

LATENT CLASS MODELS FOR DIARY METHOD DATA:
PARAMETER ESTIMATION BY LOCAL COMPUTATIONS

FRANK RIJMEN

VU MEDICAL CENTER
KATHOLIEKE UNIVERSITEIT LEUVEN

KRISTOF VANSTEELANDT

UC SINT-JOSEF KORTENBERG

PAUL DE BOECK

KATHOLIEKE UNIVERSITEIT LEUVEN

The increasing use of diary methods calls for the development of appropriate statistical methods. For the resulting panel data, latent Markov models can be used to model both individual differences and temporal dynamics. The computational burden associated with these models can be overcome by exploiting the conditional independence relations implied by the model. This is done by associating a probabilistic model with a directed acyclic graph, and applying transformations to the graph. The structure of the transformed graph provides a factorization of the joint probability function of the manifest and latent variables, which is the basis of a modified and more efficient E-step of the EM algorithm. The usefulness of the approach is illustrated by estimating a latent Markov model involving a large number of measurement occasions and, subsequently, a hierarchical extension of the latent Markov model that allows for transitions at different levels. Furthermore, logistic regression techniques are used to incorporate restrictions on the conditional probabilities and to account for the effect of covariates. Throughout, models are illustrated with an experience sampling methodology study on the course of emotions among anorectic patients.

Key words: graphical models, latent Markov model, hierarchical latent Markov model, junction tree algorithm.

In behavioural and social sciences, there is an increasing use of so-called diary methods. In diary studies, people are asked to report regularly on events and experiences as they are occurring in their daily lives. A first main advantage of these methods over traditional designs is in their higher ecological validity. Second, the longitudinal data that result from diary studies allow for investigating how phenomena evolve over time (Bolger, Davis, & Rafaeli, 2003).

Multilevel (random coefficient, mixed) regression models are becoming the standard tool for the statistical analysis of diary data (Bolger et al., 2003). Extending the multilevel model with a measurement model even increases its scope (Rabe-Hesketh, Skrondal, & Pickles, 2004; Skrondal & Rabe-Hesketh, 2005). In multilevel models, temporal dynamics are naturally conceived of as gradual processes, and modelled as a smooth function of time. For example, growth is typically modelled as a low-order polynomial of time, and individual differences therein are accounted for by specifying a distribution (over persons) for the trend coefficients.

An alternative framework is offered by latent class models. More in particular, in latent Markov models (Collins & Wugalter, 1992; Langeheine & van de Pol, 1990; Poulsen, 1990;

Frank Rijmen was partly supported by the Fund for Scientific Research Flanders (FWO).

Requests for reprints should be sent to Frank Rijmen, Clinical Epidemiology and Biostatistics, VU Medical Center, De Boelelaan 1118, 1007 MB Amsterdam, The Netherlands. E-mail: f.rijmen@vumc.nl

Wiggins, 1973), temporal dynamics are conceived as transitions between latent classes. Whether changes over time are gradual and smooth or more abrupt is not specified a priori but estimated from the data. Changes over time are discontinuous to the degree that classes show qualitatively distinct estimated response profiles. A typical diary study involves many measurement occasions however, which renders the standard EM algorithm for maximum likelihood estimation of the parameters of latent Markov (and related) models computationally too demanding.

After presenting a short description of the diary method data set which is a motivating example throughout the paper, we present the standard latent Markov model. Subsequently, we show how an efficient EM algorithm can be constructed by exploiting the conditional independence relations implied by the model. As we will explain, graphical model theory turns out to be very useful in this respect because it provides a *general* procedure by working on the graphical structure of a probabilistic model. The generality is illustrated by presenting an extension of the latent Markov model that incorporates latent variables at several levels and is therefore called a hierarchical latent Markov model. In the remaining part, we discuss how restrictions can be taken into account and how covariates can be incorporated into the model using logistic regression techniques. These topics are less well covered in the graphical model literature, and constitute the innovating part of the paper.

The data: An ecological momentary assessment study on the course of emotions among anorectic patients.

Ecological momentary assessment is a diary method in which participants have to report their momentary experiences or behaviours contingent to a signal or to an event. The participants in this study (Vansteelandt, Rijmen, Pieters, & Vanderlinden, 2007) were 32 female patients from an inpatient eating disorder unit. At regular time-intervals, they received a signal and were asked to fill out a behavioural questionnaire. The participants received nine signals a day during one week (9 beeps \times 7 days = 63 beeps), one in each block of 90 minutes between 8.30 and 22.30. Within each time block, the time of administration was drawn from a uniform distribution (stratified random time sampling, Delespaul, 1995). The mean intersignal time was 1 h 32 min and the standard deviation was 38 min. At each signal, patients were asked to rate themselves on a 7-point scale with respect to the intensity with which they experienced 12 emotional states. The choice of emotional states was inspired by the work of Diener, Smith, and Fujita (1995). They discerned six categories of emotions: anger, shame, fear, sadness, joy, and love. In the present study, two emotional states were taken from each emotional category: anger and irritation, shame and guilt, anxiety and tension, sadness and loneliness, happiness and joy, and love and appreciation, respectively.

Assessments that were reported more than 15 min after the administration of the signal were excluded. This resulted in a considerable amount of missing data, 24%, but the advantage is that one has a strong guarantee that data are not contaminated by retrospective bias (Delespaul, 1995).

We work on the dichotomized responses (0–2 vs. 3–6), and, unless mentioned otherwise, we treat the data as if signals are equally spaced in time, ignoring the random administration within time blocks.

1. Latent Markov Model

The multiple indicator latent Markov model (Collins & Wugalter, 1992; Langeheine & van de Pol, 1994) is defined as follows. Let y_{itj} denote the categorical response of person i at occasion t on item j , $i = 1, \dots, n$; $j = 1, \dots, J$; $t = 1, \dots, T$. $\mathbf{y}_{it} = (y_{it1}, \dots, y_{itj}, \dots, y_{itJ})'$ denotes the response vector of person i at occasion t , and $\mathbf{y}_i = (\mathbf{y}'_{i1}, \dots, \mathbf{y}'_{it}, \dots, \mathbf{y}'_{iT})'$ the complete response pattern of person i . $z_{it}, z_{it} = 1, \dots, s, \dots, S$, is the categorical latent state of person i at

occasion t , and thus $\mathbf{z}_i = (z_{i1}, \dots, z_{it}, \dots, z_{iT})'$ is the trajectory of person i through the latent space over time. Finally, $\mathbf{x}_i = (\mathbf{z}'_i, \mathbf{y}'_i)'$ denotes the “complete” data vector of person i . Corresponding random variables are denoted by capitals.

Assuming a first-order Markov chain for the latent variable, the latent state at occasion $t + 1$ depends on the past latent states through the latent state at occasion t only, $\Pr(z_{it+1}|z_{i1}, \dots, z_{it}) = \Pr(z_{it+1}|z_{it})$. Assuming conditional independence between responses, given the latent state, the marginal probability of a response pattern \mathbf{y}_i is then

$$\Pr(\mathbf{y}_i) = \sum_{\mathbf{z}_i} \Pr(\mathbf{x}_i) = \sum_{\mathbf{z}_i} \Pr(z_{i1}) \prod_{j=1}^J \Pr(y_{i1j}|z_{i1}) \prod_{t=1}^{T-1} \Pr(z_{it+1}|z_{it}) \prod_{j=1}^J \Pr(y_{it+1j}|z_{it+1}), \quad (1)$$

where the summation is over the S^T possible latent trajectories \mathbf{z}_i . The parameter vector $\boldsymbol{\theta}$ of the latent Markov model consists of three subsets:

- τ_{rs} , $r, s = 1, \dots, S$, the time homogeneous transition probabilities between latent states, $\Pr(Z_{it+1} = s|Z_{it} = r) = \tau_{rs}$ for $t = 1, \dots, T - 1$;
- α_{1s} , $s = 1, \dots, S$, the marginal state probabilities at occasion 1 (initial state probabilities);
- π_{cjs} , $c = 2, \dots, C$, $j = 1, \dots, J$, $s = 1, \dots, S$, the state-conditional response probabilities, $\Pr(Y_{itj} = c|Z_{it} = s) = \pi_{cjs}$. C is the number of response categories. The conditional response probability for the first category is $\pi_{1js} = 1 - \sum_{c=2}^C \pi_{cjs}$.

Marginal state probabilities at occasion $2, \dots, T$ are given by the recursive formula $\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}'_t \boldsymbol{\tau}$, $t = 1, \dots, T - 1$, where $\boldsymbol{\alpha}_t = (\alpha_{t1}, \dots, \alpha_{tS})'$ and $\boldsymbol{\tau} = (\tau_{rs})$.

The latent Markov model was applied to the diary data set. We treated each person-by-day combination as a separate case, measured at nine occasions. This comes down to assuming that the data stemming from different days were independent and that parameters were constant over days. Further on, we will relax the “tomorrow is another day” assumption and model dependencies between days as well.

A model with four latent states was estimated. The model contained 63 free parameters and the deviance ($-2 \times \log$ likelihood) amounted to 17623.3. Figure 1 displays the maximum likelihood estimates (MLEs) of the state-conditional probabilities for experiencing each of the emotions. State 1 is characterized by high probabilities of experiencing positive emotions (joy, happiness, appreciation, love) and low probabilities for negative emotions (sadness, anger, loneliness, shame, anxiety, tension, guilt, irritation). We can interpret state 1 as “positive mood”. The reverse pattern holds for state 3 (“negative mood”). The probabilities are low for all emotions in state 2, except for “tension”. In state 4, as in state 1, positive emotions tend to have a higher probability than negative emotions (except “tension”), but both categories of emotions are less well separated. State 4 can be considered to be a neutral to moderately positive mood. The estimated initial state and state transition probabilities are shown in Table 1, as well as the marginal state probabilities at occasion 9. The probabilities of staying in the same state are high for all states. Transitions are most likely to occur between states 2 and 1 ($\tau_{21} = .14$), between states 3 and 4 ($\tau_{34} = .18$), and between states 4 and 3 ($\tau_{43} = .14$). To a smaller degree, there are transitions as well between states 1 and 2 ($\tau_{43} = .08$), and between states 4 and 1 ($\tau_{43} = .07$). It is interesting to see that the two states with an opposite pattern of conditional probabilities, states 1 and 3, do barely communicate directly, but that there is an indirect transition from state 3 (“negative mood”) to state 1 (“positive mood”) via the emotionally more neutral state 4. Over the day, the marginal state probabilities of states 1 and 4 increase (from .27 to .33, and .16 to .24, respectively) at the expense of state 3 (from .34 to .22). So, the mood of patients tends to become better later on in the day.

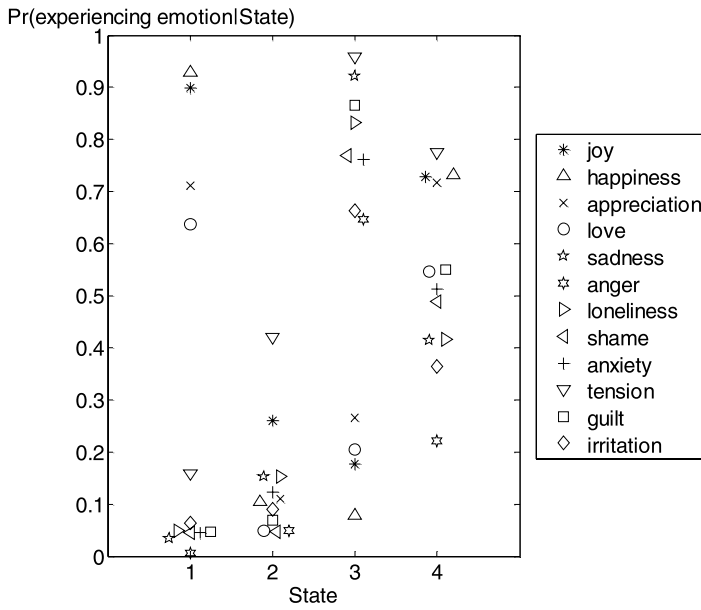


FIGURE 1.

Estimated state-conditional probabilities for the latent Markov model with four states. Days are assumed to be independent.

TABLE 1.

Estimated initial state probabilities, state probabilities at occasion 9, and state transition probabilities for the latent Markov model with four states. Days are assumed to be independent.

Initial state probabilities: α_{1s}	.27	.22	.34	.16
State probabilities at occasion 9: α_{9s}	.33	.22	.22	.24
State transition probabilities: τ_{rs}	.85	.08	.02	.05
	.14	.81	.04	.01
	.03	.04	.75	.18
	.07	.03	.14	.76

2. Computational Burden of the Standard EM Algorithm

In the standard EM algorithm, the E-step would consist of calculating, for each person i , the posterior probabilities $\Pr(\mathbf{z}_i | \mathbf{y}_i)$ for all S^T possible trajectories \mathbf{z}_i through the latent space in order to obtain the expected complete data log likelihood given the observed data and a set of provisional parameter estimates $\hat{\theta}^p$,

$$Q(\theta | \hat{\theta}^p) = \sum_{i=1}^n \sum_{\mathbf{z}_i} \Pr(\mathbf{z}_i | \mathbf{y}_i; \hat{\theta}^p) \log \Pr(\mathbf{x}_i; \theta).$$

The E-step becomes computationally too demanding when there are more than a few measurement occasions, because the number of possible trajectories increases exponentially with the number of measurement occasions. For example, for the latent Markov model applied to the diary data set, S^T equals 262144. The crucial quantities to update the parameters in the M-step,

however, are not the joint posterior probabilities of all latent variables, but the posterior marginal and pairwise consecutive state probabilities:

$$\begin{aligned} \Pr(Z_{it} = s | \mathbf{y}_i), \quad s = 1, \dots, S; \quad t = 1, \dots, T \quad \text{and} \\ \Pr(Z_{it} = r, Z_{it+1} = s | \mathbf{y}_i), \quad r, s = 1, \dots, S; \quad t = 1, \dots, T - 1. \end{aligned}$$

For example, the update equation for a conditional response probability π_{cjs} is

$$\hat{\pi}_{cjs}^{p+1} = \frac{\sum_t \sum_{i: y_{it}=c} \Pr(Z_{it} = s | \mathbf{y}_i)}{\sum_t \sum_i \Pr(z_{it} = s | \mathbf{y}_i)}.$$

Baum, Petrie, Soules, and Weiss (1970) described how the posterior marginal and pairwise successive state probabilities can be calculated efficiently, that is, without calculating the posterior probabilities of all possible trajectories through the latent space, by a set of forward–backward recursions that exploit the conditional independence relations of the latent Markov model. Before describing a generalization of the Baum–Welch algorithm, we shortly present some notions from graphical model theory. We limit ourselves to those concepts and results that are essential for an understanding of the modified EM algorithm to be presented later. The interested reader is referred to Cowell, Dawid, Lauritzen, and Spiegelhalter (1999).

3. Local Computation Based on the Junction Tree

Many statistical models can be represented graphically with a directed acyclic graph (in the context of latent class models, see Hagenaars, 1998; and Humphreys & Titterton, 2003) in which each node corresponds to a (manifest or latent) random variable. In latent class models, all variables are discrete and so we will restrict our attention to discrete variables, but note in passing that many of the stated results have been extended to continuous variables and mixed sets of discrete and continuous variables (Lauritzen, 1996). Figure 2 shows a directed acyclic graph for the latent Markov model for two items that are administered at three occasions. The (absence of) directed edges between nodes represent conditional (in)dependence relations. For example, the directed edge between Z_{i2} and Z_{i3} (Z_{i2} is called a “parent” of Z_{i3}) represents the conditional dependence of Z_{i3} on Z_{i2} , whereas the conditional independence between Z_{i1} and Z_{i3} , given Z_{i2} (the first-order Markov assumption), is represented by Z_{i2} being situated on the single directed path between Z_{i1} and Z_{i3} .

A directed acyclic graph associated with a probabilistic model for a set of discrete random variables X_1, \dots, X_M admits a recursive factorization of the joint probability function

$$\Pr(\mathbf{x}) = \prod_{m=1}^M \Pr(x_m | pa(x_m)), \quad (2)$$

where $pa(x_m)$ denotes the realization of the random variables that are parents of X_m . Applying equation (2) to the directed acyclic graph in Figure 2, one indeed obtains the same factorization of the probability of a complete data vector \mathbf{x}_i as in equation (1).

The core of the construction of efficient computational schemes relies on the transformation of a directed acyclic graph into a junction tree. For a detailed account of the algorithms for transforming a directed acyclic graph into a junction tree, see Cowell et al. (1999). We suffice by saying that the nodes of the junction tree correspond to sets of variables, called cliques, and that the intersection between two neighbouring cliques C_k and C_l is called a separator, $S_{kl} = C_k \cap C_l$. Furthermore, a junction tree constructed from a particular directed acyclic graph is not necessarily

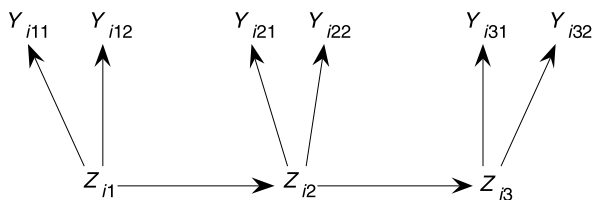


FIGURE 2.

Directed acyclic graph for the latent Markov model with three measurement occasions and two items per occasion.

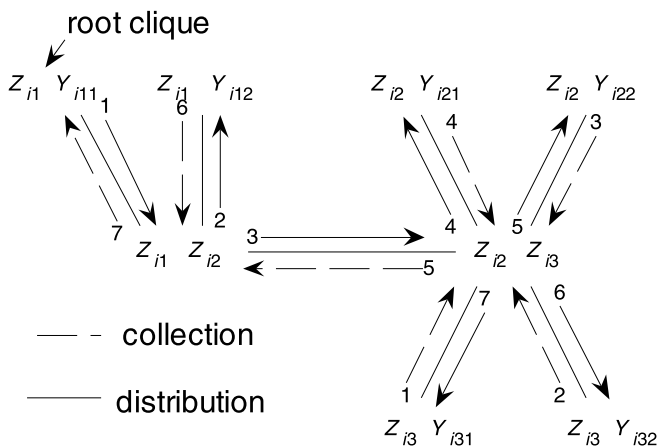


FIGURE 3.

Schedule of flows on the junction tree of cliques for the latent Markov model. Three measurement occasions and two items per occasion.

unique. Figure 3 shows a junction tree that is obtained from the directed acyclic graph in Figure 2. The cliques are the sets

$$\{Z_{it}, Y_{itj}\}, t = 1, \dots, T, j = 1, \dots, J, \quad \text{and} \quad \{Z_{it}, Z_{it+1}\}, t = 1, \dots, T - 1.$$

A crucial result is that a junction tree offers an alternative factorization of the joint probability function. More, in particular, the joint probability function of all variables can be factorized as the product of all marginal clique probabilities over the product of all marginal separator probabilities:

$$\Pr(\mathbf{x}) = \frac{\prod_C \Pr(\mathbf{x}_C)}{\prod_S \Pr(\mathbf{x}_S)}. \tag{3}$$

The factorization of equation (3) serves as the basis for an efficient computational scheme using local computations. A first step is to associate nonnegative functions or potentials Ψ to each clique and separator of the junction tree. The domain of Ψ is the set of all possible realizations of the random variables in the clique or separator. All potentials are initialized with value 1. Then, each factor of equation (2) is multiplied into a clique that contains all the nodes corresponding to the random variables in the factor. The way a junction tree is constructed implies that there is always such a clique; if there is more than one, it does not matter which one is chosen. Then,

$$\Pr(\mathbf{x}) = \frac{\prod_C \psi(\mathbf{x}_C)}{\prod_S \psi(\mathbf{x}_S)}. \tag{4}$$

Next, a schedule of flows is passed along the edges of the junction tree. Let C_k and C_l be two consecutive nodes of the junction tree, with separator S_{kl} . New potentials are defined as

$$\psi^*(\mathbf{x}_{S_{kl}}) = \sum_{C_k \setminus S_{kl}} \psi(\mathbf{x}_{C_k}),$$

and

$$\psi^*(\mathbf{x}_{C_l}) = \psi(\mathbf{x}_{C_l})\lambda(\mathbf{x}_{S_{kl}}),$$

where

$$\lambda(\mathbf{x}_{S_{kl}}) = \frac{\psi^*(\mathbf{x}_{S_{kl}})}{\psi(\mathbf{x}_{S_{kl}})}, \quad \lambda(\mathbf{x}_{S_{kl}}) \equiv 0 \quad \text{if } \psi(\mathbf{x}_{S_{kl}}) = 0,$$

and $\sum_{C_k \setminus S_{kl}}$ denotes marginalization over all random variables whose corresponding nodes are in $C_k \setminus S_{kl}$.

Jensen, Lauritzen, and Olesen (1990) proposed the following two-phase schedule: First, select an arbitrary clique of the junction tree as the root-clique. In the collection, phase flows are passed along the edges towards the root-clique. In the distribution-phase, flows are passed in the reverse direction. See Figure 3 for a scheduling of flows. After applying the two-phase schedule, equilibrium is reached and the clique and separator potentials correspond to the marginal probability functions of the cliques and separators, respectively. Numerical underflow can be avoided by normalizing the potentials of each clique after updating the potentials of that clique, for example, by dividing each potential by the sum of the clique potentials. After applying the two-phase schedule, the clique and separator potentials then become *proportional* to the marginal probability functions of the cliques and separators, respectively. The proportionality constant equals the product of the clique normalizing constants.

When some of the variables are observed, posterior probability distributions of the unobserved variables can be obtained along similar lines. The only change is that, for each observed variable, some arbitrary clique that contains the variable is selected, and the potentials of all realizations of the clique variables involving a different state for the observed variable than the observed one are set to 0. This step is called “entering the evidence”. After applying the two-phase schedule, the clique and separator potentials now are proportional to the posterior marginal probability functions of the cliques and separators, respectively. Normalizing the potentials of a clique or separator so that they sum to one yields its posterior probability distribution. Posterior distributions for individual variables are obtained by marginalizing over all other variables in the clique or separator.

4. Modified EM Algorithm

In the previous section we assumed the conditional probabilities of the factorization according to the directed acyclic graph (equation (2)) to be known. Let P denote the set of these conditional probabilities, with elements $p(x_m | pa(x_m))$. If these conditional probabilities have to be estimated from (partially) observed data, MLEs can be obtained by applying the following EM algorithm (Lauritzen, 1995):

1. Start with some initial estimates $\hat{p}^0(x_m | pa(x_m))$.

E-step:

2. Initialize the junction tree with the current estimates for the conditional probabilities

3. Enter the evidence $\mathbf{x}_i \text{ obs}$ for each case. Each case may have a different set of observed variables.
4. Apply the two-phase propagation schedule for each case to obtain posterior probabilities of the latent variables.

M-step:

5. Update the conditional probabilities:

$$\hat{p}^{\text{new}}(x_m | pa(x_m)) = \frac{\sum_{i=1}^n \Pr(x_{im}, pa(x_{im}) | \mathbf{x}_i \text{ obs}; \hat{P}^{\text{old}})}{\sum_{i=1}^n \Pr(pa(x_{im}) | \mathbf{x}_i \text{ obs}; \hat{P}^{\text{old}})}. \quad (5)$$

$\Pr(x_{im}, pa(x_{im}) | \mathbf{x}_i \text{ obs}; \hat{P}^{\text{old}})$ can always be obtained easily because there is always a clique that contains a variable and its parents. When a variable and its parents are observed for all cases, the updated conditional probability is the observed proportion. When a variable has no parents, the denominator is replaced by n .

Repeat Steps 2 to 5 until convergence. The obtained solution corresponds to a stationary point of the log likelihood (McLachlan & Krishnan, 1997). The contribution of each case to the log likelihood is obtained by summing the potentials of any arbitrary clique over its clique space, logging it, and adding the sum of the logged normalizing constants (the constants that resulted from normalizing the clique potentials to avoid numerical underflow).

If some variables share the same conditional probability distribution (e.g., transition probabilities that are assumed to be equal over time), the updated probabilities are obtained by summing the numerator and denominator in equation (5) over the variables sharing the same conditional probability distribution.

The complexity of the E-step of the modified EM algorithm scales with the sum of the clique state spaces. Hence, the smaller the clique state spaces, the more gain in efficiency is obtained by using the modified instead of the standard EM algorithm.

The Baum–Welch algorithm for the latent Markov model with a single item administered at each measurement occasion is a special case of the modified EM algorithm (Smyth, Heckerman, & Jordan, 1997). Its complexity is $O(S^2T)$, which is substantially less than the $O(S^T)$ complexity of the standard EM algorithm, except for very small T .

5. Merging Terminal Observed Nodes Sharing the Same Parents

Common to the presented models and most latent class and latent Markov models described in the literature is the existence of sets of observed variables that share the same parent(s) and appear as terminal nodes in the directed acyclic graph. This stems from the fact that, in latent class models, it is typically assumed that all dependencies between a set of observed variables are explained by a smaller set of discrete latent variables. For example, in the latent Markov model that we applied to the illustrative data set (and also in the model to be discussed next), we had 63 sets of observed variables sharing the same parents, each set consisting of 12 variables. Including them all as nodes of the associated directed acyclic graph renders the latter needlessly complex. Instead, we can merge the nodes belonging to the same set into one node, and construct a junction tree for the reduced graph. The effective state space of such a node is only of size one and equals, for each case, the observed response pattern on the corresponding set of observed variables. This is because, when entering evidence, the potentials of all configurations other than the observed one are set to zero. Including these configurations would only result in a needless propagation of zeros (Huang & Darwiche, 1996). Initialization is done in the same way as described before,

except for the fact that we now, for each case, use the conditional probabilities of the observed response *patterns* on the sets of terminal nodes sharing the same parent(s). These conditional probabilities are computed over all data that are not missing (assuming ignorable missingness, Little & Rubin, 1987). If all data on such a set of observed variables are missing, this probability is set to 1, leaving the potentials unaltered.

6. A Hierarchical Latent Markov Model

Before discussing the incorporation of constraints on the conditional probabilities and how to incorporate covariates, we go back to the motivating example and present an extension of the latent Markov model. In this extension, we drop the assumption that the latent states of consecutive days are independent. Although the assumption of independent days may be reasonable within a person, it is unrealistic to assume that the data of a person on two consecutive days are not more alike than the data of two different persons, given the well-established finding that stable individual differences in emotional experiences do exist (Feldman, 1995; Watson, Clark, & Tellegen, 1988). Therefore, we included an additional latent variable at the day-level. Transitions between the latent states at both levels (day- and signal-) were modelled as first-order time homogeneous Markov chains. The initial state and state transition probabilities of the latent Markov chain at the signal-level were specific for each state at the day-level. That is, the latent variables at the signal-level conditionally depended on the latent variables at the day-level. The response probabilities were specific for each day- and signal-state. The hierarchical latent Markov model contains five sets of parameters:

- $\tau_{r_{\text{day},s_{\text{day}}}}^{\text{day}}, r_{\text{day},s_{\text{day}}}^{\text{day}}, s_{\text{day}} = 1, \dots, S^{\text{day}}$, the time homogeneous transition probabilities between latent day-states;
- $\alpha_{1,s_{\text{day}}}^{\text{day}}, s_{\text{day}} = 1, \dots, S^{\text{day}}$, the initial day-state probabilities;
- $\tau_{r_{\text{sig},s_{\text{sig}}}}^{\text{sig}}, r_{\text{sig},s_{\text{sig}}}^{\text{sig}}, s_{\text{sig}} = 1, \dots, S^{\text{sig}}, s_{\text{day}} = 1, \dots, S^{\text{day}}$, the time homogeneous transition probabilities between latent signal-states within day-states;
- $\alpha_{1,s_{\text{sig},s_{\text{day}}}}^{\text{sig}}, s_{\text{sig}} = 1, \dots, S^{\text{sig}}, s_{\text{day}} = 1, \dots, S^{\text{day}}$, the initial signal-state probabilities within day-states;
- $\pi_{c,j,s_{\text{sig},s_{\text{day}}}}, c = 2, \dots, C, s_{\text{sig}} = 1, \dots, S^{\text{sig}}, s_{\text{day}} = 1, \dots, S^{\text{day}}$, the parameters for the conditional response probabilities within day- and signal-states.

When there are no transitions between states at the day-level allowed ($\tau_{r_{\text{day},s_{\text{day}}}}^{\text{day}} = 1$ if $r_{\text{day},s_{\text{day}}}^{\text{day}} = s_{\text{day}}$, $\tau_{r_{\text{day},s_{\text{day}}}}^{\text{day}} = 0$ otherwise), the mixed latent Markov model is obtained (Langeheine & van de Pol, 1990; except that the state at the signal-level for the first signal of a day is independent of the last signal-state of the previous day, given the day-state). Figures 4 and 5 show the directed acyclic graph and its transformation into a junction tree for a hierarchical latent Markov model for items that are administered at three occasions during two days. All items administered at a given measurement occasion are represented with a single node, the motivation for which was given in the previous section. In general, the cliques of the hierarchical latent Markov model are the sets $\{Z_{it_{\text{day}}}^{\text{day}}, Z_{it_{\text{day}+1}}^{\text{day}}\}$, $t_{\text{day}} = 1, \dots, T^{\text{day}} - 1$; $\{Z_{it_{\text{day}}}^{\text{day}}, Z_{it_{\text{day}t_{\text{sig}}}}^{\text{sig}}, Z_{it_{\text{day}t_{\text{sig}+1}}^{\text{sig}}}\}$, $t_{\text{day}} = 1, \dots, T^{\text{day}}$, $t_{\text{sig}} = 1, \dots, T^{\text{sig}} - 1$; and $\{Z_{it_{\text{day}}}^{\text{day}}, Z_{it_{\text{day}t_{\text{sig}}}}^{\text{sig}}, \mathbf{Y}_{it_{\text{day}t_{\text{sig}}}}\}$, $t_{\text{day}} = 1, \dots, T^{\text{day}}$, $t_{\text{sig}} = 1, \dots, T^{\text{sig}}$. The superscripts refers to the level of the latent variables (day- vs. signal-level).

The model was estimated with two latent states at the day-level, and two signal-states within each day-state, resulting in a total of four signal-states (number of parameters = 57). The deviance amounted to 17843.4. Figure 6 displays the MLEs of the state-conditional probabilities for experiencing each of the emotions. For the first day-state, signal-state 1 is characterized by

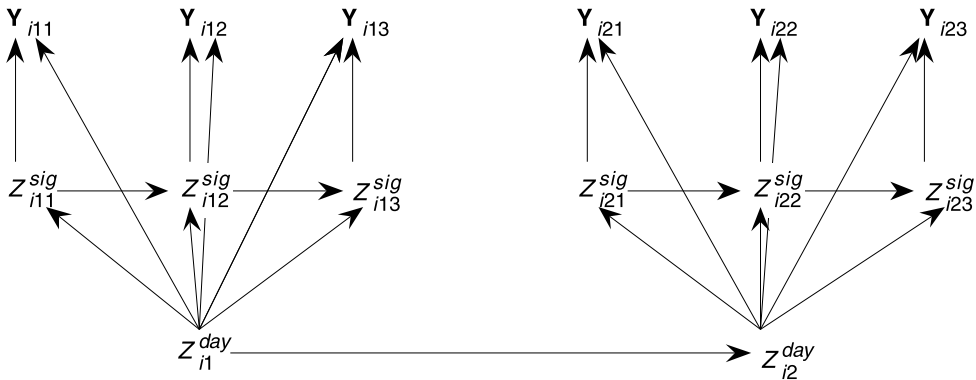


FIGURE 4.

Directed acyclic graph for the hierarchical latent Markov model. Two days and three measurement occasions per day. Items administered at the same measurement occasion are represented by a single node. The subscripts refer to person, day, and signal, respectively.

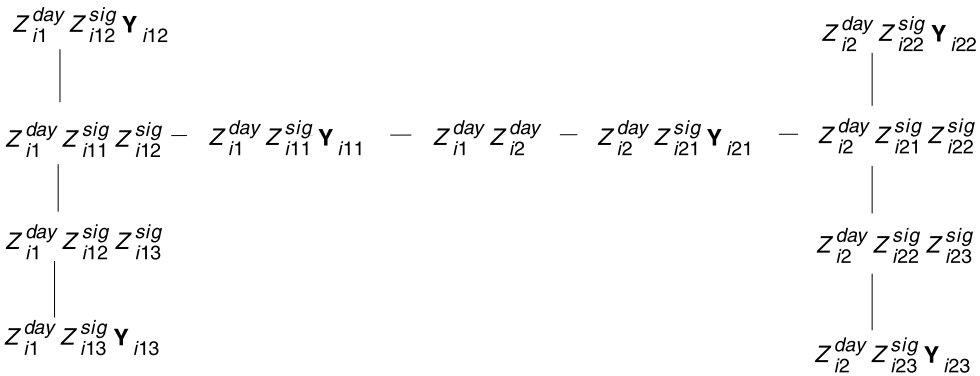


FIGURE 5.

Junction tree for the hierarchical latent Markov model. Two days and three measurement occasions per day. Items administered at the same measurement occasion are represented by a single node. The subscripts refer to person, day, and signal, respectively.

high probabilities of experiencing positive emotions and low probabilities for negative emotions. We can interpret state 1 as “positive mood”. The probabilities are low for all emotions in state 2, except for “tension”. With respect to the second day-state, the first signal-state is characterized by low probabilities of experiencing positive emotions and high probabilities for negative emotions (“negative mood”). In signal-state 2 positive emotions are not well separated from negative emotions, but the probabilities are lower for negative emotions, except for “tension”. The state can be considered to be an emotionally neutral to moderately positive state. So, overall, day-state 1 is emotionally more positive than day-state 2. The signal-states closely resemble the states of the latent Markov model (compare Figures 1 and 6).

The estimated initial state and state transition probabilities are shown in Table 2, as well as the marginal state probabilities of the day-states at day 7, and the marginal state probabilities of the signal-states at signal 9. Over days and signals, there is a tendency to experience more positive emotions. That emotions tend to become more positive at the end of the day was also found in the latent Markov analysis. That we find the same tendency over days is somewhat puzzling, however. Transitions between states at the signal-level are more likely to occur in day-state 2.

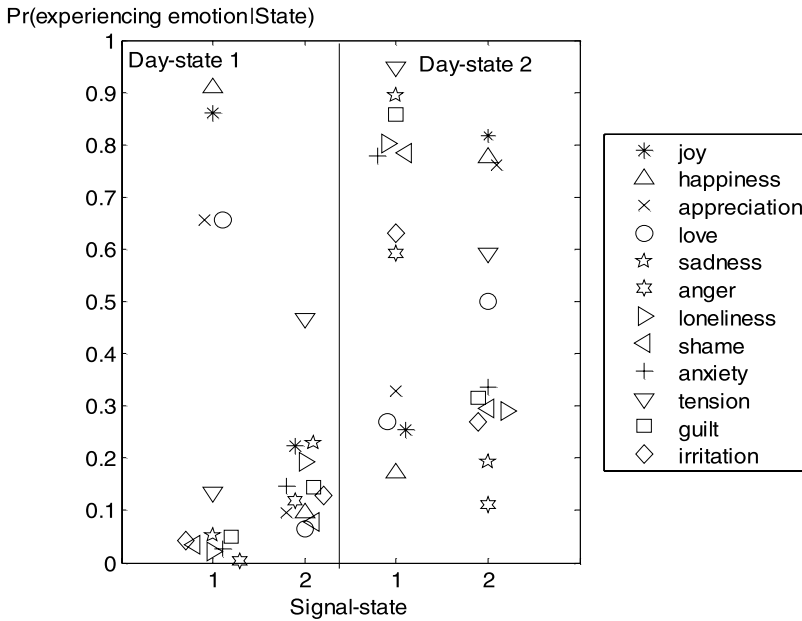


FIGURE 6.

Estimated state-conditional probabilities for the hierarchical latent Markov model with two states at signal- and day-level.

TABLE 2.

Estimated initial state probabilities, beep-state probabilities at occasion 9, and state transition probabilities for the hierarchical latent Markov model with two states at both signal- and day-level.

	Day-state 1		Day-state 2	
	Signal-state 1	Signal-state 2	Signal-state 1	Signal-state 2
Initial day-state probabilities: $\alpha_{1s,day}^{day}$.41		.59
Initial signal-state probabilities	.46	.54	.63	.37
within day-states: $\alpha_{1s, sig_s, day}^{sig}$				
Day-state probabilities at day 7: $\alpha_{7s, day}^{day}$.49		.51
Signal-state probabilities at signal 9	.57	.43	.49	.51
within day-states: $\alpha_{9s, sig_s, day}^{sig}$				
Day-state transition probabilities:		.92		.08
$\tau_{r, day_s, day}^{day}$.09		.91
Signal-state transition probabilities	.90	.10	.81	.19
within day-states: $\tau_{r, sig_s, sig_s, day}^{sig}$.14	.86	.18	.82

7. Modelling the Conditional Probabilities with Logistic Regression Models

Hitherto, we estimated a separate set of conditional probabilities for each configuration of the parent variables without imposing any restrictions on these conditional probabilities, except for equality restrictions (i.e., the transition and conditional response probabilities were assumed equal over time). However, out of a substantive hypothesis or for reasons of parsimony, one might want to model these conditional probabilities as a function of a limited number of parameters.

A straightforward choice is modelling the conditional probabilities under a logistic regression model and restricting the linear predictor to contain, for example, only the main effects of the parents. Logistic regression models can be specified for both manifest and latent variables. For categorical variables with more than two categories, multinomial (nominal variables) or ordered (ordinal variables) logistic regression techniques can be used.

Estimation proceeds along similar lines: the E-step remains the same as described above and, in the M-step, new parameter values are obtained using standard routines to fit generalized linear models such as Fischer scoring (Fahrmeir & Tutz, 2001). The conditional probabilities of the next E-step are computed from the new parameter estimates.

In a next model fitted to the diary method data set, we restricted the conditional response probabilities of experiencing the emotions using a multinomial logistic regression model with main effects for the items, interactions between day-states and positive versus negative emotions, and interactions between signal-states and positive versus negative emotions:

$$\begin{aligned}\text{logit}(\pi_{j,s,\text{sig},s,\text{day}}) &= \beta_j + \beta_{\text{pos},s,\text{sig}} + \beta_{\text{pos},s,\text{day}}, \\ \text{logit}(\pi_{k,s,\text{sig},s,\text{day}}) &= \beta_k + \beta_{\text{neg},s,\text{sig}} + \beta_{\text{neg},s,\text{day}},\end{aligned}$$

where j and k are item indices for the positive and negative emotions, respectively. These restrictions imply that both day- and signal-states only differed in terms of a shift (on the logit-scale) common to all positive emotions and a shift common to all negative emotions and, in addition, that the shifts for day- and signal-states were additive.

We estimated a model with two latent states at both the signal- and day-levels. The model contained 25 parameters: 12 item main effects; four interaction terms for the interactions between day-state and positive emotions, signal-state and positive emotions, day-state and negative emotions, and signal-state and negative emotions (there are eight interaction terms in total, but four of them were put to zero to identify the model); and nine initial state and state transition probabilities. The deviance of the estimated model amounted to 18195.8. The model is nested within the hierarchical latent Markov model without restrictions on the conditional probabilities, so that a likelihood-ratio test can be used to test the restrictions on the conditional probabilities (see the section on model selection). The model with restrictions on the conditional probabilities was rejected, $LR = 352.4$, $df = 32$, $p < .001$.

A second advantage of modelling the conditional probabilities with logistic regression models is that covariates W_1, \dots, W_p can be taken into account without much complication. Instead of factorizing $\Pr(\mathbf{x})$ according to the junction tree (see equation (3)), we now factorize $\Pr(\mathbf{x}|\mathbf{w})$. The covariates then appear in the linear predictor of the logistic regression models for the different sets of conditional probabilities. Covariates may vary over cases and/or items. Covariates can be discrete or continuous.

Alternatively, one could include the covariates in the network, and use the junction tree to factorize their joint probability function $\Pr(\mathbf{x}, \mathbf{w})$. Apart from rendering the graph and thus the process of constructing the junction tree needlessly complex, one will run into problems with continuous covariates. Probabilistic graphical models for mixed sets of discrete and continuous variables can be formulated, but the junction tree propagation requires that the continuous variables are (conditionally) Gaussian (Lauritzen, 1996). In addition, in a directed acyclic graph, continuous variables should have no discrete children (Cowell et al., 1999), which is the case when one wants to include a continuous covariate to model the conditional probabilities of a discrete random variable.

As an illustration, we included the time interval between two consecutive signals as a covariate for the probabilities of making a transition between signal-states. This way, we could test the appropriateness of our treatment of the data as if the signals were equally spaced in time, ignoring the random administration of the signals within time strata. The time interval was included

as a covariate for the hierarchical latent Markov model with no restrictions on the conditional response probabilities. The time interval was included as a linear and quadratic effect. More specifically, the transition probabilities between signal-states were, respectively, modelled as

$$\text{logit}(\tau_{r^{\text{sig}}_s^{\text{sig}}_s^{\text{day}}}^{\text{sig}}) = \beta_{r^{\text{sig}}_s^{\text{sig}}_s^{\text{day}}} + \beta_{\text{time } r^{\text{sig}}_s^{\text{sig}}_s^{\text{day}}} \times \text{time},$$

and

$$\text{logit}(\tau_{r^{\text{sig}}_s^{\text{sig}}_s^{\text{day}}}^{\text{sig}}) = \beta_{r^{\text{sig}}_s^{\text{sig}}_s^{\text{day}}} + \beta_{\text{time } r^{\text{sig}}_s^{\text{sig}}_s^{\text{day}}} \times \text{time} + \beta_{\text{time}^2 r^{\text{sig}}_s^{\text{sig}}_s^{\text{day}}} \times \text{time}^2$$

for $r^{\text{sig}} = 1, 2$; $s^{\text{sig}} = 2$, and $s^{\text{day}} = 1, 2$.

The analysis revealed no linear or quadratic trend for the time interval between two consecutive signals, $LR = 7$, $df = 4$, $p = .14$ and $LR = 12.4$, $df = 8$, $p = .14$, respectively.

8. Software

The models fitted to the diary data set were estimated with a Matlab program. The program proceeds in two steps. In the first step, the directed acyclic graph of a model for categorical variables is constructed from the specification of its edges, and transformed into a junction tree. Cliques and separators are constructed, as well as a two-phase schedule for local computations on the junction tree. This step is based on the Bayes Net Toolbox of K. Murphy (2001), which can be downloaded for free at <http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html>.

In the second step, parameters are estimated with an EM algorithm in which the E-step is carried efficiently through local computations on the junction tree following the two-phase schedule obtained in the first step. For this, a separate set of Matlab functions was written. All conditional probabilities are modelled through a logistic regression model as a function of its parents. For categorical variables with more than two categories, three link functions are available: the multinomial link function for nominal variables, and the cumulative- and adjacent-categories link functions for ordinal variables. Unrestricted conditional probabilities are modelled with a saturated multinomial logistic regression model. Additional covariates can be included. Covariates may vary over items, persons, and/or measurement occasions. The program allows for merging sets of terminal nodes corresponding to observed variables that share the same parents. Parameters can be restricted to specific values.

In the M-step, the free parameters are updated using Fischer scoring. The information matrix is approximated by numerical differentiation of the score function of the complete data in the maximum likelihood solution (the score function of the complete data equals the score function of the incomplete data; this can be shown along similar lines as presented in Fahrmeir and Tutz (2001, sect. 7.4)).

The program is not tied to the specific models presented in this paper, and can be used for a wide variety of models for discrete variables. For models of the latent Markov family, models that can be fitted include, the multiple group latent Markov model (Collins & Wugalter, 1992), the mixed latent Markov model (Langeheine & van de Pol, 1990), latent Markov models incorporating covariates (Vermunt, Langeheine, & Böckenholt, 1999), and higher-order latent Markov models (Langeheine & van de Pol, 2000). The set of Matlab functions can be downloaded at www.mathworks.com/matlabcentral/fileexchange/loadauthor.doioobjectid=1095467.

9. Discussion

In this paper we illustrated how latent class models can be used in modelling multiwave multivariate categorical data, such as data stemming from diary method studies. MLEs are obtained from an efficient EM algorithm in which the conditional independence relations of the

model are exploited during the E-step. Such efficient EM algorithms have already been proposed for specific models, with the Baum–Welch algorithm for the standard latent Markov model as the most notorious example (Baum et al., 1970; the Baum–Welch algorithm is actually a precursor of the EM algorithm). Vermunt (2003) established a similar algorithm for the estimation of multilevel latent class models. The basic algorithm we described (Lauritzen, 1995) originates from graphical model theory and operates on a junction tree that is constructed from the directed acyclic graph associated with a probabilistic model. The main advantage of relying on graphical model theory is that it offers a general approach that is not tied to the context of a specific model. A junction tree and a corresponding schedule of flows for local computation can be constructed automatically from any directed acyclic graph by applying available algorithms (Cowell et al., 1999).

The junction tree can be used in a similar way to construct efficient schemes for finding the most probable configuration of the hidden variables, given the observed data. The procedure comes down to replacing the sum-operator by the max-operator when propagating messages between cliques (Dawid, 1992). In the context of the illustrating data set on the course of emotions among anorectic patients, a thus applied computational scheme would offer insight in the sequences of emotional states of individual patients.

The innovative part of the paper is mainly situated in the use of logistic regression to model the conditional probability distributions. This way, conditional probabilities can be modelled as a function of a more limited set of parameters. Consequently, the cost of adding an edge in the graph in terms of additional parameters to be estimated can be kept low. For example, suppose that all random variables are binary. In a model with unrestricted conditional probabilities, adding a parent for a random variable results in a doubling of the conditional probabilities to be estimated for that variable. In a restricted model, the number of additional parameters can be reduced by only incorporating a main effect for the new parent, or a main effect and lower-order interaction effects between the new parent and the other parents. An additional advantage of modelling the conditional probabilities using logistic regression is that covariates can be taken into account without increasing the complexity of the associated graph. Finally, even if one is not interested in restricting the conditional probabilities, or in taking into account the effect of covariates, it can be advantageous to model them with a saturated logistic regression model. Unlike probabilities, the logistic regression weights are not bounded between zero and one, so that the distribution of the MLEs will sooner approach normality with increasing sample size, and the asymptotic standard errors obtained from the information matrix will be a better approximation to the true standard errors.

The extension of the scope of latent class models towards more complex data structures and models calls for good model testing and model selection techniques. Unfortunately, model testing for latent class models is complicated by sparseness of data when it comes to testing the global goodness-of-fit of a model, and by the parameters of the restricted model being a nonidentifiable subset of the parameters of the unrestricted model when determining the number of latent states (McLachlan & Peel, 2000). In both cases, one has to rely on the computationally demanding bootstrap technique to approximate the proper reference distribution of the test statistics under the null hypothesis (Aitkin, Anderson, & Hinde, 1981; Collins, Fidler, Wugalter, & Long, 1993; Langeheine, Langeheine, & van de Pol, 1996).

Model diagnostics can be used to assess specific aspects of the model. For the latent Markov and hierarchical latent Markov models, without restrictions on the conditional probabilities and no incorporation of covariates, model diagnostics based on pairwise log-odds ratios revealed that both models underestimated to some degree the associations within the sets of positive and negative emotions. Disadvantages of model diagnostics are that their reference distribution is usually not known, and that they can be chosen “a la tête du client”. To conclude, model testing and model selection should be a concern of future research.

References

- Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, Series A*, *144*, 419–461.
- Baum, L.E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, *41*, 164–171.
- Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*, 579–616.
- Collins, L.M., & Wugalter, S.E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*, 131–157.
- Collins, L.M., Fidler, P.L., Wugalter, S.E., & Long, S.E. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, *28*, 375–389.
- Cowell, R.G., Dawid, A.P., Lauritzen, S.L., & Spiegelhalter, D.J. (1999). *Probabilistic networks and expert systems*. New York: Springer.
- Dawid, A.P. (1992). Applications for a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, *2*, 25–36.
- Delespaul, P. (1995). *Assessing schizophrenia in daily life: The experience sampling method*. Maastricht: Maastricht University Press.
- Diener, E., Smith, H., & Fujita, F. (1995). The personality structure of affect. *Journal of Personality and Social Psychology*, *69*, 130–140.
- Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York: Springer.
- Feldman, L.A. (1995). Valence-focus and arousal-focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, *69*, 153–166.
- Hagenaars, J.A. (1998). Categorical causal modeling: Latent class analysis and directed log-linear models with latent variables. *Sociological Methods and Research*, *26*, 436–486.
- Huang, C., & Darwiche, A. (1996). Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning*, *15*, 225–263.
- Humphreys, K., & Titterton, D.M. (2003). Variational approximations for categorical causal models with latent variables. *Psychometrika*, *68*, 391–412.
- Jensen, F.V., Lauritzen, S.L., & Olesen, K.G. (1990). Bayesian updating in causal probabilistic networks by local computation. *Computational Statistics Quarterly*, *4*, 269–282.
- Langeheine, R., & van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods and Research*, *18*, 416–441.
- Langeheine, R., & van de Pol, F. (1994). Discrete time mixed Markov latent class models. In A. Dale & R.B. Davies (Eds.), *Analyzing social and political change. A casebook of methods* (pp. 170–197). London: Sage.
- Langeheine, R., & van de Pol, F. (2000). Fitting higher order Markov chains. *MPR-online 2000*, *5*. Downloaded at <http://www.mpr-online.de/issue9/> on 18/08/2005.
- Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, *24*, 492–516.
- Lauritzen, S.L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, *19*, 191–201.
- Lauritzen, S.L. (1996). *Graphical models*. Oxford, UK: Clarendon Press.
- Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
- Murphy, K. (2001). The Bayes net toolbox for Matlab. *Computing Science and Statistics*, *33*, 331–350.
- Poulsen, C.S. (1990). Mixed Markov and latent Markov modeling applied to brand choice data. *International Journal of Marketing*, *7*, 5–19.
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika*, *69*, 167–190.
- Skrondal, A., & Rabe-Hesketh, S. (2005). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
- Smyth, P., Heckerman, D., & Jordan, M.I. (1997). Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, *9*, 227–269.
- Vansteelandt, K., Rijmen, F., Pieters, G., & Vanderlinden, J. (2007). Drive for thinness, affect regulation and physical activity in eating disorders: a daily life study. *Behavior Research and Therapy*, *45*, 1717–1734.
- Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*, 213–239.
- Vermunt, J.K., Langeheine, R., & Böckenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, *24*, 178–205.

- Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070.
- Wiggins, L.M. (1973). *Panel analysis: Latent probability models for attitude and behavior processes*. Amsterdam: Elsevier.

Manuscript received 26 OCT 2005

Final version received 31 OCT 2006

Published Online Date: 4 OCT 2007