Chapter 13
Hidden Markov Models for Individual Time Series

Ingmar Visser, Maartje E. J. Raijmakers and Han L. J. van der Maas

This chapter introduces hidden Markov models to study and characterize (individual) time series such as observed in psychological experiments of learning, repeated panel data, repeated observations comprising a developmental trajectory etc. Markov models form a broad and flexible class of models with many possible extensions, while at the same time allowing for relatively easy analysis and straightforward interpretation. Here we focus on hidden Markov models with a discrete underlying state space, and observations at discrete times; however, hidden Markov models are not limited to these situations and some pointers are provided to literature on possible extensions.

Markov models have a long history in the social sciences; in psychology, for example, Markov models have been applied in analyzing language (Miller, 1952; Miller & Chomsky, 1963), in describing learning processes in paired associate learning (see Wickens, 1982, for an overview of models and techniques); in sociology, applications are mainly in the analysis of repeated measures of panel data (Langeheine & Van de Pol, 1990); similarly in political science (McCutcheon, 1987). Recently, extensions of Markov models, such as the hidden Markov model, have become increasingly popular, notably in speech recognition (Rabiner, 1989); in biology, in analyzing DNA sequences (Krogh, 1998); in econometric science, in analyzing changes in stock market prices and commodities (Kim, 1994); and finally, in machine learning and data mining (Ghahramani & Jordan, 1997). This chapter focusses on time series data from a psychological experiment in which both speed, i.e., reaction times, and accuracy are modeled simultaneously.

The rest of this chapter is organized as follows: In the next Section hidden Markov models are introduced in a conceptual fashion, and its relationship with other models is described. Following that, in Section “Likelihood, Parameter Estimation, and Inference” the main characteristics of the likelihood function, parameter optimization and

I. Visser (✉)
Developmental Psychology
University of Amsterdam
1012 ZA Amsterdam, The Netherlands
e-mail: i.visser@uva.nl

J. Valsiner et al. (eds.), *Dynamic Process Methodology in the Social and Developmental Sciences*,
DOI 10.1007/978-0-387-95922-1_13, © Springer Science+Business Media LLC 2009

inference are discussed, thereby introducing the hidden Markov in a more formal way. The next Section discusses analyses of two real life data sets thereby illustrating various characteristics of hidden Markov models and their potential to deal with individual time series. We end by summarizing and discussing the main results.

Hidden Markov Models: State Space and Transition Dynamics

Hidden Markov models, henceforth HMMs, consist of two main parts: the measurement model and the transition dynamics. The measurement model characterizes the states of the model, whereas the transition dynamics characterizes the dynamics between states over time. The states hence represent the construct of interest, whereas the transition dynamics represent the changes in the construct. For example, consider observation sequences DDDRDDRDDD and RRRDDDRDDD. It may be assumed that there are two underlying constructs, republican and democratic, which result in the corresponding voting behavior D and R (for Democrats and Republicans). The first observation sequence is likely to be from a democrat who voted R at only two occasions. It seems reasonable, given this particular observation sequence, to assume that this person is in the democratic state. The second observation sequence on the other hand is very different; here it seems reasonable that this person changes from the republican state to the democratic state, even though she/he has a single R preference on the seventh occasion of measurement.

This example illustrates an important aspect of HMMs, namely that there is no direct relationship between the state and the observations; in particular, a democratic person may vote R every now and then due to any number of reasons, e.g., measurement error, living in Florida, a temporary disapproval of the Democratic candidate etc. When analyzing such data, the question hence is to separate real change, i.e., change in the underlying variable, from the absence of change, in the face of measurement error (see e.g., Eid & Langeheine, 2003, for an application of this type). Below, the state space and the transition dynamics, and their relationship are discussed.

State Space and Measurement Model

In above example, the underlying state of interest was political preference of persons. The number of different possible preferences in this example determines the cardinality of the state space of an HMM that is used to model such data. Here it was assumed that people can be in either of two discrete states, democratic or republican. In this chapter, HMMs with a finite and discrete state space are considered.

Characteristic of HMMs is that the state space is not directly observable; if the example data above is read with D representing dry and R representing rain as observations on consecutive days, merely observing a D or an R can not inform us as to the state of the weather. Assuming that sunny and rainy periods are fairly stable and last for at least a number of days, observing rain on any given day is not sufficient evidence that the weather is in a rainy period; it may just happen to rain a bit during a period of otherwise stable and sunny weather (see Lystig & Hughes, 2002, for example of the analysis of rainfall data).

The relationship between the states of a hidden Markov model and the observations under consideration is governed by a measurement model. If this relationship is deterministic, i.e., if there is a one-to-one relationship between observations and states, the HMM reduces to a simple Markov model. For example, if voting behavior is taken to directly indicate overall political preference, then observing a D vote indicates someone is democratic. In social science research the relationship between observations and underlying constructs, states in this case, is not usually that straightforward. In particular, measurement error may obscure the relationship, or multiple observations are used to measure a construct, but none of them are assumed to correlate perfectly with the underlying construct. For example, if a person indicates that she/he likes going to parties, this may be taken as evidence of being extraverted, but such an item does not capture all there is to extraversion. A measurement model captures the relationship between observations and states.

Transition Dynamics

The second main part of interest in HMMs is the transition dynamics, that is, the model that governs the changes occurring in the states over time. For example, participants in a categorization learning experiment (Ashby & Ell, 2001) are assumed to pass through a number of stages. At the outset of the experiment, they have no knowledge of the task, and hence their performance is expected to be at chance level. In an HMM this can modeled by means of a guessing state, in which the probability of providing a correct answer is 0.5. At the end of learning, participants have full knowledge of the task, and hence do not make any errors. In an HMM, this can be modeled by means of a so-called learned state, in which the probability correct is 1. Depending on their learning strategy, participants may pass through a number of intermediate states, in which they have partial knowledge of the task at hand. The model that only consists of the guessing and the learned states, is called the all-or-none model. The transition dynamics of this model is fairly simple; it is assumed that at each trial of the learning experiment, a participant has a fixed probability of learning the task. This probability is hence the probability of moving from the guessing state to the learned state; in the all-or-none model, this is called the learning parameter. See Wickens (1982) for an overview of such learning models. See Schmittmann, Visser, and Raijmakers (2006), Visser, Schmittmann, and Raijmakers (2007) for applications of hidden Markov models in categorization learning.

Above discussion illustrates a number of interesting characteristics of HMM states. The learned state in the above example is called an absorbing state: once it is entered, one cannot leave that state. This incorporates the assumption that once the task is mastered there is no unlearning. The guessing state on the other hand is a so-called transient state: the process passes through that state but eventually leaves the state and there is no probability of returning. The transition probabilities between states determine such characteristics of states.

A (hidden) Markov model is called ergodic when there are no absorbing states and each of the states can be reached from any other states. The rainy weather/sunny weather model forms an ergodic model; the process continues forever changing from rainy spells to sunny spells and back. The transition dynamics provides information about how stable each of these states is, e.g., whether sunny periods last longer than rainy periods.

In the discrete state hidden Markov models under consideration in this chapter, the transition dynamics consists of a matrix of transition probabilities between states. These transition probabilities can be made dependent on other variables. For example, in the weather case, the transition from a sunny period to a rainy period could be dependent on air pressure such that, when the air pressure drops, the probability of transitioning from a sunny period to a rainy period increases. In the Illustrations section an example of such dependence is provided.

Relationships with Other Models

HMMs are part of a larger class of stochastic process models or state-space models (Fahrmeir, Tutz, & Hennevogl, 2001). In particular, the HMMs that we consider here are state-space models with a discrete state-space and discrete equidistant measurement occasions. Hidden Markov models are formally equivalent to latent Markov models. However, in practice the literatures on each of these models are largely separated. HMMs originate from engineering applications such as speech recognition (Rabiner, 1989), whereas latent Markov models originate in sociology and political science (Langeheine & Van de Pol, 1990). HMMs are typically applied to long univariate time series, such as speech streams or stock market prices. In contrast, latent Markov models were considered as extensions of latent class models (McCutcheon, 1987) with repeated measurements. In latent class models, the goal is to classify persons into a finite number of distinct types. The latent Markov model then is applicable whenever, for example, questionnaires are administered repeatedly and the goal is to study changes in e.g., political preferences of large groups of people. Summarizing, in latent Markov models, the focus is on short multivariate time series with many cases, whereas HMMs are mostly applied to long (univariate) time series of a single process or individual. In this chapter, we consider both a univariate and a bivariate time series. The next section provides a formal definition of hidden Markov models along with likelihood function, which is used to estimate parameters and draw inferences on (relative) goodness-of-fit of competing models.

Likelihood, Parameter Estimation, and Inference

Below, we first give a description of the parameters and distribution functions that together constitute a hidden Markov model. After that, the likelihood and parameter estimation are discussed. To be able to interpret the examples in the Illustrations section, the description of the parameters is essential. However, the full description of how to compute and optimize the likelihood may be skipped by non-technical readers.

Formally a hidden Markov model consists of the following elements (here we adopt notation by Rabiner, 1989):

1. A finite state space S with states S_i, $i = 1,...,n$.
2. A transition model A providing transition probabilities a_{ij}.
3. A measurement model for each state in S, denoted by B_i, $i = 1,...,n$, which relates the state to the observation O.
4. The initial state probabilities π_i, $i = 1,...,n$.

Here n is the number of states of the model, i.e., the number of possible values the state variable S_t can assume. π denotes the initial state distribution at $t = 1$, which is a probability vector with $\Sigma_i \pi_i = 1$. Next, $B_i(.)$ is the distribution of the responses or observations O conditional on the current state $S_t = i$. For example, for a binary item O we have $b_i(O = 1) + b_i(O = 2) = 1$, for each i. Finally, a_{ij} is the transition probability of moving from state $S_t = i$ to state $S_{t+1} = j$, which is written as a probability matrix A. That is, for each state S_i the transition probabilities sum to one, $\Sigma_j a_{ij} = 1$.

To fix ideas, consider an HMM with two states as depicted in Fig. 13.1. The data to be modeled are the observations O; in particular, there are two observations at each measurement occasion, O^1 and O^2, hence we are considering a bivariate time series. The underlying states are S_1 and S_2. The transition probabilities are denoted a_{11}, a_{12}, a_{21}, and a_{22}, respectively. The initial state probabilities π are the probabilities that the process starts in a particular state S_i (not depicted in Fig. 13.1).

The model in Fig. 13.1 could be a model for the weather example discussed above. One state then corresponds to sunny weather, and the other to rainy weather. The transition probabilities are interpreted as the stability of the weather; for example, the closer a_{11} is to 1, the more stable the corresponding weather state is.

A core assumption of Markov models is that the current state only depends on the previous state, and not on the entire history of previous states. Formally this (first-order) Markov assumption is expressed as:

$$p(S_t | S_1, S_2, \ldots, S_{t-1}) = p(S_t | S_{t-1}), \tag{1}$$

and hence the conditional distribution of S_t only depends on S_{t-1}, and not on S_{t-2} etc. Here S_t is short for $S_t = i$, meaning the process is in the i-th state of the model at time t of measurement. This Markov assumption is customarily made in many applications. If the assumption is not met, it is usually possible to increase the number of latent states in such a way that the assumption is met. This amounts to fitting so-

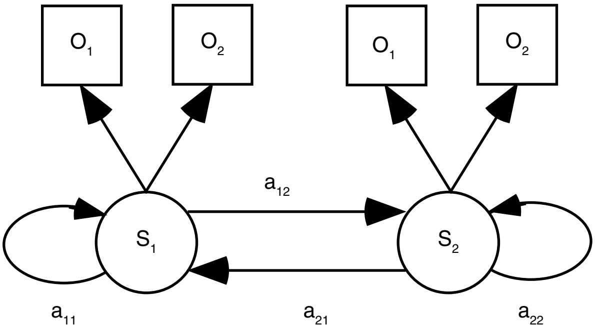

Fig. 13.1 Hidden Markov model with 2 states

called higher-order latent Markov models (see Langeheine & Van de Pol, 2000, for discussion).

To summarize, the parameters of the hidden Markov model are the following. First, the vector π, the initial state probabilities. Second, the parameters that determine the transition probabilities a_{ij}, either multinomial parameters that sum to unity, i.e., $\Sigma_j a_{ij} = 1$ or other parameterizations thereof. In our second illustrative example, we specify a multinomial logistic distribution for the transition probabilities, such that covariates can be included that influence those probabilities. Third, the parameters of the observation distributions $B_i(.)$, which specify the relationship between the unobserved states S_t and the observations O_t. For example, in a gaussian HMM, each $B_i(.)$ has two parameters, the mean and standard deviation of a gaussian or normal distribution. Below the full likelihood function of the HMM is specified along with recursive functions for efficiently computing the likelihood. It is not necessary to understand all the details of computing and optimizing the likelihood to be able to interpret the parameters in the Illustrations section. Hence, the non-technical reader may wish to skip this section and refer back to it when needed.

Likelihood

The data we are considering here has the general form $\boldsymbol{O}_1,...,\boldsymbol{O}_T$, where $\boldsymbol{O}_t = O^1,...,O^m$, an m-variate observation at time t. Using above notation, the likelihood for such a time series can be expressed as follows:

$$L(O_T|\lambda) = \Sigma_S \pi_1 B_{S_1=i}(O_1) \Pi_t a_{S_{t-1}=i,S_t=j} B_{S_t=j}(O_t), \tag{2}$$

where λ is the parameter vector containing the parameters to model π, A, and B. The sum runs over all possible sequences $S_1,..., S_T$ of the latent or hidden state sequence, and the product runs from $t = 2$ to T.

When local independence is assumed among the items, the distribution functions $B_i(\mathbf{O}_t)$ can be factored as follows:

$$B_i(O_t) = \Pi_{j=1...m} B_i(O^j). \tag{3}$$

The assumption of local independence is very common in so-called latent variable models. Some even claim that local independence is *the* defining feature of latent variable models (see Bollen, 2002, for discussion). In many applications, local independence is indeed a reasonable assumption. A particular state, e.g., a knowledge state, an economical state or an attitude, is supposedly the common cause of the observed variables. This assumption means that the underlying variable that we are interested in, the states in the case of Markov models, causes the observed variables to have the values that they have. As a consequence, when conditioning on that underlying variable, the observed variables are independent, which is expressed in the local independence assumption. Throughout the rest of the chapter, local independence is assumed for models that are fitted.

Note that so far we have not mentioned any particular assumptions about the distributions $B_i(.)$. The state variable S is distributed multinomially by the parameter vector π, and so are the transition probabilities. For the observation distributions $B_i(.)$, there is no compelling reason to make any assumptions. As a consequence, they can be any estimable density function, including the multinomial distribution for categorical responses, but also the gaussian distribution if, for example, reaction times are included. In such a case, when there is a categorical response and a continuous response, the local independence assumption proves very valuable, because there is no need to deal with the possible correlation structures among these different variables. In the second example in the Illustrations section, we use mixed variables in this way.

Parameter Estimation and Inference

To prevent underflow and to make computing the likelihood more efficient, below a recursive algorithm is described developed by Lystig and Hughes (2002). This estimation procedure differs in three ways from the standard latent Markov estimation procedures. First, scaling is used to prevent underflow problems. Second, the raw data likelihood is computed instead of likelihood based on a sufficient statistic, such as a contingency table. An advantage of this is that missing data can be easily dealt with in a similar vein as is done in for example the Mx-program (Neale, Boker, Xie, & Maes, 2003). Third, a recursive scheme is used to compute the likelihood which is known as the forward algorithm such that the number of computations is limited.

To deal with underflow problems, the joint probability of the data is first rewritten as a product of conditional probabilities as follows (Lystig & Hughes, 2002):

$$L_T = p(O_{1,\dots,}O_T) = \Pi_{t=1\dots T}\, p(O_t|O_{1,\dots,}O_{t-1}), \tag{4}$$

where $p(O_1|O_0) := p(O_1)$. Note that rewriting the joint likelihood in this way does not depend on any particular assumption of (latent) Markov models. Therefore, the dependence on the model parameters is dropped in the above equation. The log-likelihood can now be expressed as:

$$l_T = \sum_{t=1\ldots T} log\left[p\left(O_t \mid O_{1,\ldots,} O_{t-1}\right)\right]. \tag{5}$$

This formulation of the likelihood prevents underflow to occur for long time series because the conditional probabilities $p(O_t|O_{1,\ldots,}O_{t-1})$ are computed, rather than the usual probabilities $p(O_{1,\ldots,}O_T)$. Next we need to compute these conditional probabilities.

The so-called forward algorithm can be used to compute the likelihood (Baum & Petrie, 1966). Below version of the algorithm is due to Lystig and Hughes (2002). Define the forward recursion variables as follows:

$$\varphi_1(j) = p(O_{1,}S_1 = j) = \pi_j B_j(O_1). \tag{6}$$

$$\begin{aligned}\varphi_t(j) &= p(O_{t,}S_t = j|O_1, \ldots, O_{t-1})\\ &= [\Sigma_{i=1\ldots n}\varphi_{t-1}(i)a_{ij}B_j(O_t)] \times (\phi_{t-1})^{-1}\end{aligned} \tag{7}$$

where $\phi_t = \Sigma_{i=1\ldots n}\varphi_t(i)$. Note first that the sum over n in Eq. (7) is simply an enumeration of all the states of the model. Here $\varphi_t(j)$ is the probability of observing O_t in state S_j conditional on having observed $O_1,\ldots,O_{t-1}$. Hence, $\phi_t = \Sigma_{i=1\ldots n}\varphi_t(i)$ is the probability of observing O_t conditional on having observed $O_1,\ldots,O_{t-1}$. The recursion includes an efficient enumeration of all possible latent state sequences. Note that computing the Φ_t takes in the order of S^2 computations, and hence computing the likelihood takes S^2T computations. Writing out Φ_t for $t=3$ makes explicit its relationship with Eq. (2):

$$\phi_3 = \left\{\Sigma_{i3}\left[\Sigma_{i2}\left(\Sigma_{i1}p_{i1}B_{i1}(O_1)a_{i1i2}\right)B_{i2}(O_2)\right]a_{i2i3}B_{i3}(O_3)\right\}(\phi_1 \times \phi_2)^{-1} \tag{8}$$

Note that the triple summation between braces is identical to Eq. (2) for the case that $t=3$. The multiplication of this term by $(\phi_1 \times \phi_2)^{-1}$ takes care of the scaling at each time point to prevent underflow.

Combining $\phi_t = p(O_t|O_1, \ldots, O_{t-1})$, and Eq. (5) gives the following expression for the log-likelihood:

$$l_T = \sum_{t=1\ldots T} log\,\phi_t. \tag{9}$$

Parameters for the hidden Markov model can be estimated by optimizing the log-likelihood. Baum and Petrie (1966) provided an EM algorithm for doing so. Because Lystig and Hughes (2002) also provide gradients of the parameters for the log-likelihood, instead of using an EM algorithm, also direct optimization may be applied efficiently. The above algorithm for computing the log-likelihood and the gradients are implemented in the depmix package in the R-language for statistical computing (R Development Core Team, 2008). Depmix uses direct optimization of the log-likelihood with a Newton type algorithm using the gradients whenever they are available (Visser, 2008). An advantage of using direct optimization over the EM algorithm is that it is straightforward to deal with box constraints on parameters and general linear constraints between parameters. Depmix uses the Rdonlp2 package to handle such constraints (Tamura, 2007).

A second package depmixS4 also implements hidden Markov models and extends these with more general measurement models. In particular, depmixS4 fits Markov mixtures of generalized linear models (Visser & Speekenbrink, 2008). More about possible extensions of ordinary hidden Markov models in the Discussion section.

Determining goodness-of-fit of hidden Markov models is a notoriously difficult task, but see for example Altman (2004) for some recent developments. Instead of determining absolute goodness-of-fit, such as χ^2-measures for contingency tables (Wickens, 1989), in the current chapter we use relative goodness-of-fit measures to compare various candidate models with one another. The Akaike and Bayesian Information Criteria are the most common such measures that are used (Akaike, 1973; Schwarz, 1978). These criteria provide a way of balancing a measure of the goodness-of-fit of a model, such as the log likelihood with a measure of the parsimony of the model, such as the number of parameters. In the illustrations below, these two criteria are used. Lower AICs and BICs indicate better fitting models. In addition, we use the log likelihood ratio whenever this is applicable, i.e., whenever models are nested. In comparing nested models, the log likelihood ratio between two models has a known distribution, the χ^2-distribution with *df* the number of constrained parameters, and can hence be used to test models (Wald, 1943).

Illustrative Applications

Perth Water Dams

In this section we provide a brief example of a so-called left-right hidden Markov model; that is, a model with a number of transient states that can only be accessed in one direction, ending in a final absorbing state. The aim of this example is to show that there is a number of stages underlying the data which are best characterized as discrete stages with sudden transitions between them rather than a gradual, continuous change. Similar models have, for example, been used in studying development of math skills (Collins & Wugalter, 1992) and in medical applications (Reboussin, Reboussin, Liang, & Anthony, 1998); Kaplan (2008) provides an overview of such models, that are called stage-sequential models in the developmental psychology literature. The important difference between those applications and ours is that they considered only a few measurement occasions and many participants, whereas here we study a single time series.

Data

The data that are used in this illustration concern the water inflow into a number of dams surrounding Perth, Western Australia. The data are depicted in Fig. 13.2. Measurements are annual yearly totals of water inflow into a number of dams in the area of Perth. The data are kindly provided by the Water Corporation of Western Australia and concern the years 1911–2005 (*Water Corporation of Western Australia*).

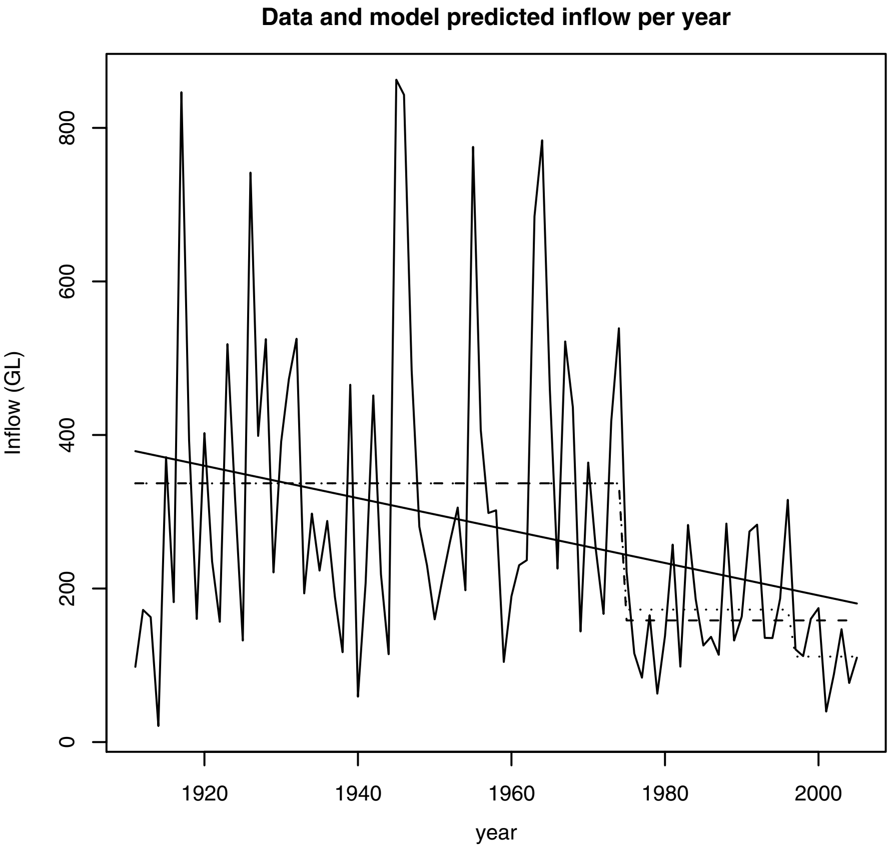

Fig. 13.2 Perth dams water inflow data in gigaliter per year from 1911 to 2005. The solid line represents the data; the solid straight line is the prediction by a simple linear model; the dashed line represents the predicted values by a 2-state (1 switch point) hidden Markov model; the dotted line represents the predictions by a 3-state (2 switch points) hidden Markov model. The first switch of the 3-state model occurs in the same year as the switch in the 2-state model

Models

The main point of interest in these data is whether there are significant trends indicating a decrease in water inflow. We consider two possible hypotheses here: First, there is a decrease, and it is linear; second, there is a decrease, and it is stepwise, i.e., there are sudden changes rather than gradual changes. In economic science for example this is referred to as a regime change (Kim, 1994), and in psychology these sorts of changes are referred to as phase transitions or catastrophic changes (van der Maas & Molenaar, 1992).

Here we compare three models, a linear model, and hidden Markov models with either 2 or 3 states to allow for either 1 or 2 change points in the data. In terms of parameters of the earlier given example model in Fig. 13.1, this means that transi-

Table 13.1 Goodness-of-fit measures for the Perth water data models. Lin denotes the linear model, and 2 and 3 denote the 2- and 3-state models respectively

Model	logl	AIC	BIC	nfree
Lin	634.29	1272.58	1277.69	2
2	612.34	1234.68	1247.45	5
3	611.01	1240.03	1263.01	9

Note: Model indicates the type of model that is fitted, lin for linear model, 2 and 3 for the 2- and 3-state models respectively; logl is the log likelihood, AIC and BIC are the Akaike and Bayesian information criteria respectively; nfree denotes the number of free parameters estimated in each model.

tion parameter a_{21} is equal to zero. In other words, there is a progression from state 1 to state 2 at some point in time but no possibility of going back to that earlier state. In the 3-state model, transition parameter a_{32} is equal to zero. Model goodness-of-fit statistics are provided in Table 13.1.

As can be read in Table 13.1, the 2- and 3-state models have much better goodness-of-fit statistics than the linear model, supporting the conclusion that change in water inflow is sudden rather than gradual. The difference between the 2- and 3-state models is relatively small and hence only future data can clarify whether there is a second change point or not. Both the 2- and 3-state models agree exactly on the time of the first change point in the mid-seventies. The extent to which these changes reflect effects of climate change has been an issue of debate (see *Marine and Atmospheric Research*, for various reports on this). All three fitted models' predictions are plotted in Fig. 13.2.

Speed-Accuracy Trade-Off

In this section we provide an example of an HMM that is used to analyze a single time series resulting from a reaction time experiment in which the speed-accuracy trade-off (SAT) was manipulated. In this example, we extend the traditional latent Markov model in two important ways. First, the latent states can have mixed indicators, e.g., a gaussian and a binomial indicator, for the reaction time and the accuracy of a trial, respectively. In much experimental research, reaction times and accuracy of trials are analyzed separately, whereas they are known to be dependent. Using both as indicators of a single latent variable allows us to explore the relationships that exist between them. In the application that we present, the relationship between speed and accuracy is explored. Second, we use covariates on the transition parameters to test the effects of experimental manipulation. More generally, the use of covariates can help account for heterogeneity, either between cases, or within a single case over time.

Heterogeneity in time can be accommodated by specifying separate distribution functions for each measurement occasion. In this general case, the distributions depend on t and we write e.g., $a_{ij} = a_{ij}(t)$. As a result, the number of parameters depends on the number of measurement occasions, which quickly becomes complicated when analyzing long time series. In particular, in the application that we consider, with only a single time series this is not feasible. Therefore, none of the distributions depends on t in this general way. Instead, we deal with heterogeneity in time in a more parsimonious way by specifying parameters of distributions to be functions of time-dependent covariates z_t. When that is the case, we write e.g., $a_{ij}(t) = P(S_t = \mathrm{j}|S_{t-1} = i,z_t)$. The transition probabilities are then modeled as a multinomial logistic regression (Agresti, 2002). In particular, we use a baseline category logit model for the transition probabilities from state i:

$$log(a_{ij}/a_{i1}) = \alpha_j + \beta_j z_t,\ j = 2, \ldots n, \tag{10}$$

where a_{ij} is the transition probability from state i to state j and n is the number of states in the model, and z_t a vector of covariates; in this example, the baseline category is 1, and the corresponding parameters α_1 and β_1 are set to zero. Note that this only works if the transition probability a_{i1} is not equal to zero; however, if this is the case, changing the baseline category can solve this problem. Recently, a latent Markov model with time-dependent covariates for the transition probabilities was presented in Chung, Walls, and Park (2007).

In this section we illustrate the use of HMMs by analyzing data from an experiment in which the speed-accuracy trade-off is manipulated by varying pay-offs for speed and accuracy in a reaction time experiment. Before presenting the data and a number of models, we briefly sketch the reasons for collecting these data.

Theoretical Background

The use of reaction times as behavioral measure in experimental psychology is pervasive. In experimental research on choice behavior it is common to analyze the reaction times only, and to consider accuracy data as a nuisance. Usually, accuracy scores are only analyzed to check whether they do not differ between conditions of an experiment. The random walk model (Laming, 1968) or diffusion model (Ratcliff, 1978) is especially suitable for simultaneously analyzing reaction times and accuracy of trials in experimental situations and it has been applied successfully in a large variety of experimental data.

The random walk model (RWM) for choice reaction times predicts a continuous trade-off between speed and accuracy; that is, it is assumed that accuracy and RT drop gradually in response to certain experimental manipulations, e.g., instructions to respond faster. An alternative to the RWM is a phase transition model (PhTM), which holds that participants switch between two modes of responding, the fast-guessing mode and the stimulus-controlled mode. In the fast-guessing mode, accuracy is at chance level and reaction times (RT) are short. In contrast, in the stimulus controlled mode, accuracy is close to 100% and RTs are relatively longer. The PhTM predicts that as pressure to respond faster and faster increases, participants switch to fast-guessing rather than respond at intermediate levels of accuracy and speed. Providing insight into the trade-off between speed and accuracy in choice reaction time tasks is important because the use of different strategies in the trade-off may threaten the validity of comparing RTs between, say, participants or experimental conditions.

The goal in the current illustration of HMMs is to test a number of predictions that differentiate the RWM and PhTM. There are two specific predictions that are explored here. First, the RWM predicts a single response strategy, which has a single parameter for adapting the trade-off, whereas the PhTM predicts the existence of two different strategies and a switching regime between those strategies. Second, in addition to the existence of two states in the PhTM, it also makes specific predictions about the switching regime dependent on the changes in the pay-offs for speed and accuracy. In particular, it predicts a certain asymmetry in the switching process between the fast-guessing state and the stimulus-controlled state. These two hypotheses translate into specific hypotheses about hidden Markov models that are outlined below in the modeling section.

Data

Figure 13.3 depicts the first 168 trials of an experiment in which the SAT was manipulated by continuously varying the pay-off for accuracy. The data at each trial consists of a reaction time (log-transformed to make the distribution normal) and whether the trial was correct or not. The data result from a lexical decision task. The third part of the Figure depicts the relative pay-off for accuracy that was manipulated experimentally. The data that are analyzed here are from a single participant (two other participants were tested with similar results). This pay-off increases and decreases over trials with the aim that the participant adapts his behavior accordingly. The depicted data in Fig. 13.3 is the first part of 168 trials; the full data of this participant has two further series of 134 and 137 trials respectively. These data form a subset of the data from Experiment 2 in van der Maas, Dutilh, Visser, Grasman, and Wagenmakers (2008).

Models

The main aim is to test the hypothesis whether there are two modes rather than one in these data; that is, the hypothesis is whether a gradual decrease in the pay-off for accuracy leads to a gradual decrease in accuracy of responding (and a corresponding decrease in RTs), or alternatively, whether a decrease in the pay-off leads to a sudden switch in the mode of responding, from stimulus-controlled responding (slow and accurate) to fast-guessing (fast responding at chance level). Furthermore, if there are two or more modes, it is interesting to find out the transition dynamics between the modes and in particular if and how those depend on the covariate, i.e., the pay-off for accuracy.

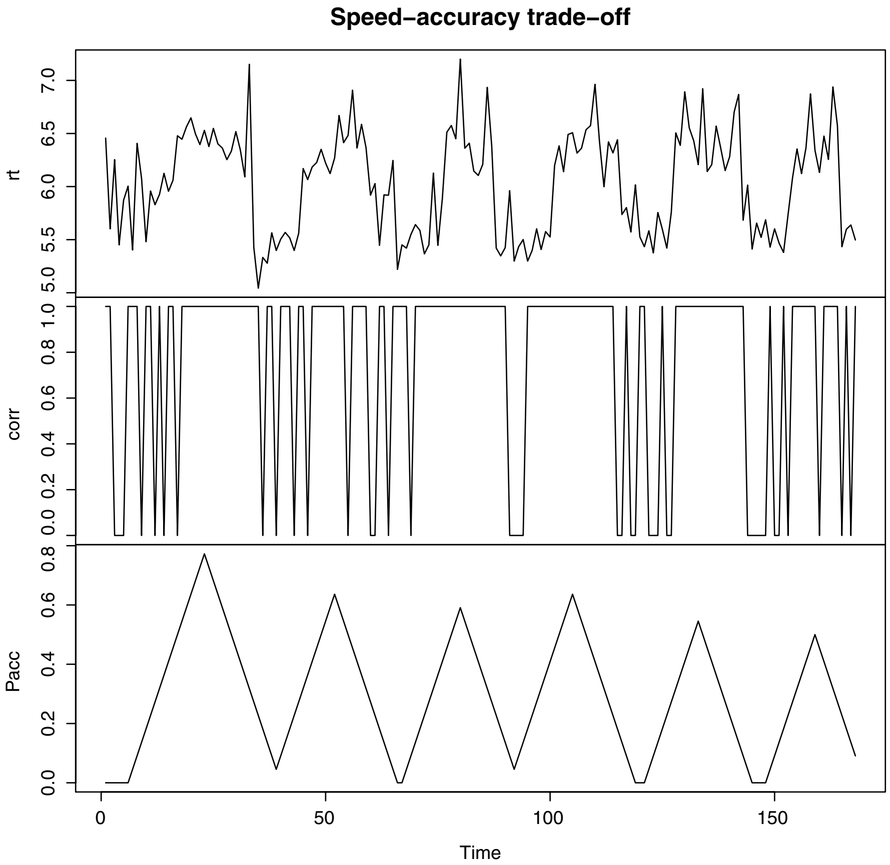

Fig. 13.3 Speed-acurracy trade-off data. Upper panel: reaction times (note that the RTs are log-transformed); middle panel: accuracy per trial (corr); lower panel: pay-off for accuracy (Pacc)

To test the first hypothesis, i.e., that there two or more modes in the data, a number of HMMs with 1–3 states are fitted to the data. Models are fitted with and without Pacc, the pay-off for accuracy, as covariate on the RTs and corr (accuracy) variables in the data. That is, for example, in each state the RTs are modeled with a linear regression model with Pacc as predictor. Similarly, the accuracy variable corr is modeled with Pacc as a predictor in a binomial logistic regression (Agresti, 2002).

Table 13.2 presents the goodness-of-fit statistics for each of these models, with a p indicating that Pacc was included as a covariate on the RT and corr variables in each state.

The AIC and BIC are reported here. The Table also includes the likelihood ratio tests for adding the predictor Pacc to the models.

As can be seen in Table 13.2, the 2- and 3-state models have much lower AIC and BIC values than the 1-state model, both with and without Pacc as covariate on the

Table 13.2 Goodness-of-fit measures for 1- to 3-state models with and without direct effects of Pacc on RT and corr

Model	logl	AIC	BIC	nfree	llr	df (*p*)
1	−554.64	1115.27	1127.53	3		
1p	−417.42	844.85	865.27	5	274.4	2 (0)
2	−296.11	610.22	646.98	9		
2p	−291.93	609.86	662.96	13	8.36	4 (0.079)
3	−265.52	565.04	634.48	17		
3p	−263.00	572.01	665.95	23	5.03	6 (0.54)

Note: Model indicates the type of fitted model, see the text for details; logl denotes the log likelihood; AIC and BIC denote the Akaike and Bayesian information criteria respectively; nfree denotes the number of freely estimated parameters of the model; llr denotes the log likelihood ratio with respect to the previous model for the models that include Pacc as covariate (models with p); df denotes the degrees of freedom for the log likelihood ratio test and between parentheses the p-value for the test is given.

responses. It is hence safe to conclude that the process consists of multiple modes of responding.

The 2- and 3-state models both contain a state which is best described as a ‘stimulus controlled’ mode of responding with relatively slow RTs and highly accurate responding. The 2-state model additionally has a fast-guessing state with fast RTs and accuracy around chance level. In the 3-state models, there are two instead of one fast-guessing state, with one state having very fast responses, and the second state having slightly slower responses. The reason that the 3-state model is slightly better than the 2-state model could be due the fact that we modeled the reaction times using a log-normal distribution, which may not be optimal. Because of this, below we further explore the 2-state model and extensions thereof rather than the 3-state models. First, however we present the parameter values of the 2-state model without covariates.

The initial state probabilities have values $p_1 = 0.99997$ and $p_2 = 0.00003$; in other words, the process starts in state S_1 with overwhelming probability. The transition matrix has values:

$$\begin{aligned} a_{11} &= 0.916 a_{12} = 0.084 \\ a_{21} &= 0.101 a_{22} = 0.899, \end{aligned} \tag{11}$$

from which it can be seen that both states are very stable, i.e., the probability of staying in either state, a_{11} and a_{22} is around 0.9. The measurement models for each of the states have the following parameters. State S_1 has a mean reaction time of 6.36 (SD = 0.24). Note that the RTs are log-transformed, so this mean is equivalent to 595 ms. The accuracy in state S_1 is equal to 0.90, which identifies this state as the stimulus-controlled state with an accuracy close to unity. State S_2 has a mean RT of 5.52 (SD = 0.20) and an accuracy of 0.53, which identifies this state as the fast-guessing state with accuracy around chance level and RTs that are on average much faster than in the stimulus-controlled state. The RTs in this state are 249 ms on average.

Table 13.3 Goodness-of-fit measures for 2-state models for the speed-accuracy data. See the text for details

Model	logl	AIC	BIC	nfree	llr	*df* (*p*)
2	−296.11	610.22	646.98	9		
2p	−291.93	609.86	662.96	13	8.36	4 (0.079)
2ptr	−248.9	519.94	564.87	11	94.27	2 (0)
2ptr-constr	−249.21	516.43	553.19	9	0.48	2 (0.78)
2ptr-hyst	−250.51	517.03	549.71	8	2.61	1 (0.11)
2ptr-nohyst	−277.11	568.21	596.81	7	55.8	2 (0)

Note: Model denotes the type of model that is fitted, see the text for details; logl denotes the log likelihood, AIC and BIC denote the Akaike and Bayesian information criteria, respectively; nfree denotes the number of free parameters in the model; llr denotes the log likelihood ratio of the model with respect to the baseline model (i.e., model 2 for models 2p, 2ptr, and 2ptr-constr, and model 2ptr-constr for the remaining models, also see the text); *df* denotes the degrees of freedom for the log likelihood ratio test and between parentheses is the corresponding *p*-value for the χ^2 -distribution.

The next model we fit is an extension of the above described model: it is a 2-state model in which the transition dynamics depend on the covariate Pacc, the pay-off for accuracy. In above fitted models, this pay-off was included as covariate on the responses (RT and corr) directly. As can be seen in Table 13.2 in the 2- and 3-state models, the likelihood ratio tests indicate that adding Pacc as a predictor for the responses does not significantly improve goodness-of-fit of these models. For example, in the 3-state model, the log likelihood ratio is 5.03 with $df = 6$ resulting in $p = 0.54$. According to the PhTM of the SAT, it is more plausible that Pacc influences the transitions between the fast-guessing and the stimulus-controlled mode of responding rather than influencing the responses directly. Consequently, Pacc is included as a covariate on the transition probabilities in the next set of the models that we fitted.

In Table 13.3, the goodness-of-fit statistics for this model (denoted 2ptr in the Table) are given along with the 2-state model without covariates (2) and the 2-state model with Pacc as predictor for the responses (2p). Note that these latter two models are identical to the ones presented in Table 13.2 and are presented here for purposes of comparison. The 2ptr model fits the data much better than either the 2-state model (2) without covariates or the 2-state model with a covariate on the responses (2p).

Two further constraints are tested in this model. First, the initial probability for starting in the stimulus controlled state is estimated at 0.999 and hence it is suspected that it is not significantly different from unity. Second, the probability for accuracy in the fast-guessing mode is 0.525 and it is interesting to test whether this differs significantly from 0.5, chance level in the task. The next model we fitted incorporates these two constraints; in Table 13.3, this model is indicated as ‘2ptr-constr’. The log likelihood ratio test (in column llr in the Table) indicates that these constraints do not lead to significant decrease in goodness-of-fit and are hence reasonable. These constraints are therefore kept in the following models.

The second hypothesis that is interesting to test concerns the phenomenon of hysteresis, i.e., whether the switching process between the states is asymmetric rather than symmetric which is a strong prediction by the phase transition model (see van der Maas & Molenaar, 1992, for an example of this in developmental psychology). Hysteresis means that each of the states has inherent stability, which means that switching from the guessing state to the stimulus controlled state occurs at a different value of the control parameter (Pacc in this case) than when switching the other way around. In terms of hidden Markov model parameters, this means that the transition matrix is asymmetric and that the values of the intercept parameter corresponding with the influence of Pacc on the transition probabilities should be different between the two states of the model. Hence, the PhTM predicts that the intercepts [α_i in Eq. (10)] of the regression models relating Pacc to the transition probabilities are different in each of the states, but not the slope parameters (β_i). Consequently, two more models were fitted, one in which only the β's were constrained to be equal, i.e., $\beta_1 = \beta_2$. This model is called the hysteresis model, because the α's are different for each state. In the Table, this model is denoted as 2ptr-hyst. In the second model, the α's are constrained to be equal, i.e., $\alpha_1 = \alpha_2$ to test the hypothesis of hysteresis. This final model is denoted 2ptr-nohyst in Table 13.3. Included in Table 13.3 are the log likelihood ratio tests of these models relative to the 2ptr-constr model as that was the best model so far. As can be seen from those tests, the 2ptr-hyst model is tenable but the 2ptr-nohyst model is not. Hence, the best model for these data is a model that incorporates hysteresis.

The measurement model parameters for state 1 are: a mean RT of 5.52 (SD = 0.21) and an accuracy of 0.5 (note that this parameter was constrained at that value). State 2 has a mean RT of 6.40 (SD = 0.24) and a mean accuracy of 0.91. Hence, state 1 is the fast-guessing state and state 2 the stimulus-controlled state. The transition model for state 1, the fast-guessing state, has parameters $\alpha_2 = -5.33$ and $\beta_2 = 12.65$ (remember that α_1 and β_1 are both equal to zero as this is baseline category logistic model). When Pacc is zero, this means that the transition probability a_{11}, i.e., the probability of remaining in the fast-guessing state is equal to 0.995. The α and β parameters for the transition model of state 2 are −2.42 and 12.65 respectively. When Pacc is 1 (the maximum value in the experiment), this means the transition probability a_{22} is 0.9999, in other words virtually equal to one.

The transition dynamics of the model are further illustrated in Fig. 13.4. Depicted is the probability of transitioning from the fast-guessing state to the stimulus-controlled state as a function of the value of Pacc (solid smooth line). The dotted smooth line is the probability of staying in the stimulus controlled state as a function of Pacc. The probabilities are computed from model 2ptr-hyst in Table 13.3 with the parameter values that are given above. The Figure clearly shows the separation between the curves which indicates that switching from the fast guessing state to the stimulus controlled state occurs at higher values of the pay-off for accuracy than switching in the other direction.

Plotted over the model predicted transition functions are the observed RTs as function of Pacc (lines with circles and triangles). The solid line with circles indicates average RTs at different levels of Pacc during runs of trials at which Pacc is increasing (that is, when a switch is expected from the fast-guessing mode to the stimulus-controlled mode); the dotted line with triangles indicates average RTs at different levels of Pacc during runs of trials in which Pacc is decreasing, that is when

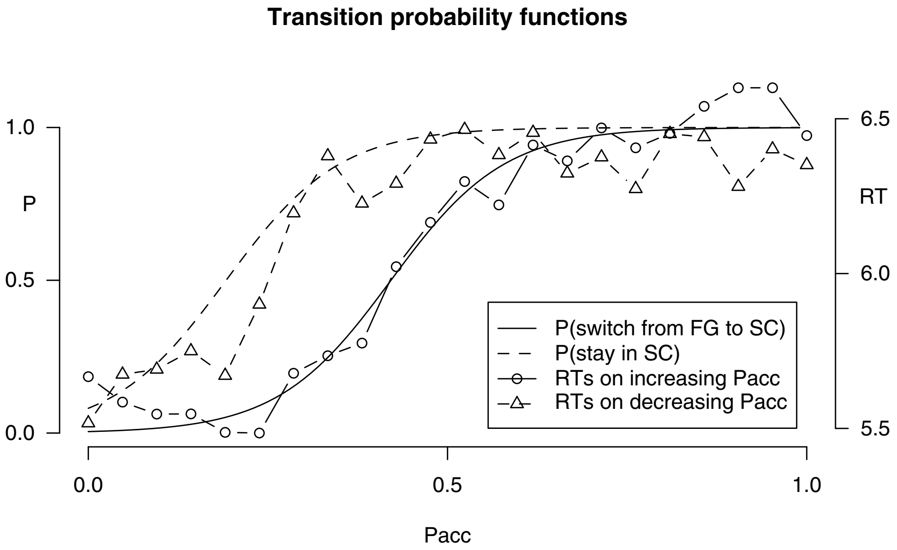

Fig. 13.4 Hysteresis in the transition probabilities (computed from the fitted model 2ptr-hyst) between states. On the left hand side scale is the transition probability. The right hand scale is for the reaction times also showing hysteresis in the observed data. See the text for details on the model and the reaction times

a switch is expected from the stimulus controlled mode to the fast-guessing mode. Here hysteresis is evident by noting that at increasing values of Pacc, the switch to slower RTs (corresponding with more accurate responding) occurs at higher values of Pacc than the switch in the other direction, i.e., when Pacc is decreasing.

Conclusion

The SAT is an important phenomenon in experimental research because strategic differences between participants may influence conclusions reached from such research. The experimental data that were analyzed here clearly indicate that there are multiple modes in responding to a simple choice reaction time task depending on the reward that participants get for responding fast versus accurate. This is in contradiction with a popular and often model for analyzing reaction time data, the random walk model. HMMs have been shown to be a useful tool to discriminate between these models, thereby showing that participants switch between two modes of responding depending on the pay-off for accuracy as a covariate.

Discussion

We have introduced hidden Markov models as an important tool in studying processes of change, in particular changes that occur suddenly rather than gradual. Stepwise or stage-wise changes are an important characteristic of many theories, including theories in experimental psychology in the analysis of reaction times, as we have illustrated. Other examples that we have studied earlier concern theories of development and learning that occur stage-wise rather than gradually. In learning, HMMs were applied to show that simple discrimination learning shows all-or-none learning processes rather than gradual stimulus-response strengthening (Raijmakers, Dolan, & Molenaar, 2001; Schmittmann et al., 2006; Visser et al., 2007). In other work, development on the balance scale task was analyzed using hidden Markov models to analyze the effects of feedback on learning (Jansen, Raijmakers, & Visser, 2007).

In this chapter we have focused on models for single time series. Such models form the basis for generalizations across participants if such is applicable. In our second example we have illustrated the use of time-varying covariates to describe changes in behavior. Also, in that example, we have illustrated the use of mixed indicators, that is, a combination of a binary and a gaussian variable measured concurrently. These extensions the hidden Markov model framework to be extremely flexible in analyzing processes of change.

Many extensions of hidden Markov models have been explored by various researchers. An important one is the possibility of dealing with continuous time measurements rather than equidistant time measurements as we have done. Bureau, Shiboski, and Hughes (2003) for example, present an analysis of disease outcomes, and Böckenholt (2005) presents an example in studying the time course of changes in emotional states.

The software framework that we used to fit the models, depmixS4 (Visser & Speekenbrink, 2008), offers a wide variety of possible models: Markov mixtures of generalized linear models. Hence, this includes the use of regression models within states of the hidden Markov model. Also, a variety of other distributions are available such as Poisson responses, gamma responses etc. The transition probabilities and the initial state probabilities within depmixS4 are as multinomial logistic distributions which allows for the possibility of including time-varying covariates on the transition parameters as we have shown in our second example. Moreover, the depmixS4 framework offers easy extensibility by the possibility of adding new response distributions.

Acknowledgement Ingmar Visser and Maartje Raijmakers were supported by an EC Framework 6 grant, project 516542 (NEST). Thanks to the *Water Corporation of Western Australia* for providing us with the Perth water dams data.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Academiai Kiado.

Altman, R. M. (2004). Assessing the goodness-of-fit of hidden Markov models. *Biometrics*, *60*, 444–450.

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *TRENDS in Cognitive Sciences*, *5*, 204–210.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, *37*, 1554–1563.

Böckenholt, U. (2005). A latent Markov model for the analysis of longitudinal datacollected in continuous time: States, durations, and transitions. *Psychological Methods*, *10*(1), 65–83.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605–634.

Bureau, A., Shiboski, S., & Hughes, J. P. (2003). Applications of continuous time hidden Markov models to the study of misclassiffied disease outcomes. *Statistics in Medicine*, *22*, 441–462.

Chung, H., Walls, T. A., & Park, Y. (2007). A latent transition model with logistic regression. *Psychometrika*, *72*(3), 413–435.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*(1), 131–157.

Eid, M., & Langeheine, R. (2003). Separating stable from variable individuals in longitudinal studies by mixture distribution models. *Measurement: Interdisciplinary Research and Perspectives*, *1*(3), 179–206.

Fahrmeir, L., Tutz, G., & Hennevogl, W. (2001). *Multivariate statistical modelling based on generalized linear models*. New York: Springer.

Ghahramani, Z., & Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning*, *29*, 245–273.

Jansen, B. R. J., Raijmakers, M. E. J., & Visser, I. (2007). Rule transition on the balance scale task: A case study in belief change. *Synthese*, *155*(2), 211–236.

Kaplan, D. (2008). An overview of Markov chain methods for the study of stage-sequential developmental processes. *Developmental Psychology*, *44*(2), 457–467.

Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, *60*, 1–22.

Krogh, A. (1998). An introduction to hidden Markov models for biological sequences. In S. L. Salzberg, D. B. Searls, & S. Kasif (Eds.), *Computational methods in molecular biology* (pp. 45–63). Amsterdam: Elsevier.

Laming, D. R. J. (1968). *Information theory of choice reaction times*. New York: Academic Press.

Langeheine, R., & Van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods and Research*, *18*(4), 416–441.

Langeheine, R., & Van de Pol, F. (2000). Fitting higher order Markov chains. *Methods of Psychological Research Online*, *5*(1), 32–55.

Lystig, T. C., & Hughes, J. P. (2002). Exact computation of the observed information matrix for hidden Markov models. *Journal of Computational and Graphical Statistics*, *11,* 678–689.

McCutcheon, A. L. (1987). *Latent class analysis* (No. 07-064). Beverly Hills: Sage Publications.

Miller, G. A. (1952). Finite Markov processes in psychology. *Psychometrika*, *17*, 149–167.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users (chap. 13). In R. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. New York: Wiley.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical modeling*. VCU Box 900126, Richmond, VA 23298.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria (ISBN 3-900051-07-0).

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE*, *77*(2), 267–295.

Raijmakers, M. E. J., Dolan, C. V., & Molenaar, P. C. M. (2001). Finite mixture distribution models of simple discrimination learning. *Memory & Cognition*, *29*(5), 659–677.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Reboussin, B. A., Reboussin, D. M., Liang, K.-Y., & Anthony, J. C. (1998). Latent transition modeling of progression of health-risk behavior. *Multivariate Behavioral Research*, *33*(4), 457–478.

Schmittmann, V. D., Visser, I., & Raijmakers, M. E. J. (2006). Multiple learning modes in the development of rule-based category-learning task performance. *Neuropsychologia*, *44*(11), 2079–2091.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Tamura, R. (2007). Rdonlp2: An R extension library to use Peter Spelluci's DONLP2 from R (version 0.3-1) [R package].

van der Maas, H. L. J., Dutilh, G., Visser, I., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2008). *Phase transitions in the trade-off between speed and accuracy in choice reaction time tasks*. Manuscript in preparation.

van der Maas, H. L. J., & Molenaar, P. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review*, *99*, 395–417. (Marine and atmospheric research. (n.d.). http://www.cmar.csiro.au/)

Visser, I. (2008). Depmix: An R-package for fitting mixtures of latent Markov models on mixed data with covariates (version 0.9.4) [R package]. Available at CRAN: http://cran.r-project.org

Visser, I., Schmittmann, V. D., & Raijmakers, M. E. J. (2007). Markov process models for discrimination learning. In K. van Montfort, H. Oud, & A. Satorra (Eds.), *Longitudinal models in the behavioral and related sciences* (pp. 337–365). Mahwah, NJ: Lawrence Erlbaum Associates.

Visser, I., & Speekenbrink, M. (2008). DepmixS4: S4 Classes for hidden Markov Models [R package latest version]. Available at CRAN: http://cran.r-project.org

Wald, A. (1943). Test of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.

Water Corporation of Western Australia. http://www.watercorporation.com.au/

Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology*. San Francisco: W. H. Freeman and Company.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.