

Summary

A Monte Carlo Simulation Study on Model Selection in Latent Markov Models

Duygu Güngör

Dokuz Eylül University

Selva Ülbe

Dokuz Eylül University

Samet Baş

Dokuz Eylül University

Time-dependent change has always been the subject of many psychological researches. In recent years, latent growth models, which are part of structural equation models, have been used for investigating time-dependent change in the case of continuous latent variable(s). However, it is not always possible to make continuous measurements in psychological research. For instance, in the study of driver behaviors, it may not be possible for a researcher who wants to rate risky behaviors to decide which of behaviors of crossing the red light or exceeding the speed limit is riskier. As in this example when observed and latent variable(s) are discrete, the latent Markov models which are also known as latent transition models, are used as an alternative for the longitudinal psychology studies. These models were used in the applied research such as substance abuse (Couden, Larsen, Donahue, & Nylund-Gibson, 2015; Guo, Aveyard, Fielding, & Sutton, 2009; La Flair et al., 2013; Lanza & Bray, 2010), eating behaviors (Cain, Epler, Steinley, & Sher, 2010; Castellini et al., 2013) and so on. Although in recent years the latent Markov models have been commonly used in the applied field by means of the development of software such as Latent Gold (Vermunt & Magidson, 2013), Mplus (Muthen & Muthen, 2013), PROC LTA (Lanza & Collins, 2008) and R (Bartolucci & Pandolfi, 2016), which statistics should be used for model selection is still unclear. In this context, the first aim of the study is to present an example of model application by using an empirical dataset with a single variable. The second objective is to examine the effects of the strength of item response probabilities, the number of times the measurement being taken and sample size on model selection by means of the dataset generated by Monte Carlo simulation method.

Latent Markov models consist of two parts that are the measurement model and the structural model. In the measurement model, the relations of observed variables to the latent variables are determined. The struc-

tural model provides the transition probabilities among time-dependent latent states. In this regard, the basic parameters, which are item response probabilities, latent states probabilities, and transition probabilities, are estimated (Collins & Lanza, 2010; Collins & Flaherty, 2012; Lanza & Bray, 2010; Vermunt, Tran & Magidson, 2008). The parameter called latent class probability in the latent class analysis is usually known as latent states probability in the latent Markov models in order to emphasize the dynamic structure of class (Collins & Lanza, 2010). Considering a hypothetical study of driver behavior, it is possible to have two latent states which can be defined as careful driving and risky driving. In such a case, if the probability of being in a state of careful driving in the first time period is determined to be, such as .70, the probability of being in the second latent state can be calculated as $1.00 - .70 = .30$ since the sum of latent state probabilities for each time period is equal to 1. Item response probabilities are similar to factor loadings in factor analysis and can be interpreted as in the latent class analysis. The part where the transition probabilities are calculated constitutes the Markov part of model. In this regard, transition probabilities are the conditional probabilities that individuals or observations in the first latent state at time $t-1$ will move to the second latent state at time t .

The latent Markov model has two basic assumptions; local independence and first order Markov assumptions. The assumption of local independence is the assumption that the variables observed at time t are only related to the current time. The first order Markov assumption is that statuses at time t are influenced only by time $t-1$ (Bartolucci, Farcomeni, & Pennoni, 2013; Vermunt, Langeheine, & Bockenholt, 1999).

Model Selection

In the model selection process, firstly the number of latent state is determined separately for each time period. If the number of latent state is known in advance,

Table 1. Model Selection Values According to Sample Size, Item Response Probability and Number of Measurement Time

Item Response Probability		Strong						Weak					
Number of Measurement Time		2		3		4		2		3		4	
Sample Size	Information Criterion	True	False Positive	True	False Positive	True	False Positive	True	False Positive	True	False Positive	True	False Positive
200	BIC	100	0	100	0	100	0	100	0	100	0	100	0
	AIC	88	12	87	13	84	16	88	12	86	14	92	8
	AIC3	98	2	99	1	100	0	99	1	100	0	99	1
	CAIC	100	0	100	0	100	0	100	0	100	0	100	0
600	BIC	100	0	100	0	100	0	100	0	100	0	100	0
	AIC	87	13	93	7	87	13	89	11	85	15	89	11
	AIC3	100	0	100	0	99	1	100	0	99	1	100	0
	CAIC	100	0	100	0	100	0	100	0	100	0	100	0
2000	BIC	100	0	100	0	100	0	100	0	100	0	100	0
	AIC	90	10	81	19	74	26	84	16	94	6	78	22
	AIC3	100	0	99	1	98	2	100	0	99	1	99	1
	CAIC	100	0	100	0	100	0	100	0	100	0	100	0

this step can be skipped by adopting a confirmatory approach (Bartolucci, Farcomeni, & Pennoni, 2013). Then conditional probabilities and latent probabilities are determined by restricting the number of latent states. By examining these models, if the restrictions on the transition and conditional probabilities can be imposed, the nested models are tested by putting various restrictions. One of the methods that is used to select the best model among these nested models is to decide it based on L² statistics. Alternatively, the use of the AIC, BIC information criteria for model selection is generally accepted.

Empirical Example

The sample of this example is composed of 511 male drivers working in a private transportation company. With the permission of the company, the speeding records of the drivers in the first half of 2013, 2014, and 2015 were accessed. A five-time dataset was obtained by dividing the years into two semi-periods. When the frequency ratios of speed limit violations over years were examined, it was observed that they varied between 4% and 8%.

The dataset with a single variable taken in five-time periods was analyzed via the Latent GOLD 5.1 Syntax version. Parameters of model with two latent states were estimated. Model-1 was tested as a model in which there were two latent states and transition probabilities were

time-homogenous while Model-2 was analyzed as a model in which two latent states and but transition probabilities were time-heterogeneous. While the BIC and CAIC information criteria pointed to the time-homogeneous model, the AIC, AIC3 information criteria pointed to the time-heterogeneous model. The heterogeneous model parameters are interpreted based on the knowledge that the AIC3 information criterion gives more results that are consistent in latent class analysis (Güngör, Korkmaz, & Sazak, 2015). When item response probabilities are examined, the probability of those in the first latent state to violate speed limit was found to be .01 and the probability of those in the second latent status was found to be .66. Based on these probabilities, the first latent state was called as those complying with speed limit while the second latent state referred to those violating speed limit. In the first half of 2013, 98% of the drivers were in the latent state of those who comply with speed limit. In the second half of 2014, this probability declined to .96 and in the first half of 2015 dropped to .93. In other words, in the first half of 2015, the proportion of those in the latent state of violation of speed limit moved up to 7%.

Method

The datasets for 18 different conditions were generated by using Monte Carlo simulation section of La-

Table 2. Parameter Estimation Bias

Item Response Probability	Sample Size	Number of Measurement Time	Latent State Probability		Latent Transition Probability		Item Response Probability	
			Mean	SD	Mean	SD	Mean	SD
Strong	200	2,3,4	.52	3.56	.13	5.28	.21	4.37
Weak		2,3,4	.05	5.84	.79	15.95	.17	2.37
Strong	600	2,3,4	.35	2.09	.49	3.20	.20	2.38
Weak		2,3,4	.24	3.63	.97	1.73	.16	1.03
Strong	2000	2,3,4	.14	1.09	.27	1.56	.13	1.36
Weak		2,3,4	.16	1.95	.19	3.61	.17	.57

tent GOLD 5.0 statistical package program (Vermunt & Magidson, 2015). The study consisted of four independent variables which are the sample size of manipulated conditions (i.e., 200, 600, and 2000), the strength of item-response probabilities (i.e., .1, .9, .3, and .7), the number of measurement occasions (i.e., 2, 3, and 4), and the information criteria used in model selection (BIC, CAIC, AIC3, AIC). The information criterion used in the model selection was independent variable. In this context, 1800 datasets generated by 18 different conditions and 100 different replications for each condition were produced and analyzed. Furthermore, fixed conditions in the research were the number of the latent state (2), the latent state probabilities (i.e., .50), the latent transition probabilities (i.e., .80 and .20), the number of items (i.e., 5), the number of item category (i.e., 2). The latent transition probabilities were time-homogeneous.

As a result of the analyzes, the information criteria were calculated by using the values of L^2 obtained for each condition. Based on these information criteria, the models were selected and reported. Since the number of the latent states was fixed to 2, the proportion of deviation from this value was calculated as a percentage. In addition, the means and the standard deviations of the parameters of each condition were calculated and interpreted.

Results

Number of models with convergence error

Convergence errors were only detected for models with three and four states. In the case of models for three states, an average of 16% convergence error ($SD = 17.12$) was found, however for models with four status the rate was 33.5% ($SD = 10.76$).

The effect of strength of item response probabilities on model selection

As indicated in the Table 1, BIC and CAIC information criteria exhibited 100% true positive results, independent of the sample size, in the case of probabilities were both strong and weak. When the AIC3 information criterion was used, the correct decision ratios are 98% and above in all conditions where the item response probability was strong or weak.

The effect of number of measurement time on model selection

In model selection using BIC and CAIC, the true positive ratios were 100% regardless of the sample size in 3 different time conditions. When the number of measurement time was 2, in model selection by using AIC the highest true positive ratio was 90%. All other ratios for 3 and 4 time of measurement were indicated in Table 1.

The effect of sample size on model selection

In the case item response probabilities were strong and sample size increased from 200 to 2000, true positive ratios of the AIC information criterion were between .78 and .94 while the BIC and CAIC information criteria shows 100% accurate estimates, independent of all conditions.

Parameter estimation bias

The values of parameter estimation bias were estimated by the following equation (Muthén & Muthén, 2002):

$$\text{Bias} = (\text{PE} - \text{PP}) / \text{PP} * 100 \quad (1)$$

PE = Average value of parameter estimates as a result of replications

PP = Pre-defined population parameter value

According to this equation, parameter bias values of latent status, latent transition, and item response probabilities were calculated in reference to the item response probabilities and sample size for three different samples as 200, 600 and 2000 (see Table 2).

Discussion

In this study, which was designed as two parts, firstly, it was aimed to introduce the model through the empirical data with only one observed variable and five time occasions. In the first study, the models with two latent states which were time-homogeneous and time-heterogeneous were estimated by using the example of speed behavior in traffic. The BIC and CAIC information criteria pointed to the time-homogeneous model, while the AIC and AIC3 information criteria indicated the time-heterogeneous model as a better model. Since AIC3 is generally considered as an information criterion giving more consistent results (Güngör, Korkmaz, Sazak, 2015), in the study the time-heterogeneous model was preferred. In addition, time-heterogeneous model allows seeing the differences among the transition probabilities over the years. The transition from the group who did not violate the speed limit in the second half of 2014 and the first half of 2015 to the group who violated the speed limit had been found to have higher probabilities compared to previous years. Furthermore, it was observed that group who violated the speed limit continued to display the same behavior in the following years.

In the second study, a simulation study on model selection was presented. The findings of the simulation study indicated that when item response probabilities were strong and the sample size was large, the parameter estimates gave results that were more accurate. Therefore, we recommend the researchers planning to use the latent Markov model to be careful in choosing indicator variables and to make comments on that item when item response probabilities were .70 and less.

As for the limitations of the study, the simulation study was limited to the variables that addressed in this research. Furthermore, the simulations involving multi-group latent Markov models, simulations of time-heterogeneous models and longitudinal equivalence studies are still some of the issues to be investigated. Another limitation of the study was to introduce a model through an example of a single package program, Latent GOLD program. Researchers can conduct the same analysis with the latent Markov extension of R program, which has become an extensively used as a free software program.