Running head: MODELLING DISCRETE CHANGE

A Framework for Discrete Change

Ingmar Visser & Maarten Speekenbrink

Department of Psychology, University of Amsterdam

Correspondence concerning this article should be addressed to:

Ingmar Visser

Department of Psychology, University of Amsterdam

Roetersstraat 15

1018 WB Amsterdam

phone: +31 (20) 5256723

fax: +31 (20) 6390279

email: i.visser@uva.nl

Abstract

A class of models is developed for measuring and detecting discrete change in learning and development. The basic model for detecting such change is the latent or hidden Markov model. Traditionally, these models were restricted to categorical, mostly binary, observed variables, placing severe restrictions on possible measurement models. In this paper, the basic model is extended to include arbitrary distributions for the observed variables, including multi-variate distributions. Moreover, there is optional support to include time-varying predictors. In effect, this model consists of mixtures of general linear models with Markovian depencies over time to model the change process. In addition, transition parameters can be made to depend on covariates as well, such that the switching regime between states depends on characteristics of the individual or the experimental situation. The model is illustrated with an example of participants' learning in the weather prediction task.

A Framework for Discrete Change

Discrete change frequently occurs in learning and development: in learning concepts, in performance on Piagetian tasks, in discrimination learning and in conditioning. This chapter is concerned with detecting the time points of change in (individual) time series. We present a framework of dependent mixture models that can be used to differentiate between gradual and discrete learning events in individual time series data. Before presenting the model in formal terms and providing some illustrations, we first review some examples in which discrete change is found.

Piagetian developmental theory (REFERENCE??) assumes step-wise changes in the strategies that children apply in all kinds of tasks such as the conservation learning and the balance scale task. Van der Maas et al. (1992) developed a catastrophe model to describe phase transitions in learning and developmental processes. They applied the catastrophe model to learning in the conservation of liquid task (REFERENCE??) in which children have to judge relative volumes of liquid in glasses of different heights and widths. Young children tend to ignore the width dimension and hence always choose the glass with the highest level of liquid (REFERENCE??). Van der Maas et al. (1992) showed that there is a sudden transition to a new strategy in which the children also take the width of the glasses into account when judging the volume of liquids.

Jansen and Van der Maas (2001) applied the catastrophe model to development of strategies on the balance scale task (Siegler, 1981). In the balance scale task participants have to judge which side of a balance goes down when the number of weights and their distances to the fulcrum are varied over trials. Younger children tend to ignore the distance dimension in this task, and instead focus solely on the number of weights on each side of the fulcrum. This strategy for solving balance scale items is called Rule 1 (Siegler, 1981). Older children include the distance dimension in determining their response to balance scale problems; however, they only do so when the weight dimension does not differ between the sides of the balance, i.e., when the number of weights is equal on both sides of the balance scale. This strategy is called Rule 2 (Siegler, 1981).

Jansen and Van der Maas (2001) found clear evidence for stage-wise transitions between Rule 1 and Rule 2 by testing criteria that were derived from the catastrophe model. In particular, they found bimodal test scores and inaccessibility. The latter means that there are no in-between strategies: children apply either Rule 1 or Rule 2 and there is no in-between option. Jansen and Van der Maas also found evidence for hysteresis: the phenomenon that switching between strategies is assymetric. Children can switch from Rule 1 to Rule 2 and back, but this occurs at different trials. In particular, if the distance dimension in the balance scale problems is made more salient by increasing the distance difference between weights on either side of the balance scale, children may switch from Rule 1 to Rule 2. If subsequently the distance difference is decreased again, children may switch back to using Rule 1. Hysteresis is the phenomenon that this switch back occurs at a different value of the control variable, in this case the distance difference.

Also in animal learning and conditioning, evidence is found for sudden changes in response behavior (Gallistel et al 2004). In particular, in their study, evidence was found for sudden onset of learning: at the start of the learning experiment, the pigeons did not learn anything and performance was stable; after a number of trials, learning kicks in and there are large increases in performance. The interest here is in modeling the distribution of onset times: that is, the trials at which learning suddenly takes off. A similar interest in process onset times is found in addiction research. For example, (REFERENCE) study the age at which children start using alcohol and how this related to eventual outcomes in terms of addiction.

Sudden transitions in learning are also observed in simple discrimination learning paradigms in which participants learn to discriminate a number of stimuli based on a single dimension such as form or color. This kind of learning is referred to as all-or-none learning or concept identification learning. Raijmakers et al (2001) found evidence for different strategies applied by children when faced with such a learning task. Schmittmann et al (2006) reanalyzed their data using hidden Markov models to show that both strategies are characterized by sudden transitions in the learning process.

In above mentioned applications, the data consist mostly of a few repeated measurements administered to large groups of participants. The focus in the current chapter is rather on data that consist of many repeated measurements, or time series, observed in only a few participants or even just a single participant. For example, Visser, Raijmakers, & Van der Maas (2008) analyzed data from a single participant in an experimental task that manipulates the trade-off between speed and accuracy. The data consisted of three time series with each around 150 repeated measurements of both reaction time and accuracy. Below we provide examples of analyzing time series from single participants from two experiments; one from the Iowa Gambling Task and one from the weather prediction task. The interest in these tasks is to show that participants develop different strategies over time in responding to the stimuli and that the transition from one strategy to the next is a discrete event. Before providing these illustrations, below we give a formalization of dependent mixture models and a brief overview of the DepmixS4 package that was developed to specify and fit such models.

Dependent Mixture Models

In this section we describe a class of models which are especially suitable for describing and testing discrete change in (individual) time series data. The dependent mixture model is similar to, but slightly different from two other types of models that are in use for modelling discrete change: the hidden and the latent Markov models.

Markov models have been used extensively in the social sciences; for example, in analyzing language learning (Miller, 1952; Miller & Chomsky, 1963), in the analysis of paired associate learning (Wickens, 1982). In these models, the focus is on survey type data: a few repeated measurements taken in a large sample. Langeheine and Van de Pol (1990) discusses latent Markov models and their use in sociology ands political science (see also McCutcheon, 1987). Latent transition models, 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.

Hidden Markov models (HMM) tend to be used in the analysis of long univariate and individual timeseries. For example, HMMs are the model of choice in speech recognition applications (Rabiner, 1989). In biology, HMMs are used to analyze DNA sequences and in econometric science, to analyze changes in stock market prices and commodities (Kim, 1994).

The dependent mixture model that we propose here spans the range from latent Markov models for few repeated measurements with many participants to hidden Markov models for individual time series. In addition, the dependent mixture model includes multivariate responses. The dependent mixture model consists of the following elements:

1. S is a collection of discrete states
2. $S_t = \mathbf{A}S_{t-1} + \xi_t, \mathbf{A}$, a transition matrix

3. $O_t = \mathbf{B}(S_t) + \zeta_t, \mathbf{B}$, an observation density

The state space, which is a set of discrete states, captures the different states that the learning or developmental process under consideration can be in. In the balance scale example mentioned above, children are applying one of two possible strategies in responding to the items. The states are characterized by their corresponding observation densities. Using for example Rule 1 in the balance scale task leads to correct answers on some items and incorrect answers on others. A different strategy may lead to correct answers on some items and to guessing behavior on other items. In analyzing categorization learning data, in which participants learn to categorize a set of objects, a typical initial state is that participants are guessing because at the start of the task they have no knowledge of which features are important in categorization.

The transition matrix **A** describes the transitions between states over repeated measures or trials. This matrix summarizes the probabilities of transitioning from one state to another which represents learning or development. The transition model contains the Markov assumption:

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

which means that the current state (at time t) only depends on the previous state S_{t-1}, and not on earlier states.

The observation densities **B** form the measurement part of the model; these describe the distributions of the observations conditional on the current state.

Hence, these distributions characterize the state, and in our examples, these characterize the strategy that a participant is using at a given measurement occasion.

The log-likelihood of DMMs is usually computed by the so-called forward-backward algorithm (Baum & Petrie, 1966; Rabiner, 1989), or rather by the forward part of this algorithm. (Lystig & Hughes, 2002) changed the forward algorithm in such a way as to allow computing the gradients of the log-likelihood at the same time. They start by rewriting the likelihood as follows (for ease of exposition the dependence on the model parameters is dropped here):

$$L_T = Pr(\mathbf{O}_1, \ldots, \mathbf{O}_T) = \prod_{t=1}^{T} Pr(\mathbf{O}_t | \mathbf{O}_1, \ldots, \mathbf{O}_{t-1}), \tag{1}$$

where $Pr(\mathbf{O}_1|\mathbf{O}_0) := Pr(\mathbf{O}_1)$. Note that for a simple, i.e. observed, Markov chain these probabilities reduce to $Pr(\mathbf{O}_t|\mathbf{O}_1, \ldots, \mathbf{O}_{t-1}) = Pr(\mathbf{O}_t|\mathbf{O}_{t-1})$. The log-likelihood can now be expressed as:

$$l_T = \sum_{t=1}^{T} \log[Pr(\mathbf{O}_t|\mathbf{O}_1, \ldots, \mathbf{O}_{t-1})]. \tag{2}$$

To compute the log-likelihood, (Lystig & Hughes, 2002) define the following (forward) recursion:

$$\phi_1(j) := Pr(\mathbf{O}_1, S_1 = j) = \pi_j b_j(\mathbf{O}_1) \tag{3}$$

$$\begin{aligned} \phi_t(j) &:= Pr(\mathbf{O}_t, S_t = j|\mathbf{O}_1, \ldots, \mathbf{O}_{t-1}) \\ &= \sum_{i=1}^{N} [\phi_{t-1}(i) a_{ij} b_j(\mathbf{O}_t)] \times (\Phi_{t-1})^{-1}, \end{aligned} \tag{4}$$

where $\Phi_t = \sum_{i=1}^{N} \phi_t(i)$. Combining $\Phi_t = Pr(\mathbf{O}_t|\mathbf{O}_1, \ldots, \mathbf{O}_{t-1})$, and equation (2) gives the following expression for the log-likelihood:

$$l_T = \sum_{t=1}^{T} \log \Phi_t. \tag{5}$$

Note that so far no assumptions have been made about the response distributions b_j, hence these can be arbitrary univariate or multivariate distributions.

DepmixS4

`depmixS4` implements a general framework for defining and fitting dependent mixture models in the R programming language (R Development Core Team, 2008). This includes standard Markov models, latent/hidden Markov models, and latent class and finite mixture distribution models. The models can be fitted on mixed multivariate data with multinomial and/or gaussian distributions. Parameters can be estimated subject to general linear constraints. Parameter estimation is done through an EM algorithm or by a direct optimization approach with gradients using the `Rdonlp2` optimization routine when contraints are imposed on the parameters.

The `depmixS4` package was motivated by the fact that Markovian models are used commonly in the social sciences, but no comprehensive package was available for fitting such models. Common programs for Markovian models include Panmark (Van de Pol, Langeheine, & Jong, 1996), and for latent class models Latent Gold (Vermunt & Magidson, 2003). Those programs are lacking a number of important features. There are currently some packages in R that handle hidden Markov models but they lack a number of features that we needed in our research. In particular, `depmixS4` was designed to meet the following goals:

1. to be able to fit transition models with covariates, i.e., to have time-dependent transition matrices

2. to be able to include covariates in the prior or initial state probabilities of models

3. to allow for easy extensibility, in particular, to be able to add new response distributions, both univariate and multivariate, and similarly to be able to allow for the addition of other transition models, e.g., continuous time observation models

Although `depmixS4` is designed to deal with longitudinal or time series data, for say $T > 100$, it can also handle the limit case with $T = 1$ in analyzing cross-sectional data. In those cases, there are no time dependencies between observed data, and the model reduces to a finite mixture model, or a latent class model. Although there are other specialized (R) packages to deal with mixture data, one specific feature that we needed which is not available in other packages is the possibility to include covariates on the prior probabilities of class membership.

Response distributions and parameters

The package is built using S4 classes (object oriented classes in R) to allow easy extensibility (Chambers, 1998).

Each row of the transition matrix and the initial state probabilities:

- is modeled as a multinomial distribution
- uses the logistic link function to include covariates
- can have time-dependent covariates

Current options for the response models are models from the generalized linear modeling framework:

- normal distribution; continuous, gaussian data
- binomial logistic; binary data
- Poisson distribution; count data
- multinomial logistic; multiple choice data

All response models have the option of including covariates Other link functions may be used; eg the probit for binary data.

Parameter Estimation

Parameters are estimated in `depmixS4` using the EM algorithm or through the use of a general Newton-Raphson optimizer. The EM algorithm however has some drawbacks. First, it can be slow to converge towards the end of optimization (although it is usually faster than direct optimization at the start, so possibly a combination of EM and direct optimization is fastest). Second, applying constraints to parameters can be problematic; in particular, EM can lead to wrong parameter estimates when applying constraints. Hence, in `depmixS4`, EM is used by default in unconstrained models, but otherwise, direct optimization is done using `Rdonlp2` (Tamura, 2007; Spellucci, 2002), because it handles general linear (in)equality constraints, and optionally also non-linear constraints.

Illustrations

Two illustrations are provided below of models that analyze single participant time series data from two common experimental paradigms. In both of these, participants learn different strategies through trial and error.

Iowa gambling task

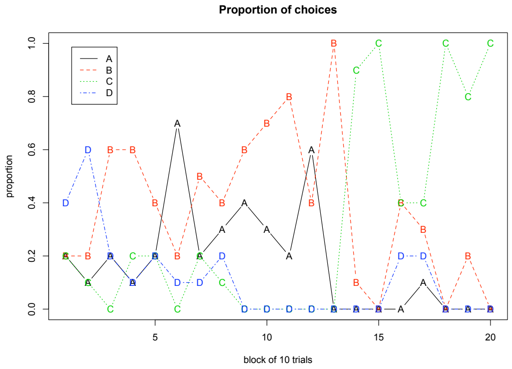
The Iowa gambling task (IGT) is an experimental paradigm designed to mimic real-life decision-making situations (Bechara, Damasio, Damasio & Anderson, 1994), in the way that it factors uncertainty, reward and punishment (Dunn, Dalgleish, & Lawrence, 2006). The task requires the selection of cards from four decks. Each deck is characterized by a certain amount of gain (delivered on each draw), frequency of loss, and amount of loss. Two decks (A and B) yield consistently high rewards, but also high, probabilistic penalties and are both (equally) disadvantageous in the long run. The other two decks (C and D) yield consistently smaller rewards, but also low, probabilistic penalties and are both (equally) advantageous in the long run. It is assumed that the ventromedial prefrontal cortex (VMPFC) is active in the IGT as VMPC patients show impaired task performance. Their preference for the decks with immediate high rewards indicates “myopia for the future”.

Crone & van der Molen (2004) designed a developmentally appropriate analogue of the IGT, the Hungry Donkey Task (HDT), with a similar win and loss schedule although the abolute amounts were redcuced by a factor of 25. The HDT is a pro-social game inviting the player to assist a hungry donkey to collect as many apples as possible, by opening one of four doors. Again, doors A and B are characterized by a high constant gain (10 apples), whereas doors C and D deliver a low constant gain (2 apples). At doors A and C, a loss of 50 apples (A) or 10 apples (C) is delivered in 50% of the trials. For doors B and D, frequency of loss is only 10%. The median loss of doors B and D is 10 and 2, respectively. Crone and van der Molen administered the HDT to children from four age groups (6-9, 10-12, 13-15, and 18-25 year-olds) and concluded that children also fail to consider future consequences.

A reanalysis of this dataset (Huizenga, Crone, & Jansen, 2007) indicated that participants might solve the task by sequentially considering the three dimensions (constant gain, frequency of loss, and amount of loss) in order to choose a door. Most youngest children in the dataset seem to focus on the dominant dimension in the task, frequency of loss, resulting in equal preference for doors B and D. Older participants seem to use a two-dimensional rule where participants first focus on the frequency of loss and then consider amount of loss, resulting in a preference for door D. A third very small subgroup seems to use an integrative rule where participants combine all three dimensions in the appropriate way. Participants using the integrative rule pick cards from doors C and D, which are advantageous in the long run.

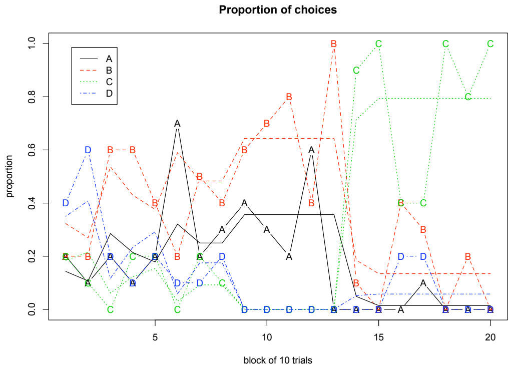
Typical analyses of these data use the last 60 trials in a series of 200 trials. A silent assumption that is made in these analyses is that behavior has stabilized after 140 trials of learning; this could very well be wrong and it is highly likely that there are individual differences in this learning process.

Single participant choices analyzed.



- N=200
- 4 choice data, displayed in blocks of 20 trials
- optimal strategy is choosing C or D

Results are the 4-state model (best by AIC/BIC), model predicted probabilities are in the figure ???. States are characterized by different types of behavior, shifting from B/D strategy to C/D (optimal) strategy.



Weather prediction task

The Weather Prediction Task (WPT, Knowlton, Squire & Gluck, 1994) is a probabilistic categorization task, in which participants learn to predict the state of the weather (sunny, or rainy) on the basis of four "tarot" cards (cards with abstract geometrical patterns). Each cue pattern is associated with a particular probability distribution over the states of the weather. In order to perform in the task, participants must predict the weather in accordance with these conditional probabilities.

There are different accounts of probabilistic category learning. According to instance or exemplar learning theories, participants learn by storing each encountered cue-outcome pairing. When presented with a cue pattern, these exemplars are retrieved from memory, and weighted according to their similarity to the probe cue pattern, to form a classification. According to associative theories, participants gradually learn by gradually associating the individual cues (or cue patterns in configural learning) to the outcomes. In rule-learning, participants are taken to extract rules by which to categorize the different cue patterns. Gluck, Shohamy and Myers (2002) proposed a number of such rules (or strategies). A main difference between these is whether responses are based on the presence/absence of a single cue, or whether responses are based on cue patterns. Gluck et al. formulated all strategies in a deterministic and optimal manner (e.g., the multi-cue strategy corresponded to giving the optimal response to each cue pattern). Meeter et al. allowed for probabilistic responding (a small probability of giving the non-optimal response).

Alternative non-strategy based analyses of the WPT (Lagnado et al, Speekenbrink et al) have estimated response strategies by variations of logistic regression.

Here, we analyze the behavior of a single individual performing the WPT for 200 trials. We let each state be characterized by a Generalized Linear Model with a Bernoulli response and logistic link function. We are interested in whether a DMM can recover a strategy model in line with Gluck et al. As we fit the data to a single subject, we must place some constraints. Specifically, we constrain the state transitions to be in a "left-right" format (states can only proceed to the immediately adjacent state and never back) and the initial state

A single state model (usual GLM) We started with a 3-state model,

- N=200
- 4 choice data, displayed in blocks of 20 trials
- optimal strategy is choosing C or D

Discussion

- depmixS4 can be downloaded from: http://r-forge.r-project.org/depmix/
- It is feasible to fit hidden Markov models in moderate length time series
- Many applications in experimental psychology
- Future developments:

1. richer measurement models, eg factor models, AR models etc
2. richer transition models, eg continuous time measurement occasions
3. explicit state durations
4. identifiability of models
5. model selection

    - standard errors of parameters

References

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W.(1994). Insensitivity to future consequences following damage to human prefrontal cortex. Cognition, 50(13), 715.

Chambers, J. M. (1998). Programming with Data: A Guide to the S Language. New York: Springer-Verlag.

Crone, E. A., & van der Molen, M. W. (2004). Developmental changes in real life decision making: Performance on a gambling task previously shown to depend on the ventromedial prefrontal cortex. Developmental Neuropsychology, 25(3), 251-279.

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. Neuroscience and Biobehavioral Reviews, 30(2), 239-271.

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the weather prediction task?: Individual variability in strategies for probabilistic category learning. Learning & Memory, 9, 408-418.

Huizenga, H. M., Crone, E. A., & Jansen, B. R. J. (2007). Decision-making in healthy children, adolescents and adults explained by the use of increasingly complex proportional reasoning rules. Developmental Science, 10(6), 814-825.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. Learning & Memory, 1 , 106-120.

Siegler, R. S. (1981). Developmental sequences within and between concepts. Monographs of the Society for Research in Child Development, 46(2, Serial No. 189).

Visser, Raijmakers, & Van der Maas (2008). Dynamics book chapter.

Author note

Ingmar Visser was supported by an EC Framework 6 grant, project 516542 (NEST). Maarten Speekenbrink was supported by the ESRC Centre for Economic Learning and Social Evolution (ELSE). Thanks to Hilde Huizenga, Brenda Jansen and Anna van Duijvenvoorde for the Iowa Gambling task data and David Lagnado for the Weather Prediction Task data.

References

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, *67*, 1554–40.

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

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.

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.

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

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. In

R. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (chap. 13). New York: Wiley.

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.

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.

Spellucci, P. (2002). Donlp2.

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

Van de Pol, F., Langeheine, R., & Jong, W. D. (1996). *Panmark 3. panel analysis using Markov chains. a latent class analysis program [user manual].* Voorburg: The Netherlands: Statistics Netherlands.

Vermunt, J. K., & Magidson, J. (2003). *Latent gold 3.0 [computer program and user's guide].* Belmont (MA), USA: Statistical Innovations Inc.

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