A longitudinal study of Polish emigration attitudes: a latent Markov model approach

Ewa Genge[1]

Abstract

Latent class analysis can be viewed as a special case of model–based clustering for multivariate discrete data. When longitudinal data are to be analyzed, the research questions concern some form of change over time. Latent Markov model is a variation of the latent class model that is applied to estimate not only the prevalence of latent class membership, but the incidence of transitions over time in latent class membership.

In 2004 Poland have joined the European Union. After this EU enlargement, many Polish citizens left the country. To mark this event we used model-based clustering approach for grouping and detecting inhomogeneities of the public attitudes to emigration from Poland. We focused especially on the latent Markov models with covariates which additionally allow us to investigate the dynamic pattern of the Polish attitudes to emigration for different demographic features. We used `depmixS4`, `Rsolnp` and `LMest` packages of R.

Keywords: *latent Markov model, panel data, model-based clustering*

JEL Classification: C33, J11

1. Introduction

On 1 May 2004 Poland and nine more nations with combined population of almost 75 million joined the EU. After Poland's accession to the European Union the number of declared emigration continued to increase. Taking into consideration only the official number of emigrants, especially the emigration rate[2], Poland ranks first in the most similar, i.e. postcommunist countries (Czech Republic, Hungary, Poland, Slovakia). In comparison with all of the countries that joined the EU in 2004, Poland is at the fifth place preceded by Cyprus, Lithuania, Latvia, Malta. However, the emigration rate in Poland is currently unprecedented, not only high, but also growing while in Lithuania and Latvia since 2010 i.e., decreasing trend has be observed (see Table 1)[3].

To mark the anniversary of over 10 years of the Polish EU membership we used model-based clustering approach for grouping and detecting inhomogeneities of the public attitudes

[1] Corresponding author: University of Economics in Katowice, Department of Economic and Financial Analysis, 1-go Maja 50, 40-287 Katowice, Poland, e-mail: ewa.genge@ue.katowice.pl.

[2] The number of registered departures from the country for a permanent stay abroad divided by a number of residents.

[3] Data that allow for international comparisons of migration flows in European Union countries are available from the European statistics agency, Eurostat.

to emigration from Poland. Despite numerous studies the situation in the field of migration in Poland is not well-known, and the actual rate of emigration is unclear (statistical data record only the officially declared migration). Therefore, we based our empirical research on sociological research conducted by Czapinski and Panek (2013) in Poland, which is carried out in a systematic manner. We focused especially on the latent Markov models with covariates, which additionally allow us to investigate the dynamic pattern of the Polish attitudes to emigration for different demographic features.

Country	2004	2005	2006	2007	2008	2009	2010	2011	2012
Czech Republic	0.34%	0.24%	0.33%	0.20%	0.50%	0.59%	0.58%	0.53%	0.44%
Estonia	0.21%	0.34%	0.41%	0.33%	0.33%	0.35%	0.40%	0.47%	0.48%
Cyprus	0.87%	1.36%	0.92%	1.50%	1.35%	1.23%	0.52%	0.58%	2.10%
Latvia	0.89%	0.78%	0.76%	0.70%	1.23%	1.77%	1.87%	1.46%	1.23%
Lithuania	1.11%	1.73%	0.98%	0.93%	0.80%	1.21%	2.65%	1.76%	1.37%
Hungary	0.04%	0.04%	0.04%	0.04%	0.10%	0.10%	0.13%	0.15%	0.23%
Malta	-	-	0.47%	1.24%	0.91%	0.94%	1.01%	0.92%	0.96%
Poland	**0.05%**	**0.06%**	**0.12%**	**0.09%**	**0.0%**	**0.60%**	**0.57%**	**0.69%**	**0.72%**
Slovenia	0.41%	0.43%	0.69%	0.74%	0.60%	0.92%	0.78%	0.59%	0.70%
Slovakia	0.12%	0.05%	0.06%	0.07%	0.09%	0.09%	0.08%	0.03%	0.04%

Table 1. The emigration ratio for countries that joined the UE in 2004.

2. Definition

The initial formulation of latent Markov (LM) model introduced by Wiggins (1973) has been developed in several directions, in connection with application in many fields (Bartolucci et al., 2015; Genge, 2014; van de Pol and Langeheine, 1990; Vermunt et al., 1999; Visser and Speekenbrink, 2010). Latent Markov model represents an important class of models for the analysis of longitudinal data, when response variables are categorical. The latent Markov model analyzes $P(\mathbf{y}_i)$ the probability function of the vector of responses over time by means of a latent transition structure defined by a first-order Markov process. For each time point *t*, the model defines one discrete latent variable constituted by *K* latent classes (which are referred to as latent states).

The model given in (1) relies on two main assumptions: first, it assumes that the latent state transitions occurring over time are modeled using the first-order Markov chain. Second, the latent states are connected to one or more observed response variables via a latent structure with conditional densities. The latter assumption implies that the observations in time *t* depend only on the latent states at time *t* and is often referred to as the local independence assumption which is the pillar of latent structure models.

Let y_{itj} denote the response of subject *i* at occasion *t* on response variable *j*, where $1 \le i \le n$, $1 \le t \le T$, $1 \le j \le J$, and $1 \le y_{itj} \le M_j$ where *n* is the number of subjects, *J* is the total number of response variables and M_j the number of categories for response variable *j*. The vector of responses for subject *i* at occasion *t* is denoted as $\mathbf{y}_{it}$ and the vector of responses at all occasions as $\mathbf{y}_i$.

The latent Markov model can be defined as follows:

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

The LMM is characterized by three probability functions:

1. $P(x_0)$ – an initial-state probability, i.e. the probability of having a particular latent initial state at *t = 0*.
2. $P(x_t \mid x_{t-1})$ – a latent transition probability, i.e. the probability of being in a particular latent state at time point *t* conditional on the latent state at time *t − 1*.

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1K} \\ \vdots & \vdots & \vdots \\ a_{K1} & \dots & a_{KK} \end{bmatrix}. \quad (2)$$

Assuming a homogenous transition process with respect to time, we achieve the latent transition matrix of transition probabilities a_{sr}, with $s, r, \dots, K$ denoting the probability of switching from latent state *s* to latent state *r*.

3. $P(y_{itj} \mid x_t)$ – a response probability, i.e. the probability of having a particular observed value on response variable *j* at time point *t* conditional on the latent state occupied at time point *t*.

When transitions are added to the latent class model, it is more appropriate to refer to the classes as states. The word class is rather more associated with a stable trait-like attribute whereas a state can change over time. This is especially useful when a model contains

covariates $\mathbf{z}_i$. In `depmixS4` package of R a generalized (multinomial) model logit link function for the effects of covariates on the transition probabilities is employed (see, for example Agresti, 2002; Vermunt, 1997). In this case, each row of the transition matrix is parameterized by a baseline category logistic multinomial, meaning that the parameter for the base category is fixed at zero.

The maximum likelihood estimation of the parameters of LM models involves maximizing the log-likelihood function $L(\mathbf{y}) = \sum_{i=1}^{n} \log P(\mathbf{y}_i)$. This problem can be solved by means of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The E step computes the joint conditional distribution of the *t+1* latent variables given the data and the current estimates of the model parameters. In the M step, standard complete data ML methods are used to update the unknown model parameters using expanded data matrix with the estimated densities of the latent variables as weights.

The EM algorithm, however, has some drawbacks. First, it can be slow to converge. Second, applying constraints to parameters can be problematic. It can be seen that computation time and computer storage increase with the number of points, which makes the standard EM algorithm impractical or even impossible to apply with more than a few time points (Vermunt et al., 1999). Therefore, `depmixS4` package of R uses special variant of EM algorithm for LM models, called Baum-Welch or forward-backward algorithm (Baum et al., 1970; Paas et al., 2007).

An important modeling issue is the choice of the number of states. The selection of the proper number of states is typically based on information criterion such as Bayesian Information Criterion (BIC) (Schwarz, 1978) or Akaike Infromation Criterion (AIC) (Akaike, 1974).

3. Example

The analyses presented below are based on Social Diagnosis questionnaires. The Social Diagnosis (Objective and Subjective Quality of Life in Poland) is a diagnosis of the conditions and quality of life of Poles as they report it.

We considered questionnaire items about the Polish emigration. The data concern one dichotomous outcome variable y_1 and one multinomial y_2 outcome variable measured at five

occasions (every two years, i.e. 2005, 2007, 2009, 2011, 2013)[4]. Unfortunately, for none of the years there is complete information. Therefore, we considered 538 complete observations at each point of time. In total, there is information on $n = 2690$ cases. The public data set, available at www.diagnoza.com (see also: Social Diagnosis, reports; Czapiński, Panek (eds.), 2013).

All computations and graphics in this paper have been done in `depmixS4` (Visser and Speekenbrink, 2014) package of R.

The following variables (questions) in years 2005, 2007, 2009, 2011 and 2013 were used in the analysis:

y_1 – Do you plan to go abroad within the next two years in order to work?

y_2 – To which country (economic emigration target)?

We also analyzed the covariates:

z_1 – education[5],

z_2 – age[6],

z_3 – social-professional status[7],

z_4 – occupation (active and inactive)[8].

In the first question respondents could choose one of two options: yes and no. In the second question the following countries were considered: Austria, Belgium, Denmark, Finland, France, Greece, Spain, Netherlands, Ireland, Luxemburg, Germany, Portugal, Sweden, UK, Italy, other UE countries, USA, Canada, Australia, other countries, Norway.

A reasonable theoretical approach might indicate that there are two latent states of survey respondents. Emigration enthusiasts and emigration sceptics. Supporters of emigration will tend to respond favorably towards leaving the country, with the reverse being the case for emigration sceptics. We might further expect that, "changing one's mind" into the other group is a function of each individual's education, occupation, social-professional status and age. We can investigate this hypothesis using a latent Markov model.

[4] Only those two variables are given at five occasions.
[5] 1-primary/no education; 2-vocational/grammar; 3-secondary; 4-higher and post-secondary.
[6] 1-up to 24 years; 2: 25-34 years; 3: 35-44 years; 4: 45-59 years; 5: 60-64 years; 6: 65+ years.
[7] 1: employees in public sector; 2: employees in private sector; 3: entrepreneurs/self-employed; 4: farmers; 5: pensioners; 6: retirees; 7: pupils and student; 8: unemployed; 9: other professionally inactive.
[8] 1: legislators, senior officials and managers; 2: professionals; 3: technicians and associate professionals; 4: clerks; 5: service workers and shop sales workers; 6: skilled agricultural and fishery workers; 7: craft and related trades workers; 9: plant and machine operators and assemblers; 10: elementary occupations.

The optimal number of states was chosen using information criteria for the basic model (Collins and Lanza, 2010), so we decided to choose two latent states.

We estimated parameters of two states using the EM algorithm. In further analysis we ran the test for significance of the coefficients. For the two states only age and occupation coefficients were significantly different from 0. By examining the estimated state-conditional response probabilities, we confirmed that the model indentified the two groups, with 8% in the pro-emigration group and 92% in the anti-emigration group. We labeled the smaller latent state emigration enthusiasts and the bigger emigration sceptics.

Latent state 1, emigration supporters, was characterized by a very high probability (98%) of a positive response to the first question about emigration. There was also the highest percentage (40%) of respondents ready to work in Ireland and in Germany (20%), 7% in the USA, 6% in Spain and 6% in the Netherlands, 2% in Austria and 2% in Denmark.

In contrast, those in latent state 2, emigration sceptics, were characterized by a low probability (1%) of a positive answer to the first variable. Almost everyone (99%) in this small group of people was ready to work in the USA[9].

A further relevant set of information provided by the latent Markov model is represented by the latent transition matrix **A** which shows the probability of switching from one latent state to another. The results related to the dynamics of the Polish attitudes to emigration are reported in Table 2. The values on the main diagonal of the transition matrix represent the state persistence that is the probabilities of remaining in a particular state. For example, the probability of staying in latent state 1 is $a_{11} = 0.18$, while the probability to remain in state 2 $a_{22} = 0.93$ is very high. The out-of-diagonal a_{sr} values indicate the probabilities of emigration state switching: for instance the attitude to emigration is not well represented by latent state 1 at time *t+1* and it is not very likely a persistence of this ready to emigration state but also a switch to the emigration sceptics state $a_{12} = 0.82$.

It is also interesting to notice that people who are not so ready to leave Poland at time *t* will not change their mind also at time *t+1* with probability $a_{22} = 0.93$, indicating stability of the behavior. They may also shift with the lowest probability to the emigration group represented by state one ($a_{21} = 0.07$).

[9] We also estimated parameters of the LM model without covariates. The estimated class-conditional response probabilities were quite similar to those for the LM model with covariates.

State	State 1	State 2
State 1	0.18	0.82
State 2	0.07	0.93

Table 2. Latent transition probabilities.

The final model that was fitted to these data was another 2 state model with the addition of a covariates effect on the transition probabilities. We were interested in whether the effect of education and age modified the probability of respondents' transitions, i.e. the change the approach to the emigration.

The hypothesis test showed that the separate variables of age and occupation had an influence on the transition probabilities (these covariates are statistically significant).

To interpret the estimated generalized logit coefficients of covariates, we calculated and plotted the transition probabilities at varying levels of age and occupations. In Fig. 1 the estimated transition probabilities are given separately for each age category and level of occupation.

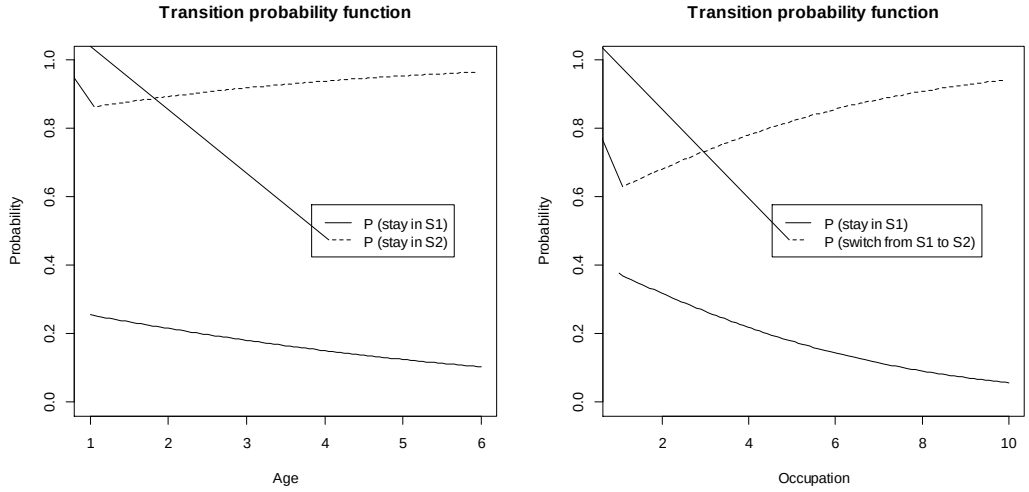

Fig. 1. Predicted transition probabilities for age and occupation covariates.

As expected, the probability of staying in State 1 is decreasing with age while the probability of staying in the second state is increasing with age (see left panel of the Fig. 1).

Due to the space limit, we do not present all the figures, but similarly, the probability of changing the attitude to willingly working abroad (switch from State 2 to State 1) is decreasing with age whereas the probability of switching from State 1 to State 2 is increasing with age. We would support the view that older people would be more critical towards emigration as they might find the adjustment to new country more difficult than younger people.

As far as the second significant covariate is concerned, respondents with higher positions are more likely to stay in the emigration state; on the other hand the lower the position, the higher the probability of switching to State 2 (see right panel of the Fig. 1). However, it is interesting to notice that regardless of a position level, respondents are very likely to belong to State 2 (a little bit higher for people with the lower position level) and to switch to State 1 (a little bit higher for people with higher position level).

Conclusion

We might suppose that emigration was a very common phenomenon because of the EU borders opening or an economic crisis. Despite two decades of uninterrupted growth, however, Polish people are still leaving. We applied a latent Markov model to analyze Polish attitudes to emigration since the EU accession.

We focused especially on the variant of LM model with covariates which additionally allowed us to investigate the dynamic pattern of the Polish attitudes to emigration for different demographic features. By examining the estimated class-conditional response probabilities, we confirmed that the society could be divided into two groups. We found two states of Poles: pro-emigration state (the smaller one) and anti-emigration state (the bigger one). We also showed the influence of covariates on the transition probabilities, representing stability of behaviors. We hope that this small group of people, ready to leave our country will change their mind after the latest statistics were released (the National Crime Agency showed the number of potential victims of trafficking last year increased by 22% on 2012). The highest number of people trafficked into the UK (the most popular country of Polish emigration) came from Romania and most of them were sexually exploited. Poland was the most likely country of origin for people facing labour exploitation.

In future research it would be worthwhile to analyze data presented above using the variant of the Latent Markov model including time-constant and time-varying covariates as well, where both initial-state and transition probabilities are allowed to differ for each latent state (Bartolucci et al., 2013).

References

Agresti, A. (2002). *Categorical data analysis*. New York: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2013). Including individual covariates and relaxing basic model assumptions. In: *Latent Markov models for longitudinal data*. Boca Raton: Chapman and Hall/CRC press.

Bartolucci, F., Montanari, G., & Pandolfi, S. (2015). Three-step estimation of latent Markov models with covariates. *Computational Statistics & Data Analysis, 83*, 287-301.

Baum, L., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics, 41*(1), 164-171.

Collins, L., & Lanza, S. (2010). *Latent class and latent transition analysis: With applications in the social behavioral, and health sciences*. Hoboken, N.J.: Wiley.

Czapiński, J., & Panek, T. (2013). Diagnoza Społeczna - Warunki i jakość życia Polaków. Retrieved from http://www.diagnoza.com.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, 39*, 1-38.

Genge, E. (2014). Zastosowanie ukrytych modeli Markowa w analizie oszczędności wśród Polaków. *Studia Ekonomiczne, Metody Wnioskowania Statystycznego w Badaniach Ekonomicznych, 189*, 58-66.

Paas, L., Vermunt, J., & Bijmolt, T. (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*(4), 955-974.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*, 461-464.

van de Pol, F., & Langeheine, R. (1990). Mixed Markov Latent Class Models. *Sociological Methodology, 20*, 213-247.

Vermunt, J. (1997). *Log-linear models for event histories*. Thousand Oaks, Calif.: Sage Publications.

Vermunt, J., Langeheine, R., & Böckenholt, U. (1999). Discrete-Time Discrete-State Latent Markov Models with Time-Constant and Time-Varying Covariates. *Journal of Educational and Behavioral Statistics, 24*, 179-207.

Visser, I., & Speekenbrink, M. (2010). DepmixS4: An R Package for Hidden Markov Models. *Journal of Statistical Software, 36*(7), 1-21.

Wiggins, L. (1973). *Panel analysis; latent probability models for attitude and behavior processes*. Amsterdam: Elsevier.