Latent Class Models in Longitudinal Research

Jeroen K. Vermunt

Department of Methodology and Statistics, Tilburg University

Bac Tran

U.S. Census Bureau

Jay Magidson

Statistical Innovations Inc.

Published in: S. Menard (ed.), Handbook of Longitudinal Research: Design, Measurement, and Analysis, pp. 373-385. Burlington, MA: Elsevier. 2008. (this version contains a few corrections)

Latent Class Models in Longitudinal Research

1 Introduction

This article presents a general framework for the analysis of discrete-time longitudinal data using latent class models. The encompassing model is the mixture latent Markov model, a latent class model with time-constant and time-varying discrete latent variables. The time-constant latent variables are used to deal with unobserved heterogeneity in the change process, whereas the time-varying discrete latent variables are used to correct for measurement error in the observed responses. By allowing for direct relationships between the latent states at consecutive time points, one obtains the typical Markovian transition or first-order autoregressive correlation structure. Moreover, each of three distinct submodels can include covariates, thus addressing separate important issues in longitudinal data analysis: observed and unobserved individual differences, autocorrelation, and spurious observed change resulting from measurement error.

It is shown that most of the existing latent class models for longitudinal data are restricted special cases of the mixture latent Markov model presented, which itself is an expanded version with covariates of the mixed Markov latent class model by Van de Pol and Langeheine (1990). The most relevant restricted special cases are mover-stayer models (Goodman, 1961), mixture Markov models (Poulsen, 1982), latent (or hidden) Markov models (Baum et al., 1970; Collins and Wugalter, 1992; Van de Pol and De Leeuw, 1996; Vermunt, Langeheine, and Böckenholt, 1999; Wiggins, 1973), mixture growth models (Nagin, 1999; Muthén, 2004; Vermunt 2006) and mixture latent growth models (Vermunt 2003, 2006) for repeated measures, as well as the standard multiple-group latent class model for analyzing data from repeated cross-sections (Hagenaars, 1990).

The next section presents the mixture latent Markov model. Then we discuss its most important special cases and illustrate these with an empirical example. We end with a short discussion of various possible extensions of our approach. The first appendix provides details on parameter estimation using the Baum-Welch algorithm. The second appendix contains model setups for the syntax version of the Latent GOLD program (Vermunt and Magidson, 2008) that was used for estimating the example models.

2 The mixture latent Markov model

Assume that we have a longitudinal data set containing measurements for N subjects at $T+1$ occasions. The mixture latent Markov model is a model containing five types of variables: response variables, time-constant explanatory variables, time-varying explanatory variables, time-constant discrete latent

variables, and time-varying discrete latent variables. For simplicity of exposition, we will assume that response variables are categorical, and that there is at most one time-constant and one time-varying latent variable. These are, however, not limitations of the framework we present which can be used with continuous response variables, multiple time-constant latent variables, and multiple time-varying latent variables. Our mixture latent Markov model is an expanded version of the mixed Markov latent class model proposed by Van de Pol and Langeheine (1990): it contains time-constant and time-varying covariates and it can be used when the number of time points is large.

Let y_{itj} denote the response of subject i at occasion t on response variable j, where $1 \leq i \leq N$, $0 \leq t \leq T$, $1 \leq j \leq J$, and $1 \leq y_{itk} \leq M_j$. Note that J is the total number of response variables and M_j the number of categories for response variable j. The vector of responses for subject i at occasion t is denoted as $\mathbf{y}_{it}$ and the vector of responses at all occasions as $\mathbf{y}_i$. The vector of time-constant and time-varying predictors at occasion t is denoted by $\mathbf{z}_i$ and $\mathbf{z}_{it}$, respectively. The time-constant and time-varying discrete latent variables are denoted by w and x_t, where $1 \leq w \leq L$ and $1 \leq x_t \leq K$. The latter implies that the number of categories of the two types of latent variables equal L and K, respectively. To make the distinction between the two types of latent variables clear, we will refer to w as a latent class and to x_t as a latent state.

The general model that we use as the starting point is the following mixture latent Markov model:

$$P(\mathbf{y}_i|\mathbf{z}_i) = \sum_{w=1}^{L} \sum_{x_0=1}^{K} \sum_{x_1=1}^{K} ... \sum_{x_T=1}^{K} P(w, x_0, x_1, ..., x_T|\mathbf{z}_i)\, P(\mathbf{y}_i|w, x_0, x_1, ..., x_T, \mathbf{z}_i), \tag{1}$$

with

$$P(w, x_0, x_1, ..., x_T|\mathbf{z}_i) = P(w|\mathbf{z}_i)\, P(x_0|w, \mathbf{z}_{i0}) \prod_{t=1}^{T} P(x_t|x_{t-1}, w, \mathbf{z}_{it}), \tag{2}$$

$$P(\mathbf{y}_i|w, x_0, x_1, ..., x_T, \mathbf{z}_i) = \prod_{t=0}^{T} P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it}) = \prod_{t=0}^{T} \prod_{j=1}^{J} P(y_{itj}|x_t, w, \mathbf{z}_{it}) \tag{3}$$

As many statistical models, the model in Equation (1) describes $P(\mathbf{y}_i|\mathbf{z}_i)$, the (probability) density associated with responses of subject i conditional on his/her observed covariate values. The right-hand side of this equation shows that we are dealing with a mixture model containing 1 time constant latent variable and $T+1$ time-varying latent variables. The total number of mixture components (or latent classes) equals $L \cdot K^{T+1}$, which is the product of the number of categories of w and x_t for $t = 0, 1, 2, ..., T$. As in any mixture model, $P(\mathbf{y}_i|\mathbf{z}_i)$ is obtained as a weighted average of class-specific probability densities – here $P(\mathbf{y}_i|w, x_0, x_1, ..., x_T, \mathbf{z}_i)$ – where the (prior) class membership probabilities or mixture proportions – here $P(w, x_0, x_1, ..., x_T|\mathbf{z}_i)$ – serve as weights (Everitt and Hand, 1981; McLachlan and Peel, 2000).

Equations (2) and (3) show the specific structure assumed for the mixture proportion $P(w, x_0, x_1, ..., x_T|\mathbf{z}_i)$ and the class-specific densities $P(\mathbf{y}_i|w, x_0, x_1, ..., x_T, \mathbf{z}_i)$. The equation for $P(w, x_0, x_1, ..., x_T|\mathbf{z}_i)$ assumes that conditional on w and $\mathbf{z}_i$, x_t is associated only with x_{t-1} and x_{t+1} and thus not with the states occupied at the other time points – the well-know first-order Markov assumption. The equation for $P(\mathbf{y}_i|w, x_0, x_1, ..., x_t, \mathbf{z}_i)$ makes two assumptions: 1) conditionally on w, x_t, and $\mathbf{z}_{it}$, the J responses at occasion t are independent of the latent states and the responses at other time points, and 2) conditionally on w, x_t, and $\mathbf{z}_{it}$, the J responses at occasion time point t are mutually independent, which is referred to as the local independence assumption in latent class analysis (Goodman, 1974).

As can be seen from Equations (2) and (3), the models of interest contain four different kinds of model probabilities:

- $P(w|\mathbf{z}_i)$ is the probability of belonging to a particular latent class conditional on a person's covariate values,
- $P(x_0|w, \mathbf{z}_{i0})$ is an initial-state probability; that is, the probability of having a particular latent initial state conditional on an individual's class membership and covariate values at $t = 0$,
- $P(x_t|x_{t-1}, w, \mathbf{z}_{it})$ is a latent transition probability; that is, the probability of being in a particular latent state at time point t conditional on the latent state state at time point $t - 1$, class membership, and time-varying covariate values,
- $P(y_{itj}|x_t, w, \mathbf{z}_{it})$ is a response probability, which is the probability of having a particular observed value on response variable j at time point t conditional on the latent state occupied at time point t, class membership w, and time-varying covariate values.

Typically, these four sets of probabilities will be parameterized and restricted by means of (logistic) regression models. This is especially useful when a model contains covariates, where time itself may be one of the time-varying covariates of main interest. In the empirical application presented below we will use such regression models. For extended discussions on logistic regression analysis, we refer to introductory texts on this topic (see, for example, Agresti, 2002; Menard, 1995; Vermunt, 1997).

The three key elements of the mixture latent Markov model described in Equations (1), (2), and (3) are that it can take into account 1) unobserved heterogeneity, 2) autocorrelation, and 3) measurement error. Unobserved heterogeneity is captured by the time-constant latent variable w, autocorrelations are captured by the first-order Markov transition process in which the state at time point t may depend on the state at time point $t - 1$, and measurement error or misclassification is accounted for allowing an imperfect relationship between the time-specific latent states x_t and the observed

responses y_{itj}. Note that these are three of the main elements that should be taken into account in the analysis of longitudinal data; that is, the inter-individual variability in patterns of change, the tendency to stay in the same state between consecutive occasions, and spurious change resulting from measurement error in observed responses.

Parameters of the mixture latent Markov model can be estimated by means of maximum likelihood (ML). For that purpose, it advisable to use a special variant of the expectation maximization (EM) algorithm that is usually referred to as the forward-backward or Baum-Welch algorithm (Baum et al., 1970; McDonald and Zucchini, 1997) which is described in detail in the first appendix. This special algorithm is needed because our model contains a potentially huge number of entries in the joint posterior latent distribution $P(w, x_0, x_1, ..., x_T|\mathbf{y}_i, \mathbf{z}_i)$, except in cases where T, L and K are all small. For example, in a fairly moderate sized situation where $T = 10$, $L = 2$ and $K = 3$, the number of entries in the joint posterior distribution already equals $2 \cdot 3^{11} = 354294$, a number which is impossible to process and store for all N subjects as has to be done within standard EM. The Baum-Welch algorithm circumvents the computating of this joint posterior distribution making use of the conditional independencies implied by the model. Vermunt (2003) proposed a slightly simplified version of the Baum-Welch algorithm for dealing with the multilevel latent class model, which when used for longitudinal data analysis is one of the special cases of the mixture latent Markov model described in the next section.

A common phenomenon in the analysis of longitudinal data is the occurrence of missing data. Subjects may have missing values either because they refused to participate at some occasions or because it is elected by the study design. A nice feature of the approach described here is that it can easily accommodate missing data in the ML estimation of the unknown model parameter. Let δ_{it} be an indicator variable taking on the value 1 if subject i provides information for occasion t and 0 if this information is missing. The only required change with missing data is the following modification of Equation (3):

$$P(\mathbf{y}_i|w, x_0, x_1, ..., x_T, \mathbf{z}_i) = \prod_{t=0}^{T} \left[P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})\right]^{\delta_{it}} .$$

For $\delta_{it} = 1$, nothing changes compared to what we had before. However, for $\delta_{it} = 0$, the time-specific conditional density becomes 1, which means that the responses of a time point with missing values are skipped. Actually, for each pattern of missing data, we have a mixture latent Markov for a different set of occasions. Two limitations of the ML estimation procedure with missing values should be mentioned: 1) it can deal with missing values on response variables, but not with missing values on covariates, and 2) it assumes that the missing data are missing at random (MAR). The first limitation may be problematic when there are time-varying covariates for which the values are also missing. However, in various special cases discussed below – the ones

that do not use a transition structure – it is not a problem if time-varying covariates are missing for the time points in which the responses are missing. The second limitation concerns the assumed missing data mechanism: MAR is the least restrictive mechanism under which ML estimation can be used without the need of specifying the exact mechanism causing the missing data; that is, under which the missing data mechanism is ignorable for likelihood-based inference (Little and Rubin, 1987; Schafer, 1997). It is possible to relax the MAR assumption by explicitly defining a not missing at random (NMAR) mechanism as a part of the model to be estimated (Fay, 1986; Vermunt 1997).

An issue strongly related to missing data is the one of unequally spaced measurement occasions. As long as the model parameters defining the transition probability are assumed to be occasion specific, no special arrangements are needed. If this is not the case, unequally spaced measurements can be handled by defining a grid of equally spaced time points containing all measurement occasions. Using this technique, the information on the extraneous occasions can be treated as missing data for all subjects. An alternative is to use a continuous-time rather than a discrete time framework (Böckenholt, 2006), which can be seen as the limiting case in which the elapsed time between consecutive time points in the grid approaches zero.

Another issue related to missing data is the choice of the time variable and the corresponding starting point of the process. The most common approach is to use calender time as the time variable and the first measurement occasion as $t = 0$, but one may, for example, also use age as the relevant time variable, as we do in the empirical example. Although children's ages at the first measurement vary between 11 and 17, we use age 11 as $t = 0$. This implies that for a child that is 12 years of age information at $t = 0$ is treated as missing, for a child that is 13 years of age information a $t = 0$ and $t = 1$ is treated as missing, etc..

3 The most important special cases

[INSERT TABLE 1 ABOUT HERE]

Table 1 lists the various special cases that can be derived from the mixture latent Markov model defined in Equations (1)-(3) by assuming that one or more of its three elements – transition structure, measurement error, and unobserved heterogeneity – is not present or needs to be ignored because the data is not informative enough to deal with it. Model I-III and V-VII are latent class models, but IV and VIII are not. Model VII differs from models I-VI in that it is model for repeated cross-sectional data rather than a model for panel data. Below we describe the various special cases in more detail.

Mixture latent Markov First of all, it is possible to define simpler versions of the mixture latent Markov model itself. Actually, the mixed Markov

latent class model proposed by Van de Pol and Langeheine (1990) which served as an inspiration for our model is the special case of our model when neither time-constant nor time-varying covariates are present. Van de Pol and Langeheine (1990) also proposed a variant in which the four types of model probabilities could differ across categories of a grouping variable (see also Langeheine and Van de Pol, 2002). A similar model is obtained by replacing the $\mathbf{z}_i$ and $\mathbf{z}_{it}$ in Equations (1)-(3) by a single categorical covariate z_i.

Mixture Markov The mixture Markov model (Poulsen, 1982) is the special case of the model presented in Equations (1)-(3) when there is a single response variable that is assumed to be measured without error. The model is obtained by replacing the more general definition in Equation (3) with

$$P(\mathbf{y}_i|w, x_0, x_1, ..., x_T, \mathbf{z}_i) = \prod_{t=0}^{T} P(y_{it}|x_t),$$

where $K = M$ and $P(y_{it}|x_t) = 1$ if $x_t = y_{it}$ and 0 otherwise. The product over the multiple response variables and the index j can be omitted because $J = 1$ and y_{it} is assumed not to depend on w and $\mathbf{z}_{it}$ but only on x_t. For this special case the number of latent states (K) is equal to the number of observed states (M) and the relationship between x_t and y_{it} is perfect, which indicates that x_t is measured without error.

A special case of this mixture Markov model is the mover-stayer model (Goodman, 1961). This model assumes that $L = 2$ and that the transition probabilities are fixed to 0 for one class, say for $w = 2$. Members of this class, for which $P(x_t|x_{t-1}, w = 2, \mathbf{z}_{it}) = 1$ if $x_t = x_{t-1}$ and 0 otherwise, are called stayers. Note that the mover-stayer constraint can not only be imposed in the mixture Markov but also in the mixture latent Markov, in which case transitions across imperfectly measured stated are assumed not to occur in the stayer class.

Because of the perfect match between x_t and y_{it}, the mixture Markov model can also be defined without latent states x_t; that is, as:

$$P(\mathbf{y}_i|\mathbf{z}_i) = \sum_{w=1}^{L} P(w|\mathbf{z}_i)\, P(y_{i0}|w, \mathbf{z}_i) \prod_{t=1}^{T} P(y_{it}|y_{it-1}, w, \mathbf{z}_{it}).$$

Latent Markov model The latent Markov, latent transition, or hidden Markov model (Baum et al., 1970; Collins and Wugalter, 1992; Van de Pol and De Leeuw, 1996; Vermunt, Langeheine, and Böckenholt, 1999, Wiggins, 1973) is the special case of the mixture latent Markov that is obtained by eliminating the time-constant latent variable w from the model, that is, by assuming that there is no unobserved heterogeneity or that it can be ignored. The latent Markov model can be obtained without modifying the formulae, but by simply assuming that $L = 1$; that is, that all subject belong to the same latent class.

The latent Markov model yields estimates for the initial-state and transition probabilities, as well as for how these are affected by covariate values, while correcting for measurement error in the observed states. The model can be applied with a single or with multiple response variables. When applied with a single categorical response variable, one will typically assume that the number of latent states equals the number or categories of the response variable: $K = M$. Moreover, model restrictions are required to obtain an identified model, the most common of which are time-homogeneous transition probabilities or time-homogeneous misclassification probabilities.

When used with multiple indicators, the model is a longitudinal data extension of the standard latent class model (Hagenaars, 1990). The time-specific latent states can be seen as clusters or types which differ in their responses on the J indicators, and the Markovian transition structure is used to describe and predict changes that may occur across adjacent measurement occasions.

Markov model By assuming both perfect measurement as in the mixture Markov model and absence of unobserved heterogeneity as in the latent Markov model, one obtains a standard Markov model, which is no longer a latent class model. This model can further serve as a simple starting point for longitudinal applications with a single response variable, where one wishes to assume a Markov structure. It provides a baseline for comparison to the three more extended models discussed above. Use of these more extended models makes sense only if they provide a significantly better description of the data than the simple Markov model.

Mixture latent growth model Now we turn to latent class models for longitudinal research that are not transition or Markov models. These mixture growth models are non-parametric random-effects models (Aitkin, 1999, Skrondal and Rabe-Hesketh, 2004; Vermunt and Van Dijk, 2002) for longitudinal data that assume that dependencies between measurement occasions can be captured by the time-constant latent variable w. The most extended variant is the mixture latent growth model, which is obtained from the mixture latent Markov model by imposing the constraint $P(x_t|x_{t-1}, w, \mathbf{z}_{it}) = P(x_t|w, \mathbf{z}_{it})$. This is achieved by replacing Equation (2) with

$$P(w, x_0, x_1, ..., x_T|\mathbf{z}_i) = P(w|\mathbf{z}_i) \prod_{t=0}^{T} P(x_t|w, \mathbf{z}_{it}).$$

This model is a variant for longitudinal data of the multilevel latent class model proposed by Vermunt (2003): subjects are the higher-level units and time points the lower-level units. It should be noted that application of this very interesting model requires that there be at least two response variables ($J \geq 2$).

In mixture growth models one will typically pay a lot of attention to the modeling of the time dependence of the state occupied at the different time

points. The latent class or mixture approach allows identifying subgroups (categories of the time-constant latent variable w) with different change patterns (Nagin, 1999). The extension provided by the mixture latent growth model is that the dynamic dependent variable is itself a (discrete) latent variable which is measured by multiple indicators.

Mixture growth model The mixture or latent class growth model (Nagin, 1999, Muthén, 2004; Vermunt, 2006) can be seen as a restricted variant of the mixture latent growth model; i.e., as a model for a single indicator measured without error. The extra constraint is the same as the one used in the mixture Markov model: $K = M$ and $P(y_{it}|x_t) = 1$ if $x_t = y_{it}$ and 0 otherwise.

A more natural way to define the mixture growth model is by omitting the time-varying latent variable x_t from the model specification, as we did for the mixture Markov model. This yields

$$P(\mathbf{y}_i|\mathbf{z}_i) = \sum_{w=1}^{L} P(w|\mathbf{z}_i) \prod_{t=0}^{T} P(y_{it}|w, \mathbf{z}_{it}),$$

Note that this model is equivalent to a standard latent class model for $T+1$ response variables and with predictors affecting these responses.

Standard latent class model When we eliminate both w and the transition structure, we obtain a latent class model that assumes observations are independent across occasions. This is a realistic model only for the analysis of data from repeated cross-sections; that is, to deal with the situation in which observations from different occasions are independent because each subject provides information for only one time point. One possible way to define this model is

$$P(\mathbf{y}_i|\mathbf{z}_{it_i}) = \sum_{x=1}^{K} P(x|\mathbf{z}_{it_i}) \prod_{j=1}^{J} P(y_{itj}|x, \mathbf{z}_{it_i}),$$

where t_i is used to denote the time point for which subject i provides information. This is a standard latent class model with covariates.

4 Application to NYS data

To illustrate the latent class models described above we use data from the nine-wave National Youth Survey (Elliott, Huizinga, and Menard, 1989) for which data were collected annually from 1976 to 1980 and at three year intervals after 1980. At the first measurement occasion, the ages of the 1725 children varied between 11 and 17. To account for the unequal spacing across panel waves and to use age as the time scale, we define a model for 23 time points ($T+1 = 23$), where $t = 0$ corresponds to age 11 and the last time point to age 33. For each subject, we have observed data for at most 9 time points

(the average is 7.93) which means that the other time points are treated as missing values.

We study the change in a dichotomous response variable "drugs" indicating whether young persons used hard drugs during the past year (1=no; 2=yes). It should be noted that among the 11 year olds in the sample nobody reported to have used hard drugs, which is something that needs to be taken into account in our model specification. Time-varying predictors are age and age squared, and time-constant predictors are gender and ethnicity.

A preliminary analysis showed that there is a clear age-dependence in the reported hard-drugs use which can well be described by a quadratic function: usage first increases with age and subsequently decreases. That is why we used this type of time dependence in all reported models. To give an idea how the time dependence enters in the models, the specific regression model for the latent transition probabilities in the estimated Markov models was:

$$\log \frac{P(x_t = k'|x_{t-1} = k, w, \text{age}_{it})}{P(x_t = k|x_{t-1} = k, w, \text{age}_{it})} = \beta_{0k'k} + \beta_{1k'k} \cdot d_{w=2} + \beta_{2k'k} \cdot \text{age}_{it} + \beta_{3k'k} \cdot (\text{age}_{it})^2,$$

where the β coefficients are fixed to 0 for $k' = k$. The variable $d_{w=2}$ is a dummy variable for the second mixture component. For the initial-state, we do not have a model with free parameters but we simply assume that all children start in the no-drugs state at age 11.

In the mixture growth models, we use the following binary logistic regression model for y_{it}:

$$\log \frac{P(y_{it} = 2|w, \text{age}_{it})}{P(y_{it} = 1|w, \text{age}_{it})} = \beta_{0w} + \beta_{1w} \cdot \text{age}_{it} + \beta_{2w} \cdot (\text{age}_{it})^2,$$

where we fix $\beta_{01} = -100$ and $\beta_{11} = \beta_{21} = 0$ to obtain a model in which $w = 1$ represents a non-user class, a class with a zero probability of using drugs at all time points.

Table 2 reports the fit measures for the estimated models, where the first set of models do not contain time-constant covariates gender and ethnicity. As can be seen from log-likelihood and BIC values, the various types of Markov models perform much better than the mixture growth models, which indicates that there is a clear autocorrelation structure that is difficult to capture using a growth model. Even with 7 latent classes one does not obtain a fit that is as good as the Markov-type models. Among the Markov models, the most general model – the mixture latent Markov model – performs best. By removing measurement error, simplifying the mixture into a mover-stayer structure, and/or eliminating the mixture structure, the fit deteriorates significantly. The last two models are mixture latent markov models in which we introduced covariates in the model for the mixture proportions. Both sex and ethnicity seem to be significantly related to the mixture component someone belongs to.

The parameters of the final model consist of the logit coefficients of the model for w, the logit coefficients in the model for the latent transition probabilities, and the probabilities of the measurement model. The latter show

that the two latent states are rather strongly connected to the two observed states: $P(y_{it} = 1|x_t = 1) = 0.99$ and $P(y_{it} = 2|x_t = 2) = 0.87$.

The most relevant coefficients in the model for the transition probabilities are the parameters for w. These show that class 2 is the low-risk class having a much lower probability than class 1 of entering into the use state ($\beta = -2.37; S.E = 0.26$) and a much higher probability of leaving the non-use state ($\beta = 3.72; S.E. = 0.68$). Combining these estimates with the quadratic time dependence of the transitions yields a probability of moving from the non-use to the use state equal to 2.8% at age 12, 23.4% at age 21, and 0.6% at age 33 for $w = 1$, and equal to 0.3% at age 12, 2.8% at age 21, and 0.1% at age 33 for $w = 2$. The probability of a transition from the use to the non-use state equals 0.1% at age 12, 20.5% at age 26, and 6.2% for $w = 1$, and 4.1% at age 12, 91.4% at age 26, and 73.1% at age 33 for $w = 2$.

The parameters in the logistic regression model for w shows that males are less likely to be in the low-risk class than females ($\gamma = -0.58; S.E. = 0.14$) and that blacks are more likely to be in the low-risk class than whites ($\gamma = 0.79; S.E = 0.22$). Hispanics are less likely ($\gamma = -0.46; S.E. = 0.33$) and other ethnic groups more likely ($\gamma = 0.25; S.E = 0.52$) to be in class 2 than white, but these effect are non significant.

5 Discussion

We presented a general framework for the analysis of discrete-time longitudinal data and illustrated it with an empirical example in which the Markov-like models turned out to perform better than the growth models.

The approach presented here can be expanded in various ways. First, while we focused on models for categorical response variables, it is straightforward to apply most of these models to variables of other scale types, such as continuous dependent variables or counts. Other extensions include the definition of multiple processes with multiple x_t or of higher-order Markov processes. Models that are getting increased attention are those that combine discrete and continuous latent variables. Finally, the approach can be expanded to deal with multilevel longitudinal data, as well as with data obtained from complex survey samples. Each of these extensions is implemented in the Latent GOLD software that we used for parameter estimation.

Appendix A: Baum-Welch algorithm for the mixture latent Markov model

Maximum likelihood (ML) estimation of the parameters of the mixture latent Markov model involves maximizing the log-likelihood function:

$$L = \sum_{i=1}^{N} \log P(\mathbf{y}_i|\mathbf{z}_i),$$

a problem that can be solved by means of the EM algorithm (Dempster, Laird and Rubin, 1977). In the E step, we compute

$$P(w, x_0, x_1, ..., x_T|\mathbf{y}_i, \mathbf{z}_i) = \frac{P(w, x_0, x_1, ..., x_T, \mathbf{y}_i|\mathbf{z}_i)}{P(\mathbf{y}_i|\mathbf{z}_i)},$$

which is the joint conditional distribution of the $T + 2$ latent variables given the data and the model parameters. In the M step, one updates the model parameters using standard ML methods for logistic regression analysis and using an expanded data matrix with $P(w, x_0, x_1, ..., x_T|\mathbf{y}_i, \mathbf{z}_i)$ as weights.

It should be noted that in a standard EM algorithm, at each iteration, one needs to compute and store the $L \cdot K^{T+1}$ entries of $P(w, x_0, x_1, ..., x_T|\mathbf{y}_i, \mathbf{z}_i)$ for each subject or, with grouped data, for each unique data pattern. This implies that computation time and computer storage increases exponentially with the number of time points, which makes this algorithm impractical or even impossible to apply with more than a few time points (Vermunt, Langeheine and Böckenholt, 1999). However, because of the collapsibility of the mixture latent Markov model, it turns out that in the M step of the EM algorithm one needs only the marginal distributions $P(w|\mathbf{y}_i, \mathbf{z}_i)$, $P(w, x_t|\mathbf{y}_i, \mathbf{z}_i)$, and $P(w, x_{t-1}, x_t|\mathbf{y}_i, \mathbf{z}_i)$. The Baum-Welch or forward-backward algorithm obtains these quantities directly rather than first computing $P(w, x_0, x_1, ..., x_T|\mathbf{y}_i, \mathbf{z}_i)$ and subsequently collapsing over the remaining dimensions as would be done in a standard EM algorithm (Baum et al, 1970; McDonald and Zucchini, 1997). This yields an algorithm that makes the mixture latent Markov model applicable with any number of time points. Whereas the original forward-backward algorithm was for latent (hidden) Markov models without covariates and a single response variable, here we provide a generalization to the more general case with a mixture w, covariates $\mathbf{z}_i$, and multiple responses.

The two key components of the Baum-Welch algorithm are the forward probabilities α_{iwx_t} and the backward probabilities β_{iwx_t}. Because of our generalization to the mixture case, we need an additional quantity γ_{iw}. These three quantities are defined as follows:

$$\begin{aligned} \alpha_{iwx_t} &= P(x_t, \mathbf{y}_{i0}...\mathbf{y}_{it}|w, \mathbf{z}_i), \\ \beta_{iwx_t} &= P(\mathbf{y}_{i(t+1)}...\mathbf{y}_{iT}|x_t, w, \mathbf{z}_i), \\ \gamma_{iw} &= P(w, \mathbf{y}_i|\mathbf{z}_i). \end{aligned}$$

Using α_{iwx_t}, β_{iwx_t}, and γ_{iw}, one can obtain the relevant marginal posteriors as follows:

$$P(w|\mathbf{y}_i, \mathbf{z}_i) = \frac{\gamma_{iw}}{P(\mathbf{y}_i|\mathbf{z}_i)}, \tag{4}$$

$$P(w, x_t|\mathbf{y}_i, \mathbf{z}_i) = \frac{\alpha_{iwx_t}\beta_{iwx_t}}{P(\mathbf{y}_i|\mathbf{z}_i)}, \tag{5}$$

$$P(w, x_{t-1}, x_t, w|\mathbf{y}_i, \mathbf{z}_i) = \frac{\gamma_{iw}\alpha_{iwx_{t-1}}P\left(x_t|x_{t-1}, w, \mathbf{z}_{it}\right)P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})\,\beta_{iwx_t}}{P(\mathbf{y}_i|\mathbf{z}_i)} \tag{6}$$

where $P(\mathbf{y}_i|\mathbf{z}_i) = \sum_{w=1}^{L} \gamma_{iw}$, and $P\left(x_t|x_{t-1}, w, \mathbf{z}_{it}\right)$ and $P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})$ are model probabilities.

The key element of the forward-backward algorithm is that $T+1$ sets of α_{iwx_t} and β_{iwx_t} terms are computed using recursive schemes. The forward recursion scheme for α_{iwx_t} is:

$$\begin{aligned}
\alpha_{iwx_0} &= P(x_0|w, \mathbf{z}_{i0}) P(y_{i0}|x_0, w, \mathbf{z}_{i0}), \\
\alpha_{iwx_t} &= \left\{ \sum_{x_{t-1}=1}^{K} \alpha_{iwx_{t-1}} P\left(x_t|x_{t-1}, w, \mathbf{z}_{it}\right) \right\} P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it}),
\end{aligned}$$

for $t = 1$ up to $t = T$. The backward recursion scheme for β_{iwx_t} is:

$$\begin{aligned}
\beta_{iwx_T} &= 1, \\
\beta_{iwx_t} &= \sum_{x_{t+1}=1}^{K} \beta_{iwx_{t+1}} P\left(x_{t+1}|x_t, w, \mathbf{z}_{it}\right) P(\mathbf{y}_{it+1}|x_{t+1}, w, \mathbf{z}_{it}),
\end{aligned}$$

for $T-1$ down to $t = 0$. The quantity γ_{iw} is obtained as:

$$\gamma_{iw} = \sum_{x_t=1}^{K} P(w|\mathbf{z}_i)\, \alpha_{iwx_t}\, \beta_{iwx_t},$$

for any t. So, first we obtain α_{iwx_t} and β_{iwx_t} for each time point and subsequently we obtain γ_{iw}. Next, we compute $P(w|\mathbf{y}_i, \mathbf{z}_i)$, $P(w, x_t|\mathbf{y}_i, \mathbf{z}_i)$, and $P(w, x_{t-1}, x_t|\mathbf{y}_i, \mathbf{z}_i)$ using Equations (4), (5), and (6). In the M step, these quantities are used to obtain new estimates for the mixture latent Markov model probabilities appearing in Equations (2) and (3) using standard methods for logistic regression analysis.

The only change required in the above formulas when there is missing data is that $P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})$ is replaced by $P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})^{\delta_{it}}$ in each of the above equations, where $\delta_{it} = 1$ if $\mathbf{y}_{it}$ is observed and 0 if $\mathbf{y}_{it}$ is missing. This implies that $P(\mathbf{y}_{it}|x_t, w, \mathbf{z}_{it})$ is "skipped" when $\mathbf{y}_{it}$ is missing. In the M step, cases with missing responses at occasion t do not contribute to the estimation of the response probabilities for that occasion, but they do contribute to the estimation of the other model probabilities.

Appendix B: examples of Latent GOLD syntax files

The Latent GOLD 4.5 software package (Vermunt and Magidson, 2008) implements the framework described in this article. In this appendix, we provide examples of input files for estimation of mixture latent Markov models, mixture Markov, latent Markov, and mixture growth models.

The data should be in the format of a person-period file, where for the Markov type models periods with missing values should also be included in

the file since each next record for the same subject is assumed to be the next time point. The definition of a model contains three main sections: "options", "variables" and "equations".

An example of the most extended model, the mixture latent Markov model is the following:

```
options
  missing includeall;
  output parameters=first standarderrors;
variables
  caseid id;
  dependent drugs nominal;
  independent gender nominal, ethnicity nominal, age numeric, age2 numeric;
  latent W nominal 2, X nominal markov 2;
equations
  W <- 1 + gender + ethnicity;
  X[=0] <- (-100) 1;
  X <- (a~tra) 1 | X[-1] + (b~tra) W | X[-1] + (c~tra) age | X[-1] + (d~tra) age2 | X[-1];
  drugs <- (e~err) 1 | X;
```

In the `options` section, only the two commands for which we changed the default setting is shown. The statement "`missing=includeall`" indicates that all records with missing values should be retained in the analysis. The output option "`parameters=first`" request dummy coding for the nominal variables using the first category as the reference category.

In the `variables` section we define the caseid variable connecting the multiple records of one person, the latent, dependent (or response) and independent variables to be used in the analysis, as well as various attributes of these variables, such as their scale type and, for categorical latent variables, their number of categories and whether they vary over time (indicated with the statement `markov`).

The `equation` section contains 4 equations: one for the mixture variable (`W`), one for the initial state (`X[=0]`), one for the state at time point t (`X`) conditional on the state at $t-1$ (`X[-1]`), and one for the response variable. With more response variables, one would have a separate equation for each response variable. The logit model for `W` contains an intercept (the term "`1`") and effects of gender and ethnicity. The parameter labels, `a`, `b`, `c`, `d`, and `e` are given in parentheses. The model for `X[=0]` contains an intercept that is fixed to -100, which means that everyone starts in latent state 1. The model for `X` is parameterized in such a way that the intercept and the effects of `W`, `age`, and `age2` can be interpreted as effects on the logit of a transition (as in the equation provided in the text). This is achieved by the conditioning "`| X[-1]`" combined with "~tra" in the parameter label, which yields a special transition coding of logit coefficients in which the no change category serves as the reference category. The model for the response variable `drugs` contains an intercept which varies across latent states, with the same type of coding as used for the transition (for the dependent variable called error coding).

A mixture Markov is obtained with the extra line "`e = -100;`". This fixes the logit parameters in the model for the response variable to -100, which

because of the special error coding (induced with "~err") yields a perfect relationship between `X` and `drugs`. The 2-class mixture can be changed into a mover-stayer structure with the additional line "`b = -100;`" which fixes the transition probabilities to 0 for the second class. This restriction can be used in the mixture Markov and in the mixture latent Markov model. A latent Markov model is obtained either by removing `W` from the `variables` and `equations` sections or by setting its number of categories to 1.

A mixture growth model is obtained by removing `X` from the variables section and replacing the `equations` section with the following:

```
equations
  W <- 1 + gender + ethnicity;
  drugs <- (a) 1 | W + (b) age | W + (c) age2 | W;
  a[1] = -100;
  b[1] = 0;
  c[1] = 0;
```

The constraint on the intercept indicates that the first mixture component does not use drugs with probability 1. The other two constraints fix the redundant `age` and `age2` effects for class one equal to 0.

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

Aitkin (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55, 218-234.

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

Böckenholt, U. (2005). A latent Markov model for the analysis of longitudinal data collected in continuous time: States, durations, and transitions, *Psychological Methods*, 10, 65-82.

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

Elliott, D.S., Huizinga, D., and Menard, S. (1989). *Multiple Problem Youth: Delinquency, Substance Use, and Mental Health Problems*. New York: Springer-Verlag.

Everitt, B.S., and Hand, D.J. (1981) *Finite Mixture Distributions*. London: Chapman & Hall.

Fay, R.E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, 81, 354-365.

Goodman, L.A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56, 841-868.

Goodman, L.A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable: Part I - A modified latent structure approach. *American Journal of Sociology*, 79, 1179-1259.

Hagenaars, J.A. (1990). *Categorical Longitudinal Data - Loglinear Analysis of Panel, Trend and Cohort Data.* Newbury Park: Sage.

Langeheine, R. and Van de Pol, F. (2002). Latent Markov chains. J.A. Hagenaars and A.L. McCutcheon (eds.), *Applied Latent Class Analysis*, 304-341. Cambrigde University Press.

Little, R.J., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

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

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

Menard, S, (1995). *Applied Logistic Regression Analysis.* Thousand Oakes, CA: Sage.

Muthén, B. (2004). Latent variable analysis. Growth mixture modeling and related techniques for longitudinal data. D. Kaplan (ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Chapter 19, 345-368. Thousand Oakes: Sage Publications.

Nagin, D.S. (1999). Analyzing developmental trajectories: a semiparametric group-based approach. *Psychological Methods*, 4, 139-157.

Poulsen, C.S. (1982). *Latent structure analysis with choice modeling applications.* Arhus: The Arhus School of Business Administration and Economics.

Schafer, J.L. (1997). *Statistical Analysis with Incomplete Data.* London: Chapman & Hall.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models.* London: Chapman & Hall/CRC.

Van de Pol, F., and De Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research*, 15, 118-141.

Van de Pol, F., and Langeheine, R. (1990) Mixed Markov latent class models. *Sociological Methodology*, 213-247.

Vermunt, J.K. (1997). *Log-linear Models for Event Histories*. Thousand Oakes: Sage.

Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 33, 213-239.

Vermunt, J.K. (2006). Growth models for categorical response variables: standard, latent-class, and hybrid approaches. K. van Montfort, H. Oud, and A. Satorra (eds.). *Longitudinal Models in the Behavioral and Related Sciences*. Erlbaum.

Vermunt, J.K. Langeheine, R., and Böckenholt, U. (1999). Latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24, 178-205.

Vermunt, J. K. and Magidson, J. (2008). *LG-syntax user's guide: Manual for Latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K. and Van Dijk. L. (2001). A nonparametric random-coefficients approach: the latent class regression model. *Multilevel Modelling Newsletter*, 13, 6-13.

Wiggins, L.M. (1973). *Panel analysis.* Amsterdam: Elsevier.

Table 1: Classification of latent class models for longitudinal research

Model name		Transition structure	Unobserved heterogeneity	Measurement error
I.	Mixture latent Markov	yes	yes	yes
II.	Mixture Markov	yes	yes	no
III.	Latent Markov	yes	no	yes
IV.	Standard Markov*	yes	no	no
V.	Mixture latent growth	no	yes	yes
VI.	Mixture growth	no	yes	no
VII.	Standard latent class	no	no	yes
VIII.	Independence*	no	no	no

*: This model is not a latent class model.

Table 2: Fit measures for the estimated models with the nine-wave National Youth Survey data set

Model	Log-likelihood	BIC	# Parameters
A. Independence	-5089	10200	3
B. Markov	-4143	8330	6
C. Mixture Markov with L=2	-4020	8108	9
D. Mover-stayer Markov	-4056	8165	7
E. Latent Markov with K=2	-4009	8078	8
F. Mixture latent Markov with L=2 and K=2	-3992	8066	11
G. Mover-stayer latent Markov with K=2	-4000	8068	9
H1. Mixture growth with L=2 ($w = 1$ non-users)	-4381	8792	4
H2. Mixture growth with L=3 ($w = 1$ non-users)	-4199	8457	8
H3. Mixture growth with L=4 ($w = 1$ non-users)	-4113	8315	12
H4. Mixture growth with L=5 ($w = 1$ non-users)	-4077	8273	16
H5. Mixture growth with L=6 ($w = 1$ non-users)	-4037	8223	20
H6. Mixture growth with L=7 ($w = 1$ non-users)	-4024	8227	24
I. F + Gender effect on W	-3992	8066	12
J. F + Gender and Ethnicity effect on W	-3975	8061	15