THREE-STEP RECTANGULAR LATENT MARKOV MODELING BASED ON ML CORRECTION

Rosa Fabbricatore [1], Roberto Di Mari[2], Zsuzsa Bakk[3], Mark de Rooij[3] and Francesco Palumbo[4]

[1] Department of Social Sciences, University of Naples Federico II, (e-mail: rosa.fabbricatore@unina.it)

[2] Department of Economics and Business, University of Catania, (e-mail: roberto.dimari@unict.it)

[3] Department of Methodology and Statistics, Leiden University, (e-mail: z.bakk@fsw.leidenuniv.nl, rooijm@fsw.leidenuniv.nl)

[4] Department of Political Sciences, University of Naples Federico II, (e-mail: fpalumbo@unina.it)

ABSTRACT: Rectangular latent Markov (LM) models have been recently introduced to account for different numbers of latent states over time. This contribution proposes a three-step estimation procedure for such models, which proved useful in the LM modeling framework for flexibility. Specifically, a bias-adjusted maximum likelihood (ML) estimator is introduced for the third step. A simulation study provided preliminary encouraging results regarding the efficacy and effectiveness of the method.

KEYWORDS: Rectangular LM models, three-step estimation, ML-based correction

1 Introduction

Latent Markov (LM) models represent a primary reference to study change over time in the framework of non-parametric latent variable models (Bartolucci *et al.*, 2014). Given a set of response variables repeatedly measured at different time points, LM models allow analyzing individuals' transitions across latent states over time, assuming a first-order Markov chain for the latent process. Three types of parameters characterize LM models: *initial state probabilities*, namely state proportion at the first time point; *transition probabilities*, describing the transition from one state to another at each subsequent time point; *class-conditional parameters* accounting for the relation between latent states and observed indicators. Moreover, the effect of individual covariates on initial and transition probabilities can be considered.

One-step and multi-step approaches have been proposed for model parameter estimation. Due to their flexibility and high feasibility, step-wise ap-

proaches are usually preferred in practice. Among them, a bias-adjusted three-step approach exploiting a maximum likelihood-based (ML) correction was proposed (Di Mari *et al.*, 2016).

Rectangular LM models have been recently introduced to address the issue of possible different numbers of latent states for the considered time points (Anderson *et al.*, 2019). Indeed, over time the nature and number of latent classes tend to vary; therefore, a unique overall definition of the latent classes, as typical in classical LM models, might result in either too restrictive or redundant. Currently, only a one-step estimation procedure for this model has been proposed (Anderson *et al.*, 2019). In this vein, the present contribution aims to further generalize the bias-adjusted three-step approach based on ML correction to the case of LM models with rectangular transition matrices.

The following section outlines the proposed three-step approach. Section 3 presents the simulation study carried out to obtain a preliminary evaluation of the developed estimator. Section 4 reports some conclusions.

2 Three-step rectangular LM modeling

Let $\mathbf{Y}_s^{(t)} = (Y_{s1}^{(t)}, \ldots, Y_{sK_t}^{(t)})'$ be the vector of responses for individual $s = 1, \ldots, N$ on the K_t indicators measured at time point $t = 1, \ldots, T$, with a realization $\mathbf{y}_s^{(t)}$. It is worth noting that the set and the number of indicators K_t varies over time. Denote with $X_s^{(t)}$ the categorical latent variable at time t taking value $i = 1, \ldots, I_t$, producing rectangular transition matrices wherever $I_{t-1} \neq I_t$.

In Step 1, the measurement part of the model is estimated for each time point exploiting a latent class model. This step connects the latent states $i = 1, \ldots, I_t$ to the response variables $\mathbf{Y}_s^{(t)}$, providing for each individual s and time t, the posterior class probability $P(X_s^{(t)} = i | \mathbf{Y}_s^{(t)} = \mathbf{y}_s^{(t)})$. In Step 2, state membership $W_s^{(t)}$ is obtained according to the modal assignment rule, namely allocating individuals in the class for which they present the largest posterior probability. Accordingly, the classification error probabilities included in the time-specific $\mathbf{D}^{(t)}$ matrix are defined as the conditional probability of the estimated class value conditional on the true one $P(W_s^{(t)} = g | X_s^{(t)} = i)$, with $g, i = 1, \ldots, I_t$. In Step 3, a rectangular LM model is estimated with the vector of class assignments $\mathbf{W}_s = (W_s^{(1)}, \ldots, W_s^{(T)})$ as single indicators and known error probabilities included in the $\mathbf{D}^{(t)}$ matrices. Keeping out of consideration the effect of covariates, the third-step log-likelihood is $\ell(\eta) = \sum_{s=1}^{N} \log\{P(\mathbf{W}_s)\}$, where η is the vector of free model parameters. The probability $P(\mathbf{W}_s)$ can be

expressed for rectangular transition matrices as

$$P(\mathbf{W}_s) = \sum_{i^{(1)}=1}^{I_1} \sum_{i^{(2)}=1}^{I_2} \cdots \sum_{i^{(T)}=1}^{I_T} P\big(X_s^{(1)} = i^{(1)}\big) \prod_{t=2}^{T} P\big(X_s^{(t)} = i^{(t)} | X_s^{(t-1)} = i^{(t-1)}\big) \prod_{t=1}^{T} P\big(W_s^{(t)} = g^{(t)} | X_s^{(t)} = i^{(t)}\big),$$

where the state-dependent distributions (given by classification errors) are considered fixed parameters, and thus they are not estimated.

A generalization of the Baum–Welch algorithm (Rabiner, 1989) for rectangular LM, which exploits forward and backward probabilities during estimation, was implemented in the R statistical software. The proposed estimator allows for both time-varying and time-invariant measurement models.

3 Simulation study for the developed bias-adjusted estimator

A simulation study was carried out to evaluate the performance of the bias-adjusted maximum likelihood estimator. Different scenarios were considered, mainly concerning class separation and sample size. In particular, three simple latent class models (one per time point) with 3-3-2 latent classes were considered for the measurement part of the model. Class separation was modeled through response probabilities to ten dichotomously-scored items, setting a probability of 0.8 and 0.9 for the most likely responses in the case of moderate and large class separation, respectively. Four sample sizes were considered: 200, 500, 2000, and 10000 observations. Finally, equal size was imposed for initial probabilities and persistent Markov chains for transition probabilities. For each condition, 500 replications were carried out. The bias in the model parameters estimates (initial and transition probabilities) was used to compare the estimator's performance under different conditions.

The results support the overall good performance of the proposed third-step bias-adjusted estimator. The data log-likelihood increases monotonically according to the number of iterations and the algorithm reaches convergence within 20 iterations. The variability of the estimated bias distribution for both initial and transition probabilities becomes smaller as class separation and sample size increase. Figure 1 shows an example of the estimated bias for the transition matrix from Time 2 to Time 3, with $\gamma_{tij} = log\frac{P(X_s^{(t)}=i|X_s^{(t-1)}=j)}{P(X_s^{(t)}=j|X_s^{(t-1)}=j)}$. Note that more accuracy for initial probabilities estimates emerged, which reported an average bias close to 0 in all the considered conditions. Conversely, as the

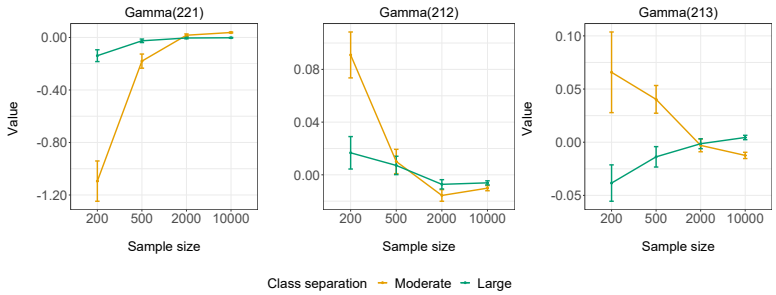

Figure 1. *Mean and standard error of transition probabilities bias (T2 to T3).*

figure shows, a small sample size ($n = 200$) strictly affects transition probabilities estimation due to the presence of very small probabilities in the transition matrix cells that can easily end up in an estimate close to the boundary. Of course, this rarely happens with large samples.

4 Conclusion

A bias-adjusted three-step rectangular LM modeling approach was proposed. In particular, a new estimator for an ML-based correction was developed for the third step. The proposed estimator proved to perform well asymptotically, with a larger estimation bias for small samples and lower class separation. Current developments aim at also considering the covariates' effect on initial and transition probabilities. Empirical applications could provide further insights into the practical advantages of the proposed method.

References

ANDERSON, G., FARCOMENI, A., PITTAU, M. G., & ZELLI, R. 2019. Rectangular latent Markov models for time-specific clustering, with an analysis of the well-being of nations. *J Roy Stat Soc C-App*, **68**, 603–621.

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2014. Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *Test*, **23**, 433–465.

DI MARI, R., OBERSKI, D. L., & VERMUNT, J. K. 2016. Bias-adjusted three-step latent Markov modeling with covariates. *Struct Equ Modeling*, **23**(5), 649–660.

RABINER, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc of the IEEE*, **77**(2), 257–286.