Proceedings of the

34th International Workshop on Statistical Modelling

Volume II

July 7-12, 2019

Guimarães, Portugal

Proceedings of the 34th International Workshop on Statistical Modelling
Volume II,
Guimarães, July 7-12, 2019,
Luís Meira-Machado, Gustavo Soutinho (editors),
Guimarães, 2019.

Editors:

Luís Meira-Machado,lmachado@math.uminho.pt
University of Minho
Dep. Mathematics and Applications
4810-058 Azurém - Guimarães
Portugal

Gustavo Soutinho, gustavo.soutinho@ispup.up.pt
EPIUnit, ICBADS, University of Porto
Rua das Taipas 135
4050-600 Porto
Portugal

ISBN 978-989-20-9630-8

A multivariate hidden Markov model: prospects for the course of public trust in Poland

Fulvia Pennoni[1], Ewa Genge[2]

[1] Department of Statistics and Quantitive Methods, University of Milano-Bicocca, Milano, Italy
[2] Department of Economic and Financial Analysis, University of Economics, Katowice, Poland

E-mail for correspondence: ewa.genge@ue.katowice.pl

Abstract: We propose to analyse longitudinal survey data by using survey weights and missing responses via a two step procedure to estimate the parameters of the Hidden Markov model with covariates affecting the latent process. The joint estimated posterior probabilities are employed to make predictions on the latent trajectories of the course of public trust of the Polish society.

Keywords: Panel data; Missing responses; Survey weights; Predictions

1 Introduction

A multidimensional phenomena like public trust is object of many sociological studies since it is a human attitude connecting the individual dimension to culture and society. It is frequently described as an invisible "institution" that lies between governors and governed and it is a typical latent concept since it is not directly measurable. Individual responses to items of survey questionnaires are generally employed to assess levels of public trust among the society.
We propose a multivariate Hidden Markov Model (HMM, Bartolucci *et al.*, 2013) able to account for the repeated responses over time along with longitudinal survey weights and missing responses. We model the response category of "no opinion" among "yes" and "no" in order to include the absence of expression or indecision and provide a multidimensional picture of this phenomena.
We aim to identify similar typology of individuals sharing common perceptions towards public and financial institutions and to explore how these perceptions are evolving over time. We explain the resulting variability according to the available time-varying socio-economic features of the respondents. We show the proposal

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), University of Minho, Portugal, 7-12 July 2019.

by analysing data arising from the Social Diagnosis surveys carried out in Poland from 2009 to 2015 (Social Diagnosis, 2015). This study is relevant since Poland is one of East-Central European countries with citizenships showing the lowest level of public trust according to the recent Eurobarometer survey (Eurobarometer 81, 2014).

2 Proposed method

Let $\mathbf{Y}_{it}$ be the observed response vector for individual i, $i = 1, \ldots, n$, at each time occasion t, $t = 1, \ldots, T$ and let Y_{ijt} be the single response variable provided to item $j, j = 1, \ldots, r$ by individual i, $i = 1, \ldots, n$ at time occasion t. A time-varying latent trait denoted as $\mathbf{U} = (U_1, \ldots, U_T)$ represents trust and it is assumed as a hidden stochastic process of first-order having a discrete distribution with k support points. We assume local independence between responses: for each individual i at time occasion t the responses collected in the vector $\mathbf{Y}_{it}$ are conditionally independent given the latent variable U_{it} and we assume that $\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{iT}$ are independent one another conditionally to the latent process $\mathbf{U}_i$, so that they remain only marginally dependent.

We propose a two step procedure to estimate the model parameters by accounting for the missing responses. First, we fitted a basic HMM by considering the sampling weights of all respondents in order to estimate the parameters of the measurement model. Then, by fixing these parameters we fit a HMM with survey weights and covariates influencing the initial hidden states of the Markov chain as well as the transition probabilities. Time-varying covariates are denoted by $\boldsymbol{X}_t$, $t = 1, \ldots, T$. At the second step, the parameters of the latent model that are the initial $\pi_{u|\boldsymbol{x}}$ and the transition probabilities $\pi_{u|\bar{u},\boldsymbol{x}}$ conditional to the covariates are parameterized using the following multinomial model:

$$\log \frac{\pi_{u|\boldsymbol{x}}}{\pi_{1|\boldsymbol{x}}} = \beta_{0u} + \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_{1u}, \quad u = 2, \ldots, k, \tag{1}$$

$$\log \frac{\pi_{u|\bar{u}\boldsymbol{x}}}{\pi_{\bar{u}|\bar{u}\boldsymbol{x}}} = \delta_{\bar{u}u} + \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\delta}_{1u\bar{u}}, \ \bar{u} \neq u, \tag{2}$$

for $t \geq 2$, and $\bar{u}, u = 1, \ldots, k$, where and $\boldsymbol{\delta}_{11} = \mathbf{0}$ to ensure model identifiability, $\boldsymbol{\beta}_{1u}^{\mathrm{T}}$ and $\boldsymbol{\delta}_{1\bar{u}u}^{\mathrm{T}}$ define the influence of the covariates.

The Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977) represents the main tool to estimate the HMM models. It is based on the *complete data likelihood* that for the proposed model is given by

$$\begin{aligned} \ell_1^*(\boldsymbol{\theta}) \quad = \quad & \sum_{i=1}^{n} \Bigg[\sum_{u=1}^{k} \sum_{j=1}^{r} \sum_{t=1}^{T} \sum_{y=0}^{2} w_i a_{iujty} \log \phi_{jy|u} + \sum_{u=1}^{k} \sum_{t=1}^{T} w_i b_{iu1} \log \pi_{u|\boldsymbol{x}} \\ & + \sum_{\bar{u}=1}^{k} \sum_{u=1}^{k} \sum_{t=2}^{T} w_i b_{i\bar{u}ut} \log \pi_{u|\bar{u},\boldsymbol{x}} \Bigg], \end{aligned} \tag{3}$$

where $\boldsymbol{\theta}$ is the vector of all free parameters, w_i denotes the survey weight for individual i, a_{ujty} corresponds to the (weighted) frequency of people responding

to the j-th item and belonging to latent state u at occasion t; $\phi_{jy|u}$ is the conditional probability of the response y given the latent state u, b_{iu1} is an indicator variable equal to 1 if individual i belongs to latent state u at the beginning of the period, and $b_{i\bar{u},t} = b_{i\bar{u},t-1}b_{iut}$ is an indicator variable equal to 1 if the same respondent moves from state $\bar{u}$ to state u at occasion t.
Since the latent configuration is not known for each respondent, the EM algorithm maximizes the *observed data log-likelihood* $\ell(\boldsymbol{\theta})$ by alternating two steps until convergence:

- **E-step**: compute the posterior expected value of the frequencies and indicator variables in equation (3) by suitable forward-backward recursions so as to obtain the expected value of $\ell_1^*(\boldsymbol{\theta})$;
- **M-step**: update $\boldsymbol{\theta}$ by maximizing the value obtained at the E-step.

The HMM model needs to be estimated several times by considering both deterministic and random starting values for the EM algorithm since the log-likelihood function may be multi-modal. It is important to explore the entire parameter's space for each model with a different number of hidden states. In order to select the suitable model we use the Bayesian Information Criterion (Schwarz, 1978) as well as parsimony and interpretability criteria. Standard errors for the parameters are computed according to the observed or expected information matrix at the maximum likelihood estimate. The allocation of each individual to each latent state is based on the maximum a-posteriori probability and it is performed by using the Viterbi algorithm (Viterbi, 1967). Suitable `R` code and functions to estimate the model parameters are adapted from the `R` package `LMest` (Bartolucci et al., 2017) and are available from the authors upon request.

3 Results

At the first step, the results of the HMM estimated without covariates lead us to choose a HMM with $k = 4$ latent states showing a maximum log-likelihood equal to $\hat{\ell} = -295,923.7$ with 127 free parameters. On the basis of the estimated probabilities of the manifest model ($\hat{\phi}_{jy|u}$) referred to the joint responses we classify Poles according to four homogenous latent subpopulations: people predominantly discouraged toward all the institutions U_D, people reluctant to express their own opinions U_{Nop}, people showing predominant trust in both public and financial institutions U_T and finally people reporting trust mainly towards selected institutions U_{ST}, such as insurance companies, government, police and social insurance institutions.
At the second step, the results of the HMM with covariates in equations (1) and (2) suggest for example that, at the beginning of the period, the Poles are equally distributed between clusters U_D, U_{Nop} and U_T. After the first occasion, higher-educated Poles show higher probability of supporting all the institutions U_T or of remaining in the cluster of those with selective confidence U_{ST} compared to people with only primary education. Lower educated people show a higher probability to remain in the subpopulation of Poles not supporting the institutions U_D or to stay in the group of those with no opinions U_{Nop} compared to those higher-educated. According to the predictive probabilities people showing predominant trust U_T at the beginning of the period become more and more selective over time by belonging to the cluster of selective trust U_{ST}.

Acknowledgments: Ewa Genge acknowledges the financial support from the grant SONATA 12, UMO-2016/23/D/HS4/00989 of the National Science Centre, Poland; Fulvia Pennoni acknowledges the support from the grant of the Italian project FIRB RBFR12SHVV.

References

Bartolucci, F., Farcomeni, A., and Pennoni F. (2013). *Latent Markov Models for Longitudinal Data*. London: Chapman & Hall.

Bartolucci, F., Pandolfi, S., and Pennoni F. (2017). `LMest`: An `R` Package for Latent Markov Models for Longitudinal Categorical Data. *Journal of Statistical Software*, **81**, 1 – 38.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1 – 38.

Eurobarometer-European Commission (2014). Public Opinion in the European Union. *Report, Standard Eurobarometer Spring 81.*

Social Diagnosis (2015). Objective and subjective quality of live in Poland. Czapinski J., Panek T. (eds.). Warszawa, Social Monitoring Council.

Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, **13**, 260–269.