

A latent Markov model from a new perspective with an application*

Francesco Bartolucci
Department of Economics
University of Perugia (IT)
email: bart@stat.unipg.it

Fulvia Pennoni and Giorgio Vittadini
Department of Statistics and Quantitative Methods
University of Milano-Bicocca (IT)
e-mail: fulvia.pennoni@unimib.it, giorgio.vittadini@unimib.it

1 Abstract

We propose the use of the latent Markov model in a context of the estimation of multiple causal effects when dealing with observational studies and there are unobserved baseline differences between individuals. The proposed model, tailored for longitudinal data analysis in its basic formulation, has been first introduced by Wiggins (1951) and then formalized in his Ph.D thesis, Wiggins (1955). In Bartolucci et al. (2013) several extensions of the first basic formulation are given and new models have been proposed. The fact that the assumptions encoded by the model may be represented with the help of a path diagram contributes to make such class of models a powerful tool for the analysis of statistical data. In fact, as stated in Pennoni (2014), such models may be seen as built on the foundation of graphical causal models first proposed by Wright (1921) in genetics. Many statistical models tailored for the estimation of the causal effects have been proposed from that period. The potential outcome framework resulted to be one of the most useful tool. However, in a longitudinal setting the latter it is not still well developed as well as for some powerful models developed in the econometric context, see also Romeo (2014).

Building on the foundation of the above models and on a recent proposal of Lanza et al. (2013), we introduce a new use of the propensity score weighting (Rosenbaum and Rubin, 1983) when dealing with a multivariate responses observed at multiple time occasions. We show some assumptions which have to be sustainable for the use of the proposed approach in the context of study. The use of the latent Markov model helps to get a reliable estimate of the average causal effect. An interesting feature of the proposed approach is its flexibility given by the adopted parameterization which allows us to deal with any kind of response variable. The model is fitted by a maximum likelihood estimation procedure based on first estimating a multinomial logit model for the probability of taking each type of treatment given suitably chosen pretreatment covariates. Then, a weighted log-likelihood of the LM model, with weights computed on the basis of the estimates computed at the previous step, is maximized so as to obtain final parameter estimates. This second step relies on the EM algorithm (Baum et al., 1970; Dempster et al., 1977) and reliable standard errors for the model parameters are obtained by using a nonparametric bootstrap method (Davison and Hinkley, 1997).

The proposed application is particularly suitable to show the model formulation as that it concerns the evaluation of human capital development which is related to a critical period of

*Presented at the meeting of the FIRB (“Futuro in ricerca” 2012) project “Mixture and latent variable models for causal-inference and analysis of socio-economic data”, Roma (IT), January 01-23, 2015

university-to-work transition during the first years of the economic crisis. The human capital as stated in Harpan and Draghici (2014) is defined by generic knowledge and skill gained by working experiences and education. By applying the proposed model to a regionally representative longitudinal dataset regarding graduated cohort in the labour market, we show some results which are useful for policy makers such as how the attaining different types of academic degree affects the work path.

References

- Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman and Hall/CRC Boca Raton.
- Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Applications*. Cambridge University Press, Cambridge, MA.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.
- Harpan, I. and Draghici, A. (2014). Debate on the multilevel model of the human capital measurement. *Procedia - Social and Behavioral Sciences*, 124:170–177. Challenges and Innovations in Management and Leadership 12th International Symposium in Management.
- Lanza, S. T., Coffman, D. L., and Xu, S. (2013). Causal inference in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 20:361–383.
- Pennoni, F. (2014). *Issues on the estimation of latent variable and latent class models*. Scholars' Press, Saarbücken.
- Romeo, F. (2014). *Evaluation of graduates first coherent job on labour market history*. Scholars' Press, Saarbücken.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.
- Wiggins, L. (1951). The application of latent structural analysis to the problem of reliability. In *The use of mathematical models in the measurements of attitudes*. Santa Monica: The RAND corporation.
- Wiggins, L. (1955). Mathematical models for the analysis of multi-wave panels. In University, C., editor, *Ph.D. Dissertation*, Ann Arbor. University microfilms.
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20:557–58.