Time Varying Covariates in Markov Latent Class Analysis: Some Problems and Solutions

Marcus Berozfsky[1], Paul P. Biemer[2], and William Kalsbeek[3]
[1]RTI International, 3040 Cornwallis Rd., Research Triangle Park, NC 27709
[2]RTI International, 3040 Cornwallis Rd., Research Triangle Park, NC 27709 and
University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599
[3]University of North Carolina at Chapel Hill, 730 MLK Jr. Blvd., Chapel Hill, NC 27599

Abstract

Markov latent class analysis (MLCA) is a modeling technique for panel or longitudinal data that can be used to estimate the classification error rates for categorical outcomes with categorical predictors (i.e., false positive and false negative rates for dichotomous items) when gold standard measurements are not available. Because panel surveys track respondents over time, explanatory variables (called grouping variables) can be either time varying or time invariant (static). Time varying grouping variables can be useful in explaining differences in the latent construct. However, they generate a large number of model parameters that can cause problems with data sparseness, make model diagnostics invalid, and model convergence less reliable. This paper discusses alternative coding schemes for time varying grouping variables and proposes a set of procedures for determining the best coding scheme for a particular set of data. This process is then illustrated using data from the National Crime Victimization Survey (NCVS). We found that for the NCVS, when parsimony is taken into account, a coding scheme that uses fewer model parameters has better fit than the more traditional coding scheme and another alternative and does not negatively affect the estimates of the classification error.

Key Words: Markov Latent Class Analysis, National Crime Victimization Survey, time varying covariates, measurement error, classification error, screener questions

1. Introduction

Markov latent class analysis (MLCA) is a modeling technique used to assess the classification error in survey items from a panel or longitudinal survey (Wiggins, 1973; Poulson, 1982; Van de Pol & de Leeuw, 1986; Van de Pol & Langeheine, 1990). Often in surveys, there is no gold standard (or error-free) data source available to evaluate the error in survey responses. Therefore, gold standard techniques cannot be used. Furthermore, even when so-called gold standard sources, like administrative records or tests using hair or blood samples, do exist, studies have found that these sources are flawed as well and, therefore, it is not safe to assume that they are error free (Visher & McFadden, 1991). MLCA does not require that a gold standard exist in order to estimate both the true prevalence of the latent variable and the corresponding measurement error. However, since it is a modeling technique it is constrained by the number of parameters that can be included before the model fit is weakened, due to data sparseness (Biemer & Berzofsky, in press), to the point where the estimates are unreliable. This paper looks at ways in which the number of parameters used for time varying grouping variables in the structural component of a MLCA model can be minimized and how to test that these approaches are appropriate for the particular data being modeled.

The MLCA model consists of two components: a *structural component* and a *measurement component*. The structural component describes interrelationships among the latent variables. In latent class analysis, a single latent variable is used to represent the true value of the dependent variable. Since MLCA deals with panel or longitudinal data, there is a latent variable to represent the value of the dependent variable at each time point. For our work, both the dependent variable and the latent construct have the same number of known, fixed number of levels. The dependent variable is defined the same way at each time point except the reference period shifts with the change in time. For example, the Current Population Survey is a monthly panel survey that tracks employment status over time and divides the population into one of three categories: employed, unemployed, or not in the labor force. Thus, if the true employment status is being modeled, there is a separate latent variable representing a person's true employment status during the previous month.

MLCA requires that at least three time points are included in the model. In a three time point model, which will be the focus of this paper, the latent variables are represented by X_1, X_2, and X_3 where X_1 is the latent value for the first time period, X_2 is the latent value for the second time period, and X_3 is the latent value for the third time period. Because the latent variables occur in a chronological fashion, each latent variable is dependent on the previous values (i.e., X_2 given X_1, hereby notated as $X_2|X_1$, and $X_3|X_2X_1$). However, when there are only three time points, in order for the models to be identifiable, a *first-order Markov assumption* is made which presumes that the response at a given time point is only dependent on the most recent previous time point (i.e., $\pi_{x_3|x_2x_1}^{X_3|X_2X_1} = \pi_{x_3|x_2}^{X_3|X_2}$).

The measurement component estimates how well each indicator (i.e., the survey items used to measure the latent construct) does at measuring the latent variable. In other words, the measurement component estimates the probabilities of the indicators conditioned on the latent variable (for dichotomous indicators and latent variables these probabilities represent the false positive and false negative rates for an indicator). Each time point used in the MLCA must have a corresponding indicator. In a model with three time points these indicators are represented by A_1, A_2, and A_3, where A_1 corresponds to the first time point, A_2 corresponds to the second time point, and A_3 corresponds to the third time point. In order for valid model estimates, the *independent classification errors (ICE) assumption* must be made. ICE assumes that the classification errors across waves are independent. I.e., $\pi_{a_1a_2a_3|x_1x_2x_3}^{A_1A_2A_3|X_1X_2X_3} = \pi_{a_1|x_1}^{A_1|X_1}\pi_{a_2|x_2}^{A_2|X_2}\pi_{a_3|x_3}^{A_3|X_3}$. Furthermore, in order for the model to be identifiable, *time-homogeneous classification errors* must be assumed which states that the classification errors for indicator A_t are the same in all waves t=1, 2, 3. In other words, $\pi_{a_t|x_t}^{A_t|X_t} = \pi_{a|x}^{A|X}$ for $a = a_t$, $x = x_t$, t=1, 2, 3.

Given these assumptions, the likelihood kernel for the MLCA model is

$$\pi_{abc}^{ABC} = \sum_{x_1}\sum_{x_2}\sum_{x_3}\left(\pi_{x_1}^{X_1}\pi_{x_2|x_1}^{X_2|X_1}\pi_{x_3|x_2}^{X_3|X_2}\right)\left(\pi_{a_1|x_1}^{A_1|X_1}\pi_{a_2|x_2}^{A_2|X_2}\pi_{a_3|x_3}^{A_3|X_3}\right) \quad (1)$$

where $\pi_{x_1}^{X_1}\pi_{x_2|x_1}^{X_2|X_1}\pi_{x_3|x_2}^{X_3|X_2}$ is the structural component and $\pi_{a_1|x_1}^{A_1|X_1}\pi_{a_2|x_2}^{A_2|X_2}\pi_{a_3|x_3}^{A_3|X_3}$ is the measurement component. Figure 1 presents the path diagram for the basic MLCA model.

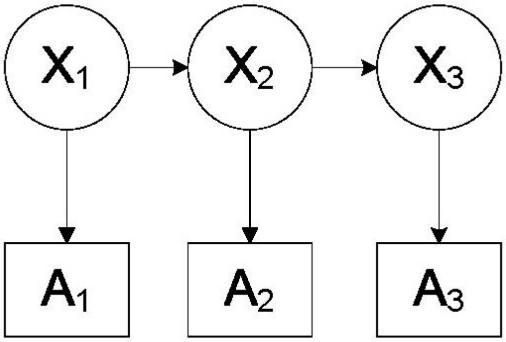

Figure 1. Path diagram for MLCA model with three time points and no grouping variables with measurement component that assume time-homogeneous classification errors (i.e., $A_1|X_1=A_2|X_2=A_3|X_3$)

1.1 Time varying covariates and time invariant covariates

In addition to the Markov, ICE, and time-homogeneous classification errors assumptions, in order to have valid estimates, MCLA models require the assumption of *homogeneous error probabilities.* This assumption requires that all individuals in the same latent class have the same probability of being misclassified. This assumption is unlikely to be met without the addition of grouping variables.

Grouping variables are manifest variables from the survey (e.g., respondent's age or race/ethnicity) or paradata about the interview characteristics (e.g., mode of interview, whether the person was alone during the interview). Because of the longitudinal nature of panel surveys, grouping variables can be either *time invariant (or static)* or *time varying*. Time invariant grouping variables do not change over time or change as simple linear function of time for all respondents such as race/ethnicity or age. Time invariant grouping variables are denoted by a single letter starting with *G*. Time varying grouping variables change over time in a potentially non-linear manner. For example, how often a person goes out in public or the mode of the data collection may change over time in different patterns for each respondent. Time varying grouping variables are denoted by G_t t=1,…,T, where t represents the wave number to which that grouping variable corresponds.

1.2 Understanding the problem

Because the changes in a person's actions or behavior that occur over time may predict the individual's current state, it is important to fully capture variations in the grouping variable in the MLCA model. Moreover, similar to the latent variables, the current value of a time varying grouping variable may be a function of its values in previous time periods. Therefore, in order to appropriately model this variation it is often necessary to condition each value on the previous values (i.e., $G_3|G_2G_1$).

Fully capturing the information provided by a time varying grouping variable dramatically increases the number of parameters used in the model. For example, a time invariant grouping variable with two levels (assuming time homogeneous error probabilities) adds six parameters to a model (i.e., $\{G\ GX_1\ GX_2X_1\ GX_3X_2\ A_1X_1\ A_2X_2\ A_3X_3\}$ has 14 parameters while $\{X_1\ X_2X_1\ X_3X_2\ A_1X_1\ A_2X_2\ A_3X_3\}$ has 8 parameters). However, a time varying grouping variable with two levels (assuming time homogeneous error probabilities) adds 28 parameters (i.e., $\{G_1\ G_2G_1\ G_3G_2G_1\ G_1X_1\ G_2G_1X_2X_1\ G_3G_2G_1X_1X_2$

A_1X_1 A_2X_2 A_3X_3} has 36 parameters). This reduces the number of degrees of freedom available to fit other parameters and can cause several problems with fitting the model, including data sparseness, weak identifiability and over-fitting (Biemer & Berzofsky, in press; Berzofsky, 2009).

Furthermore, as survey methodologists, we are mainly interested in understanding the classification error in survey questions. Understanding the classification error helps in the design of future iterations of the survey in two key ways. First, MLCA provides information on which indicators have higher error rates than others (see, for example, Biemer & Bushery, 2001; Biemer, 2004). Through this understanding, survey methodologists can reword questions to reduce the measurement error in these items. Second, through the use of grouping variables, MLCA provides information on which sub-populations have higher classification error rates. Different groups may have different classification error rates for many reasons including lower comprehension of the questions or higher propensity to falsify their responses. By understanding which sub-populations have higher classification error rates, interviewers can be better trained to try and minimize the classification error in these groups. Thus, our main interest is in the measurement component of the model. Because we want to specify the correct model for the measurement component and because that often requires an extensive set of parameters, reserving degrees of freedom for the measurement component is desirable. Using time invariant (rather than time varying) grouping variables helps in this regard.

Therefore, in terms of the structural component of the model, our ultimate goal is to have a good fitting model that uses as few parameters as possible. This will allow an analyst the ability to use more parameters in the measurement component before data sparseness issues arise. Thus, it is of interest to explore alternative coding schemes that allow one to capture all of the necessary information contained in a time varying grouping variable while using fewer parameters. In this paper, we propose a theoretical approach to develop alternative coding schemes for time varying grouping variables to be used in the structural component and a process by which an analyst can determine which is most appropriate for their data. We implement our approach using data from the National Crime Victimization Survey (NCVS).

2. Alternative Coding Schemes

2.1 Types of coding schemes

The pattern of responses for time varying grouping variables can be thought of in terms of *event characteristics* or *behavioral characteristics*. Event characteristics take into account the actual outcome at each time point and how it relates to each previous time point. The traditional approach for coding time varying covariates described earlier is a type of event-characteristics coding scheme. Alternatively, behavioral characteristics summarize all time periods simultaneously and groups respondents that had similar *patterns* of behavior across each of the time points. In other words, behavioral characteristic coding schemes assess whether modeling a person's behavior is more predictive of the dependent variable compared to the specific event that occurred at each time point. The next section will discuss how behavioral characteristic coding schemes are represented in a MLCA model.

2.2 Theoretical alternative coding schemes

Using these two general coding schemes, analysts need to develop MLCA models that make theoretical sense to their particular dataset. In this section, we present four potential models and the pros and cons to each. Each model uses a different coding scheme for representing a time varying grouping variable in the structural component and assumes a naïve measurement component. While these models are unrealistic final MCLA models, the models are constructed to easily allow comparison with each other.

2.2.1 Second-order Markov time-varying

The second-order Markov time varying coding scheme is an event characteristic coding scheme that fully takes into account all previous time points based on how events transpired over time. Under this coding scheme the likelihood kernel for the MLCA model is written as

$$\pi^{G_1G_2G_3A_1A_2A_3}_{g_1g_2g_3a_1a_2a_3} = \sum_{x_1}\sum_{x_2}\sum_{x_3} \pi^{G_1}_{g_1} \pi^{G_2|G_1}_{g_2|g_1} \pi^{G_3|G_2G_1}_{g_3|g_2g_1} \pi^{X_1|G_1}_{x_1|g_1} \pi^{X_2|G_2G_1X_1}_{x_2|g_2g_1x_1} \pi^{X_3|G_3G_2G_1X_2}_{x_3|g_3g_2g_1x_2} \pi^{A_1|X_1}_{a_1|x_1} \pi^{A_2|X_2}_{a_2|x_2} \pi^{A_3|X_3}_{a_3|x_3} \quad (2)$$

Figure 2 presents the path diagram for this model. This model utilizes all information available as it occurred and is appropriate if the actual trait at a particular time point is more related to the latent construct rather than the respondent's more general behavior patterns. However, this scheme utilizes a large number of model parameters which reduces the number of other grouping variables that can be used in the model before data sparseness issues arise.

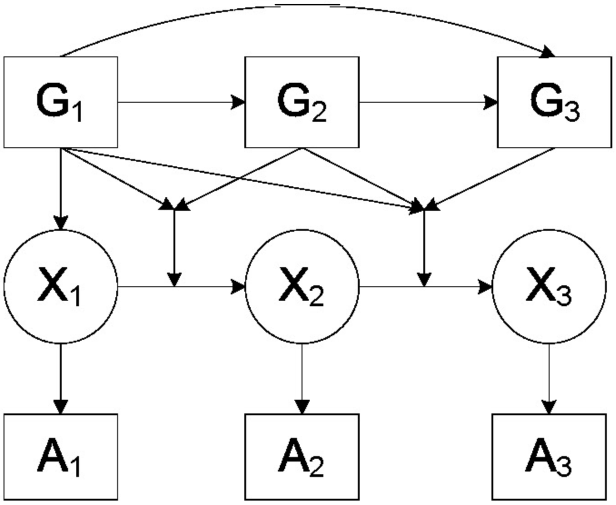

Figure 2. Path diagram for MLCA model with second-order Markov time varying grouping variable in the structural component and a simple measurement component that assumes time-homogeneous classification errors (i.e., $A_1|X_1=A_2|X_2=A_3|X_3$)

2.2.2 First-order Markov time varying

The first-order Markov time varying coding scheme is an event characteristics coding scheme that reduces the number of parameters used by assuming $\pi^{G_3|G_1G_2}_{g_3|g_1g_2} = \pi^{G_3|G_2}_{g_3|g_2}$ (i.e., the probability of having the trait only depends on the most recent previous period). Under this coding scheme the likelihood kernel for the MCLA model is written as

$$\pi^{G_1G_2G_3A_1A_2A_3}_{g_1g_2g_3a_1a_2a_3} = \sum_{x_1}\sum_{x_2}\sum_{x_3} \pi^{G_1}_{g_1} \pi^{G_2|G_1}_{g_2|g_1} \pi^{G_3|G_2}_{g_3|g_2} \pi^{X_1|G_1}_{x_1|g_1} \pi^{X_2|G_2G_1X_1}_{x_2|g_2g_1x_1} \pi^{X_3|G_3G_2X_2}_{x_3|g_3g_2x_2} \pi^{A_1|X_1}_{a_1|x_1} \pi^{A_2|X_2}_{a_2|x_2} \pi^{A_3|X_3}_{a_3|x_3} \quad (3)$$

Figure 3 illustrates the path diagram for this model. If this first-order Markov assumption is plausible then this coding scheme reduces the number of parameters used by 10 for a 2-level time varying grouping variable and 48 for a 3-level time varying grouping variable compared to the second-order Markov time varying coding scheme. However, if this assumption does not hold, the loss of information could potentially decrease the fit of the model.

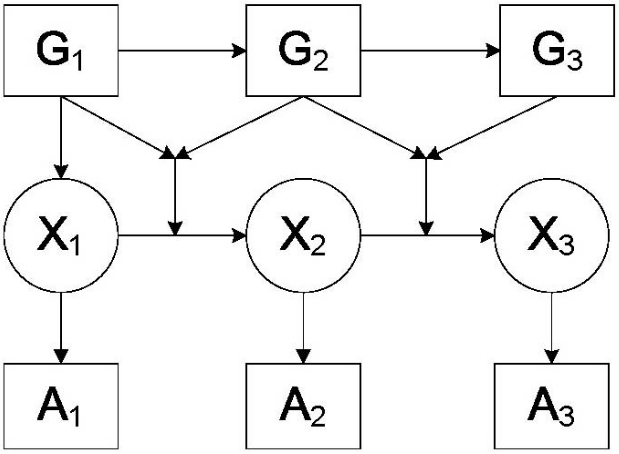

Figure 3. Path diagram for MLCA model with first-order Markov time varying grouping variable in the structural component and a simple measurement component that assumes time-homogeneous classification errors (i.e., $A_1|X_1=A_2|X_2=A_3|X_3$)

2.2.3 Time-invariant summary

The time-invariant summary coding scheme is a behavioral characteristic coding scheme that uses information from each time point to model each possible response pattern. For example, under a behavioral characteristic coding scheme, a two-level time varying covariate where a person had a particular trait or did not (e.g., used public transportation or did not) crossed with three time points (i.e., $G_1G_2G_3$) produces eight distinct behavioral patterns; namely, 111, 112, 121, and so on. Under this coding scheme the likelihood kernel for the MLCA model is written as

$$\pi^{G_1G_2G_3A_1A_2A_3}_{g_1g_2g_3a_1a_2a_3} = \sum_{x_1}\sum_{x_2}\sum_{x_3} \pi^{G_1G_2G_3}_{g_2g_2g_3} \pi^{X_1|G_1G_2G_3}_{x_1|g_2g_2g_3} \pi^{X_2|G_1G_2G_3X_1}_{x_2|g_2g_2g_3x_1} \pi^{X_3|G_1G_2G_3X_2}_{x_3|g_2g_2g_3x_2} \pi^{A_1|X_1}_{a_1|x_1} \pi^{A_2|X_2}_{a_2|x_2} \pi^{A_3|X_3}_{a_3|x_3} \quad (4)$$

Figure 4 presents the path diagram for this model. With this coding scheme is best when the behavioral pattern of a respondent is correlated to the latent construct more than the actual event that transpired at any given point in time. However, because information from all three time points is used at each time point, this scheme uses more parameters than the second-order Markov time varying coding scheme.

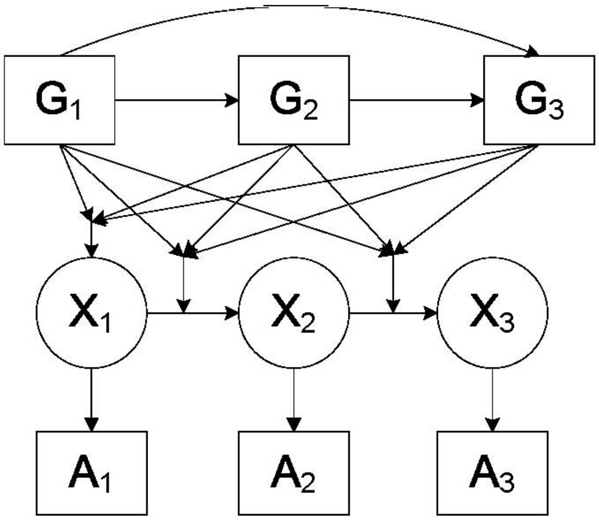

Figure 4. Path diagram for MLCA model with a time-invariant summary grouping variable in the structural component and a simple measurement component that assumes time-homogeneous classification errors (i.e., $A_1|X_1=A_2|X_2=A_3|X_3$)

2.2.4 Reduced time-invariant summary

The reduced time-invariant summary coding scheme is a behavioral characteristic coding scheme, but rather than having a level for each possible response pattern it collapses some of the patterns whose relationship with the latent construct differs by a small or negligible amount. Because this is a time invariant variable, this collapsed variable is represented by *G*. For example, respondents whose trait changes over time, regardless of the change or order of the change, may have a similar propensity for the latent construct to occur. Therefore, the reduced time-invariant summary variable would have a level for respondents that always had the same trait at each time point and a single additional level for respondents whose trait changed over time. Under this example, for a two-level time varying grouping variable, *G* would be defined as:

$$G=\begin{cases}1 & \text{if } G_1=1,\ \mathrm{G}_2=1,\ \mathrm{G}_3=1\\ 2 & \text{if } G_1=2,\ \mathrm{G}_2=2,\ \mathrm{G}_3=2\\ 3 & \text{if mixed response pattern}\end{cases}$$

Alternative collapsing schemes may split those with a mixture response pattern into more than one category. Under this coding scheme the likelihood kernel for the MLCA model is written as

$$\pi_{ga_1a_2a_3}^{GA_1A_2A_3}=\sum_{x_1}\sum_{x_2}\sum_{x_3}\pi_g^G\pi_{x_1|g}^{X_1|G}\pi_{x_2|gx_1}^{X_2|GX_1}\pi_{x_3|gx_2}^{X_3|GX_2}\pi_{a_1|x_1}^{A_1|X_1}\pi_{a_2|x_2}^{A_2|X_2}\pi_{a_3|x_3}^{A_3|X_3} \tag{5}$$

Figure 5 presents the path diagram for this coding scheme. This coding scheme can greatly reduce the number of parameters used in a model. Depending on how the collapsing is done, this scheme most likely adds fewer parameters to the model than any of the other schemes discussed. However, if the actual event that occurred at a particular time period is important or the response patterns are collapsed too much, the loss of information may make the model fit worse.

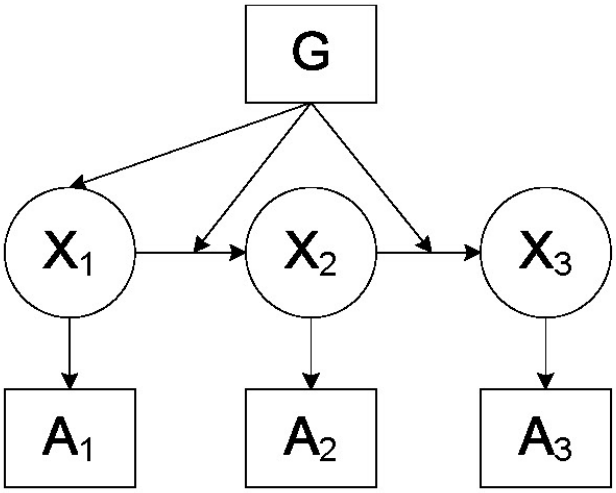

Figure 5. Path diagram for MLCA model with a reduced time-invariant summary grouping variable in the structural component and a simple measurement component that assumes time-homogeneous classification errors (i.e., A1|X1=A2|X2=A3|X3)

3. Process for determining best fitting coding scheme

We developed a five step process to compare each alternative coding scheme being considered to the second-order Markov time varying coding scheme and determine which provides the best fit for a particular set of data. This analysis should be done prior to fitting the main MLCA model so that one knows which coding scheme to use for the main model. The steps in this process are:

1. Determine which time varying grouping variables to analyze
2. Select coding schemes that make the most sense for the particular dataset being analyzed
3. Using the second-order Markov time varying coding scheme, obtain the expected frequencies table for the data
4. Using the expected frequencies data table, run a model for each coding scheme
5. Compare the BIC from each model to determine best fitting model and verify that the classification error rates do not differ by coding scheme

Our proposed process needs to be conducted on each time varying grouping variable being considered for the structural component. When determining which variables to include, it is important to only consider variables that are theoretically related to the latent construct. For example, paradata are not theoretically related to the latent construct, and, thus should not influence whether the latent construct occurred. If, by chance, a paradata variable is found to be significant it is likely a spurious finding. For example, the theory may be wrong or incomplete. One can still examine variables which do not fit the theory, but it can lead to puzzling findings if the theory becomes too far-fetched.

When determining which coding schemes to compare, there are several things to consider. For example, it is probably useful to consider at least one alternative event characteristic coding scheme (we assume that the second-order Markov time varying coding scheme will be included as the default coding scheme) and one behavioral coding scheme. Because the time-invariant summary coding scheme that accounts for all response patterns uses more parameters than the second-order Markov time varying coding scheme, an analyst needs to have a strong feeling that each behavioral pattern has a different relationship with the latent construct. Therefore, when considering which reduced time-invariant summary coding schemes to include, one needs to consider both (1) how many parameters it adds to the model and (2) whether additional response patterns can be collapsed. When deciding how to collapse, it is important to look at the distribution of the patterns. Smaller groups (less than 5% of respondents) should be collapsed into adjacent categories to avoid sparseness problems. However, even if the percentage of respondents indicating a change in behavior over time is large (20% or greater), it may still be advantageous to collapse these respondents into a small number of categories in order to reduce the number of parameters used in the structural component.

In terms of fit, the second-order Markov time-varying covariate model is best; however, as previously noted, it may not be best when model parsimony is considered. Therefore, we used the BIC to compare models (Schwarz, 1978). The BIC adds a penalty factor to the maximum log-likelihood based on the number of parameters used in the model. Thus, the model with the smallest BIC is the best fitting model when parsimony is considered.

In order to compare each coding scheme, a data table containing each possible grouping variable needs to be constructed. In other words, if a time-invariant summary variable is being considered, the $GG_1G_2G_3A_1A_2A_3$ table would need to be constructed. For the time-invariant summary variables this may create *structural zeros* (Biemer & Berzofsky, 2010) because of illogical combinations. A structural zero is a cell with an expected frequency of zero because it cannot logically occur. For example, for a two-level time varying grouping variable, suppose G is defined as 1=always has trait (i.e., 111 pattern), 2=never has trait (i.e., 222 pattern), and 3=trait varies over time (i.e., a mixture pattern). Then, the expected cell frequency table will create cells with G indicating a mixture pattern, but G_1, G_2, and G_3 indicating that the respondent has always had the trait (i.e., G=3, G_1=1, G_2=1, and G_3=1). By definition this combination cannot occur, creating a structural zero. Failure to account for structural zeros in models using the time-invariant summary grouping variables will produce incorrect test statistics because the number of parameters calculated will be wrong. Most software packages allow users to specify structural zeros. For example, LEM (Vermunt, 1997) has a weight statement that can be used to assign no weight to the cell combinations that are not logically possible.

Moreover, because the measurement component is of ultimate interest it is important to ensure that any coding scheme used in the structural component does not affect the estimates made by the measurement component of the model. While the two components of the MLCA model are discussed and interpreted separately, there is only one model being fit. Therefore, while not expected, changes to one component can change the estimates from the other component. Thus, when determining the best coding scheme for the structural component, one needs to see if the choice of coding scheme alters the estimates from the measurement component.

4. Applying procedures to the NCVS

4.1 Overview of the NCVS

4.1.1 Sample design

The NCVS is a household survey that measures crime victimization rates in the United States. The survey uses a multi-stage probability design to make inference to all persons 12 years old or older in the U.S. (U.S. Department of Justice, 2007). A sample of 50,000 households is selected every 6 months and all persons 12 years old or older in a sampled household are interviewed. The survey utilizes a rotating panel design in which each household is surveyed every 6 months and remains in the field for three and a half years. The reference period for each interview is the past 6 months. The NCVS achieves a 90% response rate.

4.1.2 Conducting an MLCA with the NCVS

For the NCVS the latent construct of interest is whether a particular crime occurred or not during the past 6 months. These latent constructs are measured through a series of *screener questions*. Screener questions are a short set of questions that help a respondent remember whether a particular event occurred during the reference period (see, for example, Biemer, 2000). The NCVS includes 10 screener questions to probe about crime victimization. However, because these screener questions ask about an overlapping set of crimes, we collapsed them into three latent constructs: victim of a less serious individual crime, victim of a more serious individual crime, and victim of a household crime. The construct of less series crimes against an individual include crimes such as theft, simple assault, and robbery. The construct of serious crimes against an individual include aggravated assault and rape or sexual assault. The construct of crimes against a household include vandalism, motor vehicle theft, and household burglary.

In order to conduct an MLCA, data from all waves for a rotation group are needed. The most recent public use file containing all waves for a set of respondents is the National Crime Victimization Longitudinal File, 1995 – 1999 (U.S. Dept. of Justice, 2007). This file has data from three rotation groups which were released in the third quarter of 1995, the first quarter of 1996, and the third quarter of 1996. These rotation groups contained 26,345 households and 66,706 unique respondents.

In this analysis, we used the first three waves after the bounding interview. Furthermore, we only included respondents that completed the screener questions in all three waves and provided an answer to the time varying grouping variable being tested in each wave. Based on these conditions, the number of respondents used for analysis ranged between 27,845 and 28,000 respondents for the individual level screeners and the number of respondents used for analysis ranged between 16,150 and 17,025 respondents for the household level screeners. Furthermore, in this analysis, we ignored the survey design and assume the data came from a simple random sample.

4.1.3 Time varying grouping variables in the NCVS

The NCVS survey consists of over 20 different pieces of information that can be used as a grouping variable. These variables consist of respondent characteristics, information from the sampling frame, and interview paradata. Of these variables, six items are time varying: three respondent characteristics and three paradata items.

The survey items include how often the respondent went out in the evening over the past 6 months, how often the respondent went shopping over the past 6 months, and how often the respondent used public transportation over the past 6 months. For each of these items the respondent could answer “don’t know” or one of five responses ranging from “never” to “almost every day (or more frequently)”. For purposes of this analysis, the respondent characteristic items were collapsed based on the distribution of the data. The frequency of going out in the evening and frequency of going shopping were collapsed to three levels: 1=almost every day, 2=at least once a week, 3=once a month or less. The use of public transportation was collapsed into two categories: 1=at least once in the past 6 months, 2=never”. For all three of these items, responses of “don’t know” were treated as missing.

The paradata items include three two-level grouping variables. These variables are the mode of the interview (face-to-face or telephone), whether the person self respondent or if a proxy respondent was used, and whether the respondent was alone while taking the interview.

4.2 Comparing time varying coding schemes for the structural component

4.2.1 Determining which variables to analyze

When modeling the structural component, it is important to identify grouping variables that best explain differences in the latent construct. In general, paradata do not make theoretical sense in the context of the structural component. Therefore, only the respondent characteristic grouping variables (i.e., frequency of going out in the evening, shopping, and use of public transportation) were analyzed.

4.2.2 Determine which coding schemes to compare

For our analysis, we considered two coding schemes to compare to the second-order Markov time varying coding scheme: the first-order Markov time varying coding scheme and the reduced time-invariant summary coding scheme. The first-order Markov scheme was considered because the interval between interviews in the NCVS is 6 months. With this length of spacing between interviews, the Markov assumption may also be plausible for these grouping variables because knowing a person’s actions over the previous year may provide a good representation of how that person acts in general. For the reduced time-invariant summary coding scheme, we considered a scheme that collapsed all response patterns where there was at least one change over the three waves. In coming to this decision, we first looked at the percentage of respondents that indicated a changing behavioral pattern. Table 1 presents the distribution of each time-invariant summary variable. While each of these had large enough change categories to consider collapsing into more than one category, we decided to collapse them into a single category because we hypothesized that change of any kind was the main piece of information influencing crime victimization and other reduced coding schemes would still have required too many parameters than we were willing to use in the structural component.

4.3 Results of comparison

Using LEM, comparisons were conducted for each latent construct (i.e., less serious individual crimes, more serious individual crimes, and household crimes) and time varying grouping variable. Table 2 presents the fit statistics for these models. For each type of crime victimization, the second-order Markov time varying model has the smallest dissimilarity index, but the reduced time-invariant summary model had the smallest BIC. In fact, the reduced time-invariant summary model saves up to 78 degrees of freedom for a 3-level time varying covariate and up to 16 degrees of freedom for a 2-

level time varying covariate. The fit statistics indicate that the second-order Markov time varying model fits the data best when parsimony is not taken into account, however, when parsimony is considered the reduced time-invariant summary has the best fit.

Table 1. Distribution of Time-Invariant Summary Grouping Variables

Time-Invariant Summary Variable	Distribution (%)
Going out in the evening	
Always responded "Every night"	5.4
Always responded "Once a week"	23.0
Always responded "Less than once a week"	15.1
Responded in a mixture pattern	56.5
Going shopping	
Always responded "Every day"	7.3
Always responded "Once a week"	39.0
Always responded "Less than once a week"	3.8
Responded in a mixture pattern	49.9
Uses public transportation	
Always responded "Used public transportation"	9.6
Always responded "Never used public transportation"	65.4
Responded in a mixture pattern	25.0

Table 2. Comparison of Coding Schemes for Time Varying Grouping Variables in the Structural Component by Type of Crime Victimization and Coding Scheme

Victimization Type/Coding	Going Out in the Evening				Going Shopping				Using Public Transportation			
Scheme	p†	df	d††	BIC‡	p†	df	d††	BIC‡	p†	df	d††	BIC‡
Less serious individual crime												
Second-order Markov time varying	104	112	0.00	1.954	104	112	0.00	1.769	36	28	0.00	1.135
First-order Markov time varying	56	160	0.12	1.974	56	160	0.10	1.783	26	38	0.07	1.149
Reduced time-invariant summary	26	190	0.02	1.947	26	190	0.01	1.761	20	44	0.01	1.134
More serious individual crime												
Second-order Markov time varying	104	112	0.00	1.162	104	112	0.00	1.430	36	28	0.00	0.799
First-order Markov time varying	56	160	0.12	1.164	56	160	0.10	1.445	26	38	0.06	0.813
Reduced time-invariant summary	26	190	0.01	1.161	26	190	0.01	1.423	20	44	0.00	0.798
Household crime												
Second-order Markov time varying	104	112	0.00	1.196	104	112	0.00	1.110	36	28	0.00	0.687
First-order Markov time varying	56	160	0.12	1.205	56	160	0.10	1.117	26	38	0.07	0.694
Reduced time-invariant summary	26	190	0.02	1.190	26	190	0.01	1.103	20	44	0.00	0.685

†p=number of parameters
††d=dissimilarity index
‡BIC is in units 100,000

Table 3 presents the classification error rates for each time varying variable by type of crime victimization and coding scheme. For the NCVS data, the classification error rates did not vary much by coding scheme for a given victimization type. These estimates are based on a simple measurement component and, therefore, may change with the inclusion of grouping variables. However, these estimates indicate that, as expected, the choice of coding scheme for time varying grouping variables in the structural component does not impact the measurement component estimates.

Table 3. Estimated Classification Error Rates in the NCVS for Each Time Varying Grouping Variable by Type of Crime Victimization and Coding Scheme

Victimization Type/Coding Scheme	Going Out in the Evening		Going Shopping		Using Public Transportation	
	$\pi_{1\|2}^{A_1\|X_1}$	$\pi_{2\|1}^{A_1\|X_1}$	$\pi_{1\|2}^{A_1\|X_1}$	$\pi_{2\|1}^{A_1\|X_1}$	$\pi_{1\|2}^{A_1\|X_1}$	$\pi_{2\|1}^{A_1\|X_1}$
Less serious individual crime						
Second-order Markov time varying	0.0213	0.6833	0.0298	0.6204	0.0305	0.6432
First-order Markov time varying	0.0213	0.6832	0.0298	0.6204	0.0305	0.6432
Reduced time-invariant summary	0.0213	0.6836	0.0298	0.6206	0.0305	0.6430
More serious individual crime						
Second-order Markov time varying	0.0000	0.8780	0.0000	0.8863	0.0000	0.8954
First-order Markov time varying	0.0000	0.8922	0.0000	0.8708	0.0000	0.8878
Reduced time-invariant summary	0.0000	0.8760	0.0000	0.8883	0.0000	0.8951
Household crime						
Second-order Markov time varying	0.0168	0.6743	0.0052	0.6735	0.0285	0.6327
First-order Markov time varying	0.0167	0.6750	0.0065	0.6880	0.0262	0.6364
Reduced time-invariant summary	0.0170	0.6756	0.0055	0.6755	0.0286	0.6322

Table 4 presents the estimates for the structural component variables (i.e., X_1, X_2, and X_3) by type of crime victimization and coding scheme. These estimates are based on a model with a single grouping variable in the structural component and, therefore, may change with the inclusion of additional grouping variables. Under each crime victimization type the classification error rates are the same or nearly the same under each coding scheme.

An interesting observation of the preliminary model results is that the false negative error rates (i.e., $\pi_{2|1}^{A_1|X_1}$ in Table 3) seem implausibly high at first glance. However, this is consistent with the pattern of the observed prevalence rates (not shown in the tables), which suggests that crime victimization is affected by a respondent's time in sample; in particular, it suggests that crime victimization prevalence decreases after each panel wave. A more plausible explanation is that crime victimization is independent of a respondent's time in sample and, therefore, the rate of underreported crimes is increasing with each interview. This phenomenon is plausible because, as shown in other research (see for example, Kalton, et. al., 1989), respondents can be "conditioned" by prior interviews to misreport in ways that will shorten the interview. Denying a victimization will shorten the interview by avoiding further questioning about the victimization, thus reducing respondent burden.

Note, however, that the MLCA model with only three time points must restrict the error probabilities to be equal across the three time points. In other words, although the false negative error probabilities are increasing over time, the MLCA model must assume they are constant over time. As a consequence, only the average of the time 1, time 2 and time 3 false negative error probabilities can be estimated (cf, for example, Biemer, 2010). This creates a higher false negative rate for the first time point and lower false negative rate for the later two time points. This averaging also inflates the estimated true rate at the first time point and deflates it at the later two time points. For example, note from Table 4 that the estimated true crime victimization rate decreases over time. Again, victimization theory would suggest no relationship between a respondent's propensity to be victimized and their time in the panel.

In summary, we believe the high false negative error rates are plausible, but the decreasing true victimization rates are likely incorrect because of a violation in the time homogeneous classification errors assumption. Because NCVS has six waves, an analysis using all waves can be conducted which will allow for the validity of this assumption to be tested.

Table 4. Estimated Probabilities for Being a Victim of a Crime in the NCVS at Each Time Point for Each Time Varying Grouping Variable by Type of Crime Victimization and Coding Scheme

Victimization Type/Coding Scheme	Going Out in the Evening			Going Shopping			Using Public Transportation		
	$\pi_1^{X_1}$	$\pi_1^{X_2}$	$\pi_1^{X_3}$	$\pi_1^{X_1}$	$\pi_1^{X_2}$	$\pi_1^{X_3}$	$\pi_1^{X_1}$	$\pi_1^{X_2}$	$\pi_1^{X_3}$
Less serious individual crime									
Second-order Markov time varying	0.263	0.186	0.153	0.179	0.120	0.096	0.191	0.127	0.100
First-order Markov time varying	0.263	0.186	0.153	0.179	0.120	0.096	0.191	0.127	0.100
Reduced time summary invariant	0.265	0.187	0.154	0.179	0.120	0.096	0.191	0.127	0.100
More serious individual crime									
Second-order Markov time varying	0.086	0.076	0.062	0.091	0.081	0.065	0.100	0.090	0.072
First-order Markov time varying	0.098	0.088	0.071	0.080	0.072	0.058	0.093	0.083	0.066
Reduced time summary invariant	0.084	0.075	0.061	0.093	0.083	0.066	0.100	0.090	0.071
Household crime									
Second-order Markov time varying	0.284	0.229	0.198	0.323	0.267	0.236	0.207	0.161	0.134
First-order Markov time varying	0.285	0.230	0.201	0.337	0.278	0.245	0.218	0.171	0.144
Reduced time summary invariant	0.285	0.229	0.199	0.324	0.268	0.236	0.206	0.160	0.134

5. Conclusions

Time varying grouping variables can be very useful in explaining the latent construct and providing good model fit for the structural component of the model. However, their use of a large number of parameters can cause data sparseness which can make model diagnostic tests invalid and lead to model identifiability issues. This paper proposed alternative coding schemes and develops a set of procedures that can be used to determine which is the most appropriate for a particular set of data.

When applying these procedures to the NCVS, we found that the second-order Markov time varying coding scheme fit the data best when parsimony was not taken into account. However, when parsimony was accounted for, the reduced time-invariant summary coding scheme fit best. This finding was true regardless of the latent construct or the time varying grouping variable being analyzed. Therefore, when the measurement component is the main focus of analysis, the reduced time-invariant summary coding scheme can be used to reduce the number of parameters used in the structural component freeing up more degrees of freedom for fitting the measurement component.

However, these results are specific only to the NCVS data. Other latent constructs may be more related to the event characteristics. Therefore, prior to conducting an MLCA this process should be applied to all appropriate time varying grouping variables. To fully understand the relationship between event characteristic coding schemes or behavioral characteristic coding schemes and the latent construct, we plan to conduct a simulation study.

References

Berzofsky, M. E. (2009). Survey classification error analysis: Critical assumptions and model robustness. Presented at the annual meeting of the Classification Society and Interface Society, St. Louis, MO.

Biemer, P.P. (2010). *Latent Class Analysis of Survey Error*. Hoboken, NJ: John Wiley & Sons

Biemer, P. (2004), An analysis of classification error for the revised Current Population Survey employment questions, *Survey Methodology*, *30*(2), 127–140.

Biemer, P. P. (2000), An Application of Markov Latent Class Analysis for Evaluating Reporting Error in Consumer Expenditure Survey Screening Questions, RTI Technical Report for the US Bureau of Labor Statistics, RTI International, Research Triangle Park, NC.

Biemer, P. (2004), An analysis of classification error for the revised Current Population Survey employment questions, *Survey Methodology*, *30*(2), 127–140.

Biemer, P., & Berzofsky, M. (in press). Some issues in the application of latent class models for questionnaire design. In J. Madans, K. Miller, G. Willis, & A. Maitland (Eds.) Questionnaire evaluation methods, Hoboken, NJ: John Wiley & Sons.

Biemer, P., & Bushery, J. (2001), On the validity of Markov latent class analysis for estimating classification error in labor force data, Survey Methodology, 26(2), 136–152.

Kalton, G., D. Kasprzyk, and D. McMillen. 1989. Nonsampling errors in Panel Surveys. In *Panel Surveys*, edited by G. Kalton, D. Kasprzyk and D. McMillen. New York: Wiley

Poulsen, C. A. (1982), Latent Structure Analysis with Choice Modeling Applications, Aarhus School of Business Administration and Economics, Arhus, Denmark.

Schwarz, Gideon E. (1978). "Estimating the dimension of a model". Annals of Statistics 6 (2): 461–464

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

Van de Pol, F., & Langeheine, R. (1990), Mixed Markov latent class models,in C. C. Clogg, ed., Sociological Methodology, Blackwell, Oxford, pp. 213–247.

Wiggins, L. M. (1973). Panel Analysis, Latent Probability Models for Attitude andBehavior Processing, Elsevier SPC, Amsterdam.

U.S. Dept. of Justice, Bureau of Justice Statistics. NATIONAL CRIME VICTIMIZATION SURVEY LONGITUDINAL FILE, 1995-1999 [Computer Bibliographic Citation: file]. Conducted by U.S. Dept. of Commerce, Bureau of the Census. ICPSR04414-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 2007-03-14. doi:10.3886/ICPSR04414

Vermunt, J. K. (1997). *LEM: A general program for the analysis of categorical data*. Tilburg Netherlands: Department of Methodology and Statistics, Tilburg University.

Visher, C. A., & K. McFadden. (1991). *A comparison of urinalysis technologies for drug testing in criminal justice*. National Institute of Justice Research in Action. Washington, DC: U.S. Department of Justice.