Estimating the Measurement Error in the Current Population Survey Labor Force–A Latent Class Analysis Approach With Sample Design[1]

Bac Tran, Justin Nguyen
U.S. Census Bureau, 4700 Silver Hill Road, Washington, D.C. 20233-1912
Bac.Tran@census.gov*, *Justin.D.Nguyen@census.gov

Key Words: Latent Class Models, Panel Data, Unemployment, Rotation Group, Month-in-Sample, Classification Probability, Interview, Census Bureau.

I. Introduction

The Current Population Survey (CPS) is a U.S. national household survey conducted by the U.S. Census Bureau for the Bureau of Labor Statistics. It is designed to generate national and state-level estimates of labor force characteristics such as: employed (E), unemployed (UE) and not in the labor force (NILF); demographic characteristics; and other characteristics of the non-institutionalized civilian population. Previous papers (Biemer 2000, Tran 2003, Tran 2004) applied traditional first-order Latent Markov models to estimate measurement error in CPS labor force. However, those models could not deal with the unobserved heterogeneity that meant that there were groups of sample persons having different transition and error probabilities. Furthermore, the CPS sample design was not taken into account in the analysis. This resulted in overestimating the measurement error by a substantial amount. This conclusion is supported by both our empirical analysis of the complexity of CPS data as well as by our simulation results. Also, the analysis showed that month-in-sample 1 has more measurement error in estimating the unemployment rate than the other months in sample (2-8). Past research (Causey, 1976) indicated that month-in-sample 1 likely produces less bias in estimating the labor force.

This paper will present a validation of applying Latent Markov models to estimate measurement error by a thorough simulation. This paper introduces the Mover-Stayer Latent Markov model (a mixture latent Markov model) and its application to the CPS data in order to estimate measurement error for the labor force status. The analysis used LatentGold4.5, software developed by Statistical Innovations, to implement the model estimation.

The CPS uses addresses from the most current U.S. Census, adding new construction, as the frame. The total sample size is about 72,000 assigned households per month. The CPS uses a 4-8-4 rotating panel design, i.e. 4 months in, 8 months out, and 4 months in. In the CPS, the same respondents are interviewed at several points following the pattern 4-8-4. For any given month, the CPS sample is grouped into eight sub-samples corresponding to the eight rotation groups.

[1] *This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.*

II. Mover-Stayer Latent Markov Model- A Mixture Latent Markov Model

Latent Class Analysis (LCA) treats the true classification of the labor force status as an unobserved variable. The observed variables (A, B, C, D: labor force for four consecutive months in our study, see Figure 1) obtained from the CPS survey in a panel survey are fallible indicators of the latent variable X. LCA suggests a relationship between observed variables and latent variables through a mathematical equation. Under the equation the table of observed data is viewed as a partial table from a full table of observed and unobserved data. The Markov assumption is employed (see below), hence the so-called Markov Latent Class Analysis (MLCA). With the homogeneity in the distribution of the labor force in the population we have one Markov chain, as in a traditional first-order Markov chain. With heterogeneity we have more than one chain. A simple case is two chains as in our study, the Mover-Stayer model.

As mentioned above, a small portion of the population that becomes unemployed and stays unemployed for a long time could violate the Markov assumption. First-order Markov models cannot deal with the heterogeneity of the underlying latent class. Before presenting the Mover-Stayer Markov model (M-S model) we would like to show the traditional first-order Markov model (MLC). The MLC has the following form (Van De Pol and Langeheine, 1990):

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

where T +1 is the number of time points ($0 \le t \le T$), y_i is response vector of i^{th} observation of length T+1 (in our application T=3), y_{it} is the t^{th} component of y_i, x_t denotes a possible value of a latent variable at time t where x_t = 1, 2, .., K (in our study K=3, representing the three categories of labor force Employed, Unemployed, and Not In Labor Force which are abbreviated by E, U, and NILF). The Markov assumption states that

$$P(x_t| x_{t-1}) = P(x_t| x_{t-1}, x_{t-1}, \dots, x_0).$$

The assumptions of the first-order Markov Latent Class in the equation (1) are:

1. x_t is independent of $x_{t-2}, x_{t-3}, \dots, x_0$.
2. There is no unobserved heterogeneity.
3. Classification errors are independent across time points.

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

They are parameterized as follows:

$$
\begin{aligned}
P(x_0 = s) &= \frac{\exp(\alpha_s)}{\sum_{k=1}^{K} \exp(\alpha_k)}, \\
P(x_t = r | x_{t-1} = s) &= \frac{\exp(\gamma_{rs}^t)}{\sum_{k=1}^{K} \exp(\gamma_{ks}^t)} \\
P(y_{it} = m | x_t = s) &= \frac{\exp(\beta_{rs})}{\sum_{m=1}^{M} \exp(\beta_{ks})}
\end{aligned}
$$

where α_s, γ^t_{rs}, and β_{rs} are parameters from the logit models for $P(x_0=s)$, $P(x_t=r|\ x_{t-1}=s)$ and $P(y_{it}=m|\ x_t=s)$ respectively.

The Mover-Stayer model, a two-class mixed Markov Latent Class model, assumes that there are two unobserved subgroups with different transition probabilities. The M-S model has the form

$$P(y_i) = \sum_{w=1}^{2} \sum_{x_0=1}^{K} \sum_{x_1=1}^{K} \cdots \sum_{x_T=1}^{T} P(w) P(x_0 \mid w) \prod P(x_t \mid x_{t-1}, w) \prod P(y_{it} \mid x_t, w) \quad (2)$$

where w=1 or 2 denotes two classes of latent variables and w=2 represents the stayer class, i.e.

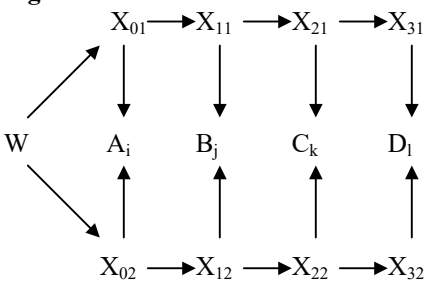
$$P(x_t = r | x_{t-1} = s, w = 2) = 0$$

for $r \neq s$, otherwise

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

In our study we assume classification error is time homogeneous. Therefore, in the parameterization equations there is no t term for the βs. The path diagram for model (2), applied to the CPS labor force with four consecutive time periods, looks like below.

Figure 1



The latent variable W represents two groups of sample persons (Unemployed and the other). The Xs are latent variables represent the true labor force status at four time points. A, B, C, D are the observed labor force variables used as the indicators of the Xs.

III. Validation Process

The main purpose of the use of MLCA is to estimate classification errors, and then estimate the true labor force status distribution under the correct model. We validated the method through a simulation. Simulate data was created from CPS labor force data collected over time. Basically, we considered three different model-type settings:

1. Traditional first-order homogeneous Markov
2. Mixture model with violation of homogeneity
3. Second-order Markov model violating the Markov assumption

The three settings will answer the question on how sensitive estimates from Markov Latent Class models are when certain model assumptions are violated. The first and second models were presented above. The third Markov model has the form:

$$P(\mathrm{y}_i) = \sum_{x_0=1}^{K} \sum_{x_1=1}^{K} \cdots \sum_{x_T=1}^{K} P(x_0) P(x_1|x_0) \prod_{t=2}^{T} P(x_t|x_{t-1}, x_{t-2}) \prod_{t=0}^{T} P(y_{it}|x_t) \quad (3)$$

The simulate data sets were created based on seven sets of transition probabilities and three sets of misclassification probabilities:

1. One set is in agreement with a homogeneous first-order Markov process.
2. Three sets are in agreement with a second-order Markov process, where the three sets represent small, moderate, and large violations of the first order Markov assumption.
3. Three sets are in agreement with a Mover-Stayer process, where the three sets represent small, moderate, and large violations of the homogeneous Markov assumption.

Three sets of misclassification errors correspond to the conditions of small, moderate, and high proportions of misclassification. These three conditions are based on the summary presented by Tran and Winters (2003). In total, we have 21 design cells and generated 1,000 replicates for each cell. Each replicate contains eight rotations, each rotation with about 6,000 sample units. Therefore, each replicate has about 48,000 sample units. For each of the data sets, the parameters of setting # 1 were estimated. When this is not the correct model, the correct model is also estimated. Depending on the design cell, this is either the Mover-Stayer model or the second-order Markov LC model (settings #2 or 3).

The four questions we want to answer with the simulation study are:

1. Is it is possible to detect whether model assumptions are violated?
2. Are the estimated misclassification probabilities unbiased when the correct model is specified?
3. Are the estimated misclassification probabilities biased when model assumptions are violated when an incorrect model is specified?
4. Are the estimated class sizes biased when model assumptions are violated when an incorrect model is specified?

The results from the simulation give the answers as follows:

1. Yes, it is possible to detect that model assumptions are violated, but only for large violations.
2. Yes, estimates of the misclassification probabilities are unbiased when the right model is specified.
3. Yes, there is an upward bias in the estimates of the misclassification probabilities, but it is surprisingly small. Only with a very extreme (and unrealistic) second-order process do we see substantial bias in the estimated

misclassification probabilities obtained with an incorrect first-order model.

4. Yes, estimates of the class sizes are biased downwards. With weak violations, this bias is negligible.

IV. Application to the CPS Data

We used CPS data from June 2006 to September 2006. The labor force status for four months was used as indicators of the latent variable X. The survey variables used were rotation, proxy, and interview mode. The grouping variable used was SEX (Male/Female). As in Tran (2004) the rotation variable was collapsed into two categories, rotation 1 versus the other rotations (2-8). The proxy variable had two categories, self and proxy. Variable MODE contained CATI, CAPI personal visit, and CAPI telephone.

We ran eight different models that are summarized as follows:

Table 1

	Model	LL/BIC(LL)/#parameters/ dof	p-value
1	X_0, $X_1\|X_0$, $X_2\|X1$, $X3\|X_2$, $A\|X_0$, $B\|X_1$, $C\|X_2$, $D\|X_3$	-215234.214 430785.640 26 216	0
2	$X_0\|S$, $X_1\|X_0S$, $X_2\|X_1S$, $X_3\|X_2S$, $A\|X_0$, $B\|X_1$, $C\|X_2$, $D\|X_3$	-212863.559 426141.936 34 450	0
3	X_0, $X_1\|X_0$, $X_2\|X_1$, $X_3\|X_2$, $A\|X_0$PRM{AX_0,AP,AR,AM} $B\|X_1$PRM{BX_1,BP,BR,BM} $C\|X_2$PRM{CX_2,CP,CR,CM} $D\|X_3$PRM{DX_0,DP,DR,DM}	-214718.750 430047.5254 50 106584	1
4	$X_0\|S$, $X_1\|X_0S$, $X_2\|X_1S$, $X_3\|X_2S$ $A\|X_0$PRM{AX_0,AP,AR,AM} $B\|X_1$PRM{BX_1,BP,BR,BM} $C\|X_2$PRM{CX_2,CP,CR,CM} $D\|X_3$PRM{DX_0,DP,DR,DM}	-212323.593 425354.8146 58 179026	1
5	W $X_0\|W$, $X_1\|X_0W$, $X_2\|X_1W$, $X_3\|X_2W$, $A\|X_0$, $B\|X_1$, $C\|X_2$, $D\|X_3$	-215099.933 430553.6795 29 213	0
6	W|S $X_0\|WS$, $X_1\|X_0WS$, $X_2\|X_1WS$, $X_3\|X_2WS$, $A\|X_0$, $B\|X_1$, $C\|X_2$, $D\|X_3$	-212695.790 425879.6005 40 444	0
7	W $X_0\|W$, $X_1\|X_0W$, $X_2\|X_1W$, $X_3\|X_2W$, $A\|X_0$PRM{AX_0,AP,AR,AM} $B\|X_1$PRM{BX_1,BP,BR,BM} $C\|X_2$PRM{CX_2,CP,CR,CM} $D\|X_3$PRM{DX_0,DP,DR,DM}	-214593.312 429833.2497 53 106581	1
8	W|S $X_0\|WS$, $X_1\|X_0WS$, $X_2\|X_1WS$, $X_3\|X_2WS$ $A\|X_0$PRM{AX_0,AP,AR,AM} $B\|WX_1$PRM{BX_1,BP,BR,BM} $C\|WX_2$PRM{CX_2,CP,CR,CM} $D\|X_3$PRM{DX_0,DP,DR,DM}	-212149.942 425080.7164 64 179020	1

Notes:
CATI: Computer Assisted Telephone Interviewing
CAPI: Computer Assisted Personal Interviewing
X_0, X_1, X_2, and X_3: latent variables for labor force
W: latent mover-stayer variable
A, B, C, D: labor force indicators for June – September 2006
P: Proxy (Proxy/Self)
R: Rotation (1 versus 2-8), rotation 1 is not conducted in CATI
M: Mode (CATI, CAPI personal visit, CAPI telephone)

We used LatentGold4.5 (LG4.5), software developed by Statistical Innovations (2007), to estimate the model parameters. All the models were identifiable. We use the following criteria to identify a good model.

- The model is identifiable.
- The p-value of the likelihood ratio p-value should be greater than 0.01.
- The Bayesian information criterion (BIC), defined as L^2 – log (N) degrees of freedom, should be the smallest among all competing models.

Based on those criteria model, the best model for these data is model 8.

V. Classification Probabilities

We compared our estimates of the CPS classification probabilities with similar estimates obtained from previous papers (Biemer & Bushery 2000, Tran & Winters 2003, and Tran & Mansur 2004). The results are summarized in Table 2.

The correct classification probability Pr(observed= Unemployed| True = Unemployed)) for the Unemployed category is estimated as 79.16 percent. This figure is close to 81.81 percent, as found in Biemer and Bushery (1993 data). However, in this study we use the sample design with weight (see **VI**). The estimates and their standard errors are included in Table 2. Below we give the probabilities Pr(response| Proxy, Rotation, Mode, True status)

Output 1

				Probability		
P	R	M	X	E	U	NILF
2	1	2	1	0.9822	0.0094	0.0084
2	1	2	2	0.0385	0.7666	0.1950
2	1	2	3	0.0388	0.0240	0.9373
1	1	3	1	0.9961	0.0026	0.0013
1	1	3	2	0.1464	0.7657	0.0879
1	1	3	3	0.0063	0.0064	0.9873
2	1	3	1	0.9901	0.0039	0.0059
2	1	3	2	0.0895	0.7127	0.1978
2	1	3	3	0.0283	0.0135	0.9582
1	2	2	1	0.9900	0.0064	0.0035
1	2	2	2	0.0237	0.8557	0.1206
1	2	2	3	0.0041	0.0068	0.9891
2	2	2	1	0.9739	0.0096	0.0165
2	2	2	2	0.0134	0.7359	0.2507
2	2	2	3	0.0187	0.0145	0.9668
1	1	2	1	0.9920	0.0062	0.0018
1	1	2	2	0.0647	0.8462	0.0891
1	1	2	3	0.0087	0.0115	0.9798
2	2	3	1	0.9842	0.0041	0.0117
2	2	3	2	0.0320	0.7056	0.2623
2	2	3	3	0.0135	0.0081	0.9784
1	2	3	1	0.9948	0.0027	0.0025
1	2	3	2	0.0565	0.8177	0.1258
1	2	3	3	0.0030	0.0038	0.9933
2	2	1	1	0.9827	0.0062	0.0111
2	2	1	2	0.0442	0.7880	0.1678

```
 2  2  1  3    0.0207   0.0211   0.9582
1  2  1  1    0.9935   0.0041   0.0023
1  2  1  2    0.0728   0.8521   0.0751
1  2  1  3    0.0046   0.0100   0.9854
```

Looking at the correct classification for the Unemployed category alone extracted from Output 1, we have:

```
P     R    M       Pr(response=UE|X=UE,P,R,M)

1     1    2    0.8462
1     1    3    0.7657
1     2    1    0.8521
1     2    2    0.8557
1     2    3    0.8177
2     1    2    0.7666
2     1    3    0.7127
2     2    1    0.7880
2     2    2    0.7359
2     2    3    0.7056
```

We see that, in terms of measurement error, self-reporting is better than proxy reporting. If considering self-reporting as more accurate, then month-in-sample 2-8 has less measurement error than month-in-sample 1. This also means that month-in-sample one overestimates the unemployment rate more than the other months in sample.

Table 2: Classification Probabilities

Classification		**Previous Estimates**			**Current Estimate/ (s.e)**
True (estimated)	**Reported**	**Biemer& Bushery MLCA**	**Tran& Winters MLCA (1996-1999)**	**Tran& Mansur LCA (Jan2002-Dec 2003)**	**Tran &Nguyen Mover-Stayer (June06-Sep06)**
EMP	**EMP**	98.77 (1993) 98.73 (1995) 98.73 (1996)	98.74	97.35	98.89 (0.1)
	UE	0.34 (1993) 0.49 (1995) 0.37 (1996)	0.37	0.35	0.51 (0.06)
	NILF	0.89 (1993) 0.78 (1995) 0.79 (1996)	0.89	2.29	0.60 (0.08)
UE	**EMP**	7.06 (1993) 7.86 (1995) 8.57 (1996)	9.87	11.39	5.01 (1.44)
	UE	81.81 (1993) 76.09 (1995) 74.42 (1996)	71.38	71.51	79.16 (1.82)
	NILF	11.13 (1993) 16.04 (1995) 17.00 (1996)	18.75	17.10	15.83 (1.74)
NILF	**EMP**	1.41 (1993) 1.11 (1995) 1.13 (1996)	1.26	8.93	1.15 (0.12)
	UE	0.75 (1993) 0.69 (1995) 0.87 (1996)	0.72	1.85	1.06 (0.09)
	NILF	97.84 (1993) 98.20 (1995) 98.00 (1996)	98.03	89.22	97.79 (0.15)

VI. Sample Design With Weight

A sample unit, taken from a finite population, in this research is identified by an identification code and the following four characteristics:

1. Stratum
2. Primary Sampling Unit (PSU)
3. Sampling weight

The model parameters of a mixture latent Markov model are estimated by means of pseudo-Maximum Likelihood (PM) estimation (Skinner, Holt, and Smith, 1989) through the linearization variance estimator. Readers will find the technical details in the Technical Guide from Vermunt & Magidson (2007).

VII. Limitation

There are limitations when using maximum likelihood procedure with missing values: The procedure can deal with missing values on response variables, but not with missing values on covariates, and it assumes that the missing data are missing at random (MAR). The data prepared for this study were four-consecutive month data (June 2006 to September 2006) from the CPS. We need to use more data to fit the model. This study applied a simple Mover-Stayer model in which there were two classes, mover and stayer. There could be more than two classes. The weight we used for the analysis was the average of the second stage weights from four-month CPS data. We need to figure a weighting scheme that is better than the averaging.

VIII. Conclusion

This paper utilized the suggestion from Tran (2003) that heterogeneity of the underlying class of the labor force is necessary for the model. That idea leads to the investigation of Mover-Stayer Markov Latent Class models. This study validated the method by an extensive simulation, while Tran 2003 simulation was just a partial case. Furthermore, the sample design with weight is taken into consideration for this paper. With these features and investigations we are able to estimate the measurement errors by using Markov Latent Class Analysis, specifically Mixed Markov Latent Class Analysis.

References

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

Bailar, B.A.(1975). "The Effect of Rotation Group Bias on Estimates from Panel Surveys," *Journal of the American Statistical Association,* Vol.70, pp. 23-30.

Biemer, P.P (2004). "The Twelfth Morris Hansen Lecture Simple Response Variance: Then and Now", *Journal of Official Statistics*, 20 (3): 417-439

Biemer, P. and Wiesen, C. (2002). "Measurement Error Evaluation of Self-Reported Drug Use: A Latent Class Analysis of

the US National Household Survey on Drug Abuse," *Journal of Royal Statistical Society*, Part 1, 165, pp. 97-119.

Biemer, P. and Bushery, J. (2000). "On the Validity of Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data," *Survey Methodology*, Vol. 26, No. 2, pp. 139-152.

Causey, B. (1976). Draft for the Record, Internal Memorandum, dated June 9, 1976, "Findings on CPS Rotation Group Bias," Statistical Research Division, Bureau of the Census, Washington, D.C.

Goodman, L.A. (1961). "Statistical Methods for the Mover-Stayer Model," *Journal of the American Statistical Association*, 81, 354-365

Hagenaars, J.A, and McCutcheon A.L. (2002). Applied Latent Class Analysis. Cambridge University Press.

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

Mansur, K. and Shoemaker, Jr. H. (1999). "The Impact of Changes in the Current Population Survey on Time-in-Sample Bias and Correlations between Rotation Groups," *Proceedings of the Section on Survey Methods, American Statistical Association,* pp. 180-183.

McCutcheon, A.L. (1987). Latent Class Analysis. Newbury Park, CA: Sage.

Skinner, C.J., Holt, D., and Smith, T.M.F. (eds.) (1989), Analysis of Complex Surveys, New York: Wiley.

Shockey, J. (1988). "Adjusting of Response Error in Panel Surveys, A Latent Class Approach," *Sociological Methods and research,* Vol.17, No.1, August, pp. 478-488.

Tran, Bac and Winters, Franklin (2003). "Markov Latent Class Analysis and Its Application to the Current Population Survey in Estimating the Response Error," *2003 Proceedings of the American Statistical Association*, Statistical Computing Section [CD-ROM], Alexandria, VA: American Statistical Association.

Tran, Bac and Mansur, Khandaker (2004). "Analysis of the Unemployment Rate in the Current Population Survey- A Latent Class Approach," *2004 Proceedings of the American Statistical Association*, Statistical Computing Section [CD-ROM], Alexandria, VA: American Statistical Association.

Van De Pol, F., and De Leeuw, J. (1986). "A Latent Markov Model to Correct for Measurement Error," *Sociology Method and Research*, 15, 118-141

Van De Pol, F., and Langeheine, R. (1990). "Mixed Markov Latent Class Models," *Sociology Methodology*, 213-247

Vermunt, J. (1997). Log Linear Models for Event Histories. Thousand Oaks: Sage

Vermunt, J.K., Langeheine, R., and Bockenholt, U. (1999). "Latent Markov Models With Time-Constant and Time-Varying Covariates," *Journal Education and Behavioral Statistics*, 24, 178-205

Vermunt, J.K. (2003). "Multilevel Latent Class Models," *Sociology Methodology*, 33, 213-239

Vermunt, J.K. and Magidson, J. (2007) Technical Guide to Latent Gold 4.5. Belmont Massachusetts: Statistical Innovations Inc.