Using a Mixture Latent Markov Model to Analyze Longitudinal U.S. Employment Data Involving Measurement Error

Jay Magidson,[1] Jeroen K. Vermunt,[2] and Bac Tran[3]
(1) Statistical Innovations Inc., Belmont, MA., U.S.
(2) Department of Methodology, Tilburg University, Tilburg, the Netherlands
(3) U.S. Bureau of the Census, Suitland, MD., U.S.

Abstract

Latent Markov modeling is used as an alternative to the Current Population Survey (Census, 2002) reinterviewing methodology for estimating the measurement error in the recorded employment status. This alternative methodology, which is implemented in the syntax version the Latent GOLD program, turns out to be a promising new approach for estimating measurement error in longitudinal surveys. However, it is important to take into account unobserved heterogeneity in the initial-state and transition probabilities because the size of the measurement error is overestimated when unobserved heterogeneity is not taken into account.

1. Background and Statement of the Problem

The Current Population Survey (CPS) is a national household survey conducted by the U.S. Census Bureau for the Bureau of Labor Statistics (Census, 2002). It is designed to generate monthly national estimates of labor force characteristics including employment status - i.e, counts of: (E) employed, (U) unemployed, and (N) not in the labor force. For a variety of reasons, employment status questions in the CPS may not elicit true employment status. Thus, a respondent's true employment status is unknown. The primary goal was to improve the methodology for estimating the magnitude and direction of measurement errors that exist in employment status elicited by the survey.

Some years ago the U.S. Census Bureau started exploring the solution of utilizing a latent Markov (LM) structure to estimate the measurement error as a more cost effective and timely approach than its current reinterview methodology (Biemer and Bushery, 2000; Tran & Winters, 2003). Rather than incurring a monthly cost of reinterviewing a sample of respondents, the LM approach replaces the reinterview with the repeated measurements that are readily available at no additional cost as part of the longitudinal design. On the other hand, the use of repeated measures instead of a reinterview means that differences in a respondent's state from one month to another may reflect either measurement error or a true transition from one state to another (e.g., from being employed to being unemployed). Thus, it is necessary to model both measurement and transitions simultaneously as part of the model.

The main contribution of the current project* is that it shows that it is important to use a mixture LM model instead of a standard LM model. Use of the latter may seriously overestimate the size of measurement error probabilities.

Below we first introduce the standard LM model, as well as its mixture extension. Then, details are provided on the implementation of these models in the Latent GOLD software package. Subsequently, we describe the CPS data set

*This research project was supported by contract #50 YABC-2-66060 from the U.S. Census Bureau

we used for our analysis and report the results obtained when applying the LM models to this data set.

2. Mixture Latent Markov Models

Latent Markov models - which are sometimes also referred to as hidden Markov models, latent transition models, or regime-switching models – have been increasingly used to analyze discrete-time longitudinal data where respondent observations contain measurement error (Collins & Wugalter, 1992; Hagenaars, 1990; Poulsen, 1982; Van de Pol & Langeheine, 1990; Vermunt, Magidson & Tran, 2008). The LM approach defines the true states as categories (latent classes) of a dynamic latent (unobservable) variable within a statistical model. The *latent* aspect of the LM model allows for *measurement error*. The *Markov* assumption is reflected in the LM model via *transition probabilities* which allow for correlation between a respondent's true state at times $t-1$ and t. The inclusion of both transition probabilities as well as measurement error parameters allows the LM structure to isolate changes to a respondent's true state and thus obtain pure estimates of measurement error (Hagenaars, 1990).

Let y_{it} denote the observed value of the dependent variable of interest at time point t for a person with response pattern i. In our application y_{it} is a nominal variable with $M = 3$ categories; that is, $1 \leq y_{it} \leq M$. The total number of time points is $T+1$, $0 \leq t \leq T$. The response vector of length $T+1$ containing the responses for pattern i is denoted by $\mathbf{y}_i$, and the associated model probability by $P(\mathbf{y}_i)$. A LM model is a model with $T+1$ latent variables, each having K categories. In the current study, we will only work with models in which $K = M = 3$, the number of latent labor force states is assumed to be equal to the number of observed labor force states. Let x_t denote a possible value of the latent variable at time point t, where $1 \leq x_t \leq K$. The LM model of interest has the following form

$$P(\mathbf{y}_i) = \sum_{x_0=1}^{K} \sum_{x_1=1}^{K} ... \sum_{x_T=1}^{K} P(x_0) \prod_{t=1}^{T} P(x_t|x_{t-1}) \prod_{t=0}^{T} P(y_{it}|x_t). \tag{1.1}$$

The unknown model probabilities to be estimated are the initial latent state probabilities $P(x_0)$, the latent transition probabilities $P(x_t|x_{t-1})$, and the classification error probabilities $P(y_{it}|x_t)$.

The model probabilities can be reparameterized as follows:

$$P(x_0 = s) = \frac{\exp(\alpha_s)}{\sum_{k=1}^{K} \exp(\alpha_k)}, \tag{1.2}$$

$$P(x_t = r|x_{t-1} = s) = \frac{\exp(\gamma_{rs}^t)}{\sum_{k=1}^{K} \exp(\gamma_{ks}^t)} \tag{1.3}$$

$$P(y_{it} = \ell|x_t = s) = \frac{\exp(\beta_{\ell s})}{\sum_{m=1}^{M} \exp(\beta_{ms})} \tag{1.4}$$

that is, as multinomial logistic equations (see, for example, Vermunt, Langeheine, and Böckenholt, 1999). Note that the $\beta_{\ell s}$ parameters do not contain a superscript t, which means that the classification errors are assumed to be time homogeneous, an assumption needed for the identification of a LM model for a single response variable and with time-heterogeneous transition probabilities. Our focus here is on the estimation of these time-homogeneous measurement error probabilities $P(y_t = \ell|x_t = s)$. Furthermore, note that 1, K, and M constraints are needed to

identify the set of α_s, γ_{rs}^t, and $\beta_{\ell s}$ parameters, respectively. Here, we assume that $\alpha_1 = 0$, $\gamma_{11}^t = \gamma_{22}^t = \gamma_{33}^t = 0$, and $\beta_{11} = \beta_{22} = \beta_{33} = 0$. Under these identifying constraints, we can write α_s, γ_{rs}^t, and $\beta_{\ell s}$ as

$$\alpha_s = \log\left[\frac{P(x_0 = s)}{P(x_0 = 1)}\right] \tag{1.5}$$

$$\gamma_{rs}^t = \log\left[\frac{P(x_t = r|x_{t-1} = s)}{P(x_t = s|x_{t-1} = s)}\right] \tag{1.6}$$

$$\beta_{\ell s} = \log\left[\frac{P(y_{it} = \ell|x_t = s)}{P(y_{it} = s|x_t = s)}\right], \tag{1.7}$$

respectively. This means that α_s can be interpreted as the logit (or log odds) of having initial state s rather than state 1, γ_{rs}^t as the logit of making a transition from state s to state r rather than staying in state s, and $\beta_{\ell s}$ as the logit of misclassifying someone with latent state s by assigning observed state ℓ rather than the correct observed state s. Below, we will refer to the latter two as "transition" and "error" coding, respectively. Note that the γ_{rs}^t and $\beta_{\ell s}$ will typically take on large negative values since it is much more likely to remain in the current state than to make a transition to another state, and it is much more likely to measure the correct state than an erroneous state. For example, if the transition probability $P(x_t = 2|x_{t-1} = 1) = .01$ and the probability of no transition $P(x_t = 1|x_{t-1} = 1) = .98$, $\gamma_{rs}^t = \log(.01/.98) = -4.58$.

As part of the current project, we also simulated the effects of various violation of the assumptions underlying the traditional (homogeneous, first-order) Markov model on the estimates of measurement errors. An important result of the simulation study was that a violation of the homogeneous transition probability assumption such that the heterogeneity is correlated with the initial-state probability would produce inflated estimates of measurement error. To determine whether such a violation exists in the CPS application, in addition to estimating the traditional LM model, we also used a mixture LM model where two LM chains were specified. This two-class mixture LM model has the following form (Van de Pol & Langeheine, 1990; Vermunt, Magidson & Tran, 2008):

$$P(\mathbf{y}_i) = \sum_{w=1}^{2} \sum_{x_0=1}^{K} \sum_{x_1=1}^{K} ... \sum_{x_T=1}^{K} P(w)\, P(x_0|w) \prod_{t=1}^{T} P(x_t|x_{t-1}, w) \prod_{t=0}^{T} P(y_{it}|x_t). \tag{1.8}$$

Here, w denotes one of the two categories of a time constant latent variable, $w = 1$ or 2. As can be seen, in this mixture model, both the initial-state probabilities and the transition probabilities are allowed to differ for each latent class. In addition, we have a new set of parameters, the mixture proportions $P(w)$.

A more restricted special case of the two-class mixture LM is the mover-stayer LM model (Van de Pol & Langeheine, 1990; Vermunt, Magidson & Tran, 2008). As an unrestricted mixture version of the model, it violates the traditional LM assumptions due to the existence of different transition probabilities across 2 subgroups. There is a group of movers who have nonzero probabilities of moving from one state to another over the time period, and a group of stayers who do not change. This model is obtained by assuming

$$P(x_t = r|x_{t-1} = s, w = 2) = 0, \tag{1.9}$$

for $r \neq s$. For the stayer class ($w = 2$), the transition probabilities are a priori fixed to 0, but the probabilities in the mover class – $P(x_t = r|x_{t-1} = s, w = 1)$ –

are not restricted. The restrictions on the transition probabilities for the stayer class can also be defined using our logistic formulation. That is,

$$P(x_t = r|x_{t-1} = s, w) = \frac{\exp(\gamma^t_{rsw})}{\sum_{k=1}^{K} \exp(\gamma^t_{ksw})}, \tag{1.10}$$

with $\gamma^t_{rs2} = -100$ for $r \neq s$ and 0 otherwise. Note that assuming that the logit of a transition for $w = 2$ equals -100 is the same as saying that the probability of a transition equals 0. The degree of the unobserved heterogeneity depends on the size of the stayer class $w = 2$, as well as on the difference between the transition probabilities of the mover and stayer classes.

3. Implementation in Latent GOLD program

A much more extended LM model was implemented in the syntax version of the Latent GOLD program (Vermunt and Magidson, 2007). Apart from the rather basic models described above, models that can be estimated include:

- higher-order Markov models;
- models containing time-constant and time-varying covariates such as gender, age, ethnicity, etc., affecting specific components of the models, such as initial-state probabilities, transition-probabilities, measurement error probabilities, and mixture proportions (Vermunt, Langeheine & Böckenholt, 1999; Paas, Vermunt, Bijmolt, 2007);
- models with multiple observed variables (Paas, Vermunt & Bijmolt, 2007);
- models for multiple associated Markov processes;
- models for observed response variables of other scale types, such as variables which are ordinal, continuous, or counts (Dias, Ramos & Vermunt, 2007);
- models for multilevel data similar to the multilevel latent class models proposed by Vermunt (2003).

Not only more extended models can be dealt with, but it is also possible to perform estimation with complex sampling design features such as weighting, clustering and stratification.

Assuming that n_i is the observed frequency count for data pattern i, maximum likelihood estimation of the mixture LM model parameters involves maximizing the following log-likelihood function:

$$L = \sum_{i=1}^{I} n_i \log P(\mathbf{y}_i). \tag{1.11}$$

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

$$P(w, \mathbf{x}|\mathbf{y}_i) = \frac{P(w, \mathbf{x})P(\mathbf{y}_i|w, \mathbf{x})}{P(\mathbf{y}_i)}, \tag{1.12}$$

which is the joint conditional distribution of the $T + 2$ latent variables given the data and the model parameters. In the M step we compute new estimates of

the model probabilities from an expanded table with $I \cdot 2 \cdot K^{T+1}$ "observed" cell entries $n_i \cdot P(w, \mathbf{x}|\mathbf{y}_i)$ (Vermunt, Langeheine and Böckenholt, 1999). It can easily be seen that computation time and computer storage increases exponentially with the number of time points, which makes the standard EM algorithm impractical or even impossible to apply with more than a few time points (Vermunt, Langeheine and Böckenholt, 1999).

To circumvent this problem, the Latent GOLD program uses a special variant of the EM algorithm for LM models that is called the Baum-Welch or forward-backward algorithm (Baum et al., 1970; Paas, Vermunt & Bijmolt, 2007). Use of this algorithm allows for the estimation of models for very long time series, as well as easily handling of incomplete data when such is present. Details on the implementation of the Baum-Welch algorithm – which was originally developed for hidden Markov models – in the estimation of mixture LM models can be found in Vermunt, Tran & Magidson (2008).

Below we provide the Latent GOLD syntax for standard, mixture, and mover-stayer LM models with time-heterogeneous transition probabilities. 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, which are called "`options`", "`variables`", and "`equations`". Although not shown below, the `options` section should contain the statement "`missing=includeall`", which indicates that records with missing values should be retained in the analysis.

An example of a set up for a standard LM model with time-varying transition probabilities is the following:

```
\\ standard LM model
variables
   caseid id;
   dependent Y nominal;
   independent t nominal;
   latent X nominal dynamic 3;
equations
   X[=0] <- 1;
   X <-  1 + X[-1] + t + X[-1] * t;
   Y <-  1 + X;
```

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 `dynamic`). Note that the independent variable `t` is the time variable that we included as one of the columns in the data set. This variable has a missing value for time point 0, and the values 1, 2, etc. for the other time points.

In the `equations` section, the dynamic latent variable is referred to in a different way depending on whether it is the initial state (`X[=0]`), the state at a particular time point (`X`), or the state at the previous time point (`X[-1]`). Note that these three equations are logit equations, one for the initial latent state (`X[=0]`), one for the latent state at a particular time point (`X`) conditional on the latent state at the previous time point (`X[-1]`), and one for the response variable (`Y`). If the model would contain more than a single response variable, one would have a separate equation for each response variable. The model for `X[=0]` contains only an intercept or constant term, which is denoted by `1`. The model for `X` contains an intercept, an effect of the latent state at the previous time point `X[-1]`, an

effect of the time variable `t`, as well as an interaction effect between the state at the previous time point and the time variable. The equation for the observed response `Y` contains an intercept and an effect of `X`.

Using the special coding scheme for the logit parameters introduced in equations (1.6) and (1.7), exactly the same LM model can also be defined as follows:

```
\\ standard LM model with special coding of logit parameters
equations
   X[=0] <- 1;
   X <- (~tra) 1 | X[-1] + (~tra) t | X[-1];
   Y <- (~err) 1 | X;
```

There are two important differences compared to the previous set up. The first difference is that we make use of the option to condition a parameter set on the value of other variables, which is achieved with the symbol "`|`" Note that an intercept plus an effect of `X[-1]` is the same as having an intercept that depends on `X[-1]`, and an effect of `t` plus a `X[-1] * t` interaction term is the same as having an effect of `t` that depends on `X[-1]`. The second difference is that we inserted "`(~tra)`" or "`(~err)`" before some of the regression terms. These modify the coding of the coefficients concerned into transition (error) coding; that is, a coding in which the "no transition" ("no error") category serves as the reference category. Note that this implies that the reference category depends on the origin state (true state).

A mixture LM model can be defined as follows:

```
\\ mixture LM model
variables
   caseid id;
   dependent Y nominal;
   independent t nominal;
   latent X nominal dynamic 3, W nominal 2 coding=first;
equations
   W <- 1;
   X[=0] <- 1 | W;
   X <- (~tra) 1 | X[-1] + (~tra) W | X[-1] + (~tra) t | X[-1];
   Y <- (~err) 1 | X;
```

The difference from the previous models is that we added the two-class latent variable `W` which has its own equation and which appears in the equations for the initial state and the transition probabilities. This model could be expanded in two different ways: we could include an iteraction between `W` and `t` in the model for `X`, and we could allow the measurement error to dependent on `W`.

The mover-stayer LM model is defined as follows:

```
\\ mover-stayer LM model
equations
   W <- 1;
   X[=0] <- 1 | W;
   X <- (~tra) 1 | X[-1] + (g~tra) W | X[-1] + (~tra) t | X[-1];
   Y <- (~err) 1 | X;
   g = -100;
```

The difference compared to the previous set up is that the term `W | X[-1]` is now restricted to be equal to -100 for the stayer class. This is achieved by giving a label to the parameter set concerned (here `g`) and adding the required parameter constraint at the end of the `equations` section.

Table 1. Estimated Misclassification Percentages Based on the Standard LM Model and the Mover-Stayer LM Model

True Employment Status (x_t)	Observed Employment Status (y_{it})					
	Standard LM			Mover-Stayer LM		
	E	U	N	E	U	N
E	97.8	0.6	1.6	98.5	0.6	0.8
U	9.2	70.1	20.7	4.7	78.2	17.1
N	1.9	0.9	97.2	1.1	1.0	97.8

4. Sample Data and Results

The CPS employs a rotation group design structure: Upon entering the CPS study, an individual is interviewed for 4 consecutive months, and then is out of the sample for 8 months prior to being interviewed again for another 4 consecutive months (Census, 2002). To illustrate the LM approach we utilize sample data consisting of the 14 rotation groups who were interviewed at least once during the 4 month period June - September 2006. All such observations were used to estimate the models - respondents in 4 rotation groups each contributed 1 month of data, 4 groups contributed 2 months, 4 groups contributed 3 months and 2 groups provided 4 months. In total, there were 123,147 respondents, approximately 8,000 per rotation group.[†]

Based on the Bayesian Information Criteria (BIC), the mixture LM model provided a better fit to the data than the traditional LM model. In addition, the 2 classes having different transition probabilities were highly correlated with the initial state and lower measurement error estimates were obtained than those from the traditional LM model, as suggested by our simulation.

Close examination of the transition probabilities obtained under the mixture model suggested that a mover-stayer structure, as suggested by Tran & Winters (2003), may provide a good fit. As indicated above, such a model is obtained from the 2-class mixture LM model by restricting the off-diagonal entries in the transition probability matrix to be zero (the stayer class). This mover-stayer model provided a further improvement in the BIC, supporting the mover-stayer variant restrictions on the mixture LM model.

The mover class in the mover-stayer LM model consisted of approximately 20%, the remaining 80% were stayers. The mover class was much more likely to be "Unemployed" than the stayer class during the initial time point – 20% for the movers versus less than 2% for the stayers. As mentioned above, when a strong association such as this is present, the simulation suggested that measurement error estimates obtained from the traditional LM model would be inflated (upward bias). That is, wrongly ignoring unobserved heterogeneity here (i.e., mistakenly specifying the traditional LM model) would cause measurement/ misclassification errors to be overstated. Comparison of the traditional LM model with the mover-stayer LM model in our empirical analysis provided results consistent with the simulation result.

Table 1. compares the measurement error estimates obtained from the mover-stayer LM model with those from the traditional LM model. As can be seen, under each of the 3 true states, the off-diagonal terms reflecting the estimated measurement errors are smaller under the mover-stayer LM structure. Our con-

[†]Similar results to those reported here were obtained in a more complete analysis based on 14 time points and which took into account the complex sampling design of the CPS.

clusion is that latent Markov models can be used for estimating measurement error as long as unobserved heterogeneity is taken into account, for example, with a mover-stayer structure.

References

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

Biemer, P. P., & Bushery, J. M. (2000). On the validity of Markov latent class analysis for estimating classification error in labor force data, *Survey Methodology, 26*, 139–152.

Census (2002). *Current Populations Survey: Design and Methodology*. Technical paper 63RV.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Dias, J.G., Vermunt, J.K., & Ramos, S. (2007). Latent class analysis of financial time series. Submitted from publication.

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

Paas, L. J., Vermunt, J. K., & Bijmolt, T. H, (2007). Discrete-time discrete-state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 170*, 955–974 .

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

Tran, B. & Winters, F. (2003). Markov latent class analysis and its application to the Current Population Survey in estimating the response error. Proceedings of the Section on Survey Research Methods, American Statistical Association.

Van de Pol, F., & Langeheine, R. (1990) Mixed Markov latent class models. em Sociological Methodology, 20, 213–247.

Vermunt, J. K., Langeheine, R., & 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. (2003). Multilevel latent class models. *Sociological Methodology, 33*, 213–239.

Vermunt, J.K. & Magidson, J. (2007). *LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module*. Belmont, MA.: Statistical Innovations Inc.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (Ed.), *Handbook of Longitudinal Research: Design, Measurement, and Analysis*. Elsevier. In press.