealed-bid or English auctions?. *Marketing Science*, *31*(2), pp.317-333.
Shi, S.W. and Zhang, J., 2014. Usage experience with decision aids and evolution of online purchase behavior. *Marketing Science*, *33*(6), pp.871-882.
Shi, S.W., Wedel, M. and Pieters, F.G.M., 2013. Information acquisition during online decision making: A model-based exploration using eye-tracking data. *Management Science*, *59*(5), pp.1009-1026.
Smith, Aaron, Prasad A. Naik, and Chih-Ling Tsai., 2006. "Markov-switching model selection using Kullback–Leibler divergence." *Journal of Econometrics* 134, no. 2 (2006): 553-577.
Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde., 2002. "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, no. 4 583-639.
Stüttgen, P., Boatwright, P. and Monroe, R.T., 2012. A satisficing choice model. *Marketing Science*, *31*(6), pp.878-899.
Train, K.E., 2009. *Discrete choice methods with simulation*. Cambridge university press.
van der Lans, Ralf, Rik Pieters, and Michel Wedel, 2008a. "Competitive Brand Salience," *Marketing Science*, 27 (September), 922–31.
van der Lans, R., Pieters, R. and Wedel, M., 2008b. Eye-movement analysis of search effectiveness. *Journal of the American Statistical Association*, *103*(482), pp.452-461.

Vermunt, J.K. and Magidson, J. (2015). Upgrade Manual for Latent GOLD 5.1. Belmont Massachusetts: Statistical Innovations Inc.

Viterbi, A.J., 1967, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13 (2), pp 260–269.

Wang, M.,and Chan, D. (2011), "Mixture latent Markov modeling: identifying and predicting unobserved heterogeneity in longitudinal qualitative status change," *Organizational Research Methods*, 14 (3), 411-431.

Wedel, M. and Kamakura, W.A., 2000. Market Segmentation Conceptual and Methodological Issues. Kluwer Academic Publishing, Boston

Welch, L.R., 2003. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, *53*(4), pp.10-13.

Yamato, J., Ohya, J. and Ishii, K., 1992, June. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on* (pp. 379-385). IEEE.

Zhang, J.Z., Netzer, O. and Ansari, A., 2014. Dynamic targeted pricing in B2B relationships. *Marketing Science*, *33*(3), pp.317-337.

Zhang, J.Z., Watson IV, G.F., Palmatier, R.W. and Dant, R.P., 2016. Dynamic Relationship Marketing. *Journal of Marketing*.

Zucchini, W. and MacDonald, I.L., 2009. *Hidden Markov models for time series: an introduction using R* (Vol. 150). CRC press.