University of Toronto
Department of Economics

Working Paper 589

Rectangular latent Markov models for time-specific clustering

By Gordon Anderson, Alessio Farcomeni, Grazia Pittau and Roberto Zelli

September 22, 2017

Rectangular latent Markov models for time-specific clustering

Gordon Anderson[*], Alessio Farcomeni[†],
Maria Grazia Pittau and Roberto Zelli[‡]

Abstract

A latent Markov model admitting variation in the number of latent states at each time period is introduced. The model facilitates subjects switching latent states at each time period according to an inhomogeneous first-order Markov process, wherein transition matrices are generally rectangular. As a consequence, latent groups can merge, split, or be re-arranged. An application analyzing the progress of well-being of nations, as measured by the three dimensions of the Human Development Index over the last 25 years illustrates the approach.
Key words: group merging, group splitting, Human Development Index, latent transitions.

1 Introduction

Latent Markov (LM) panel data models can be seen as mixed models in which random effects are discrete and evolve over time according to a latent Markov chain (Zucchini and MacDonald, 2009; Bartolucci *et al.*, 2013). Alternatively, they can be seen as (regression) mixtures in which there are k unknown latent intercepts/support points and weights evolve over time. Discrete and time-varying random effects can be used to flexibly take into account unobserved heterogeneity without using normality and time-trend assumptions; and/or they can be used to obtain a (possibly covariate-adjusted) time-specific clustering with respect to the outcome values. A particularly appealing feature of clustering longitudinal data via latent Markov models is that subjects at each time occasion can move to another cluster, which is particularly realistic in several applications. On these points see also Maruotti (2011), Farcomeni (2015) and the discussion of Bartolucci *et al.* (2014).

[*]Department of Economics, University of Toronto, email: anderson@chass.utoronto.ca
[†]Department of Public Health and Infectious Diseases, Sapienza University of Rome, email: alessio.farcomeni@uniroma1.it
[‡]Department of Statistical Sciences, Sapienza University of Rome, email: grazia.pittau@uniroma1.it , roberto.zelli@uniroma1.it .

This work focuses on the clustering properties of latent Markov models, where a proportion of subjects is assigned to one of the k latent states. Classification is usually performed by thresholding the posterior probabilities that a subject is in a certain latent state at a certain time. The main contribution given is that of addressing the limitation of having a fixed number of groups over time. The new model class definition involves a time-varying number of latent states k_t, $t = 1, \ldots, T$; where T is the number of measurement occasions, so that transition matrices are rectangular whenever k_t changes in subsequent occasions. For instance, when $k_{t-1} = 4$ and $k_t = 2$, groups are re-arranged according to a 4 x 2 transition matrix of the kind

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \\ \pi_{31} & \pi_{32} \\ \pi_{41} & \pi_{42} \end{bmatrix}$$

where, e.g., a proportion π_{31} of subjects in group 3 at time $t-1$ moves to the new group 1 at time t. Similarly, when when $k_{t-1} = 2$ and $k_t = 4$, groups are re-arranged according to a 2 x 4 transition matrix of the kind

$$\begin{bmatrix} \pi_{11} & \pi_{12} & \pi_{13} & \pi_{14} \\ \pi_{21} & \pi_{22} & \pi_{23} & \pi_{24} \end{bmatrix}$$

where, e.g., a proportion π_{13} of subjects in group 1 at time $t-1$ moves to the new group 3 at time t. It is important for the sake of interpretation to underline that even groups with the same label have different meaning when k_t changes; as the latent centroid will be different for instance. Accordingly, the proportion π_{11} of subjects in group 1 at time $t-1$ should not be interpreted as staying in group 1, but to move to the new group 1.

The substantive motivation lies on the issue of classification of countries according to their level of development or well-being over time. It is clear that latent Markov models are well-suited for this analysis, as a country might for instance foresee worsening conditions (e.g., due to a war) and move to a club of lower development in consequence; or more hopefully move to a club characterized by better conditions. An open and important issue (Quah, 1997; Durlauf *et al.*, 2005; Anderson *et al.*, 2016) is that of club convergence/divergence, that is, whether the number of clusters is constant, increasing or decreasing over time. This has been debated in the literature, but so far no completely formal approach has appeared to investigate the issue. To this end, we will focus on the three components (gross national income, education, life expectancy at birth) of the Human Development Index (UNDP, 1990) measured over the past 25 years at nation level for 164 nations.

This work also tackles the obvious issue of choosing, among the several possibilities, the configuration for the number of latent classes. In fact, while in classical latent Markov models one usually considers only few possibilities (e.g., $k = 1, 2, 3, 4, 5$); the same rationale would lead to a very large number of models even for moderate values of T (e.g., if five options are allowed at each time occasion, there are 5^T possibilities). This renders common information-based comparison criteria like AIC or BIC infeasible.

A penalized likelihood form is proposed where the observed likelihood is penalized by the entropy of the latent distribution. A similar penalty has been seen to work very well with latent class models (where no transitions are allowed), see for instance Chamroukhi (2016) and references therein. Maximization of the penalized likelihood in the latent Markov framework is far from being straightforward and a naive EM algorithm would basically require enumeration of the model space, similarly to the case of AIC and BIC. In order to overcome this issue, a novel Expectation-Maximization-Markov-Metropolis (EMMM) algorithm for efficient optimization of the penalized likelihood is then proposed. A simple constraint can be used to optimize the penalized likelihood also in the classical case $k_1 = k_2 = \cdots = k_T = k$. In this sense, this paper provides for the first time an automatic and computationally feasible procedure for choosing the number of latent classes in latent Markov models without having to enumerate the model class.

In the rest of the paper the next section briefly reviews latent Markov models. In Section 3 rectangular latent Markov models, where the number of clusters is time-varying, are introduced and inference is discussed. In Section 4, choice of the optimal configuration of cluster numbers by optimization of a penalized likelihood is discussed and the EMMM algorithm introduced. Section 5 reports a simulation study illustrating the performance of the EMMM algorithm for choosing the true configuration of latent states, and the MSE in comparison with more classical model specifications. In Section 6 nations' wellbeing data is analyzed and some concluding remarks are offered in Section 7.

2 Setup

Let y_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T_i$ denote an r-dimensional vector of continuous outcomes measured on the i-th subject at time t; with $T = \max_i T_i$. Let also U_{it} denote an *unobserved* discrete random variable with support $1, \ldots, k$. The basic LM model can be specified as

$$y_{it}|U_{it} = j \sim MVN_r(\xi_j, \Sigma_j),$$

where $\xi_j \in \mathcal{R}^r$ and Σ_j is a positive definite covariance matrix. The model is completed by assumptions of local independence, that is, that conditionally on U_{it} the outcome $\underline{y}_{it}$ is independent of the past measures; and on the distribution of the latent variable U_{it}. Commonly a first-order homogeneous Markov chain is specified, with $\Pr(U_{i1} = j) = \pi_j$ and $\Pr(U_{it} = j|U_{i,t-1} = h) = \pi_{hj}$. The transition probabilities are collected in a *square* transition matrix Π.

The number of parameters involved is $(k-1) + k(k-1) + 2r + r(r-1)/2$.

Several extensions, generalizations and special cases are available, including of course use of covariates (Bartolucci, 2006; Bartolucci and Farcomeni, 2009; Bartolucci *et al.*, 2013), additional categorical or continuous random effects (Altman, 2007; Maruotti, 2011), etc. Inference is carried out either by numerical maximization of the likelihood computed through a forward recursion (Turner, 2008; MacDonald, 2014) or via an Expectation-Maximization (EM) algorithm which also uses a backward recursion.

3 Rectangular LM models

The assumption that k is fixed over time might be restrictive in some applications. A rectangular LM model is obtained through inhomogeneous first-order Markov chains whose number of states is not time-fixed. Formally, U_{it} can be assumed to have support k_t, for $t = 1, \ldots, T$. Configuration-specific initial and transition probabilities can be specified by assuming $\Pr(U_{i1} = j|k_1) = \pi_{jk_1}$ and $\Pr(U_{it} = j|U_{i,t-1} = h, k_{t-1}, k_t) = \pi_{hjk_{t-1}k_t}$, where $\sum_{j=1}^{k_1} \pi_{jk_1} = 1$ and $\sum_{j=1}^{k_t} \pi_{hjk_{t-1}k_t} = 1$. In words, a proportion π_{jk_1} of subjects is assigned to the j-th group (out of k_1) at time $t = 1$. At time $t = 2$ a proportion $\pi_{hjk_1k_2}$ of subjects in group h at time 1 is assigned to group j, regardless of whether $k_1 = k_2$ or $k_1 \neq k_2$. Whenever $k_{t-1} \neq k_t$, a *rectangular* transition matrix is obtained, where subjects are re-arranged into a new grouping configuration. Each subject in group h at time $t-1$ can be assigned to any of the new groups according to the transition probabilities. Marginal probabilities are easily obtained, for instance $\Pr(U_{i2} = j) = \sum_h \pi_{hk_1}\pi_{hjk_1k_2}$.

The model is completed by specification of configuration-specific mean vectors and covariance matrices as

$$y_{it}|U_{it} = j, k_t \sim MVN_r(\xi_{jk_t}, \Sigma_{jk_t}).$$

In words, when there are k_t groups there also are k_t centroids (and scatter matrices). Note that all these parameters are not time-dependent. For the sake of identifiability we assume that the first dimension of ξ_{jk_t} is increasing in j.

The number of parameters involved depends on the number of different values for the number of latent states. Let $v \leq T$, denote the unique values in the set $\{k_1, \ldots, k_T\}$. Let also the couples $(g_1, g_2), (g_3, g_4), \ldots, (g_{l-1}, g_l)$ denote the unique consecutive couples for the number of groups from time $t-1$ to time t. For example, if $k_t = k$ constantly, a unique consecutive couple (k, k) is obtained. If for a certain time t we have possibly a unitary increase for the number of groups, three couples are obtained: (k, k) (for the first $t-1$ occasions), $(k, k+1)$ (for the t-th) and $(k+1, k+1)$ (for the remaining). The number of parameters is then $(k_1 - 1) + \sum_{j=1}^{l/2} g_{2j-1}(g_{2j} - 1) + v(2r + r(r-1)/2)$.

Of course, several extensions are possible simply by making assumptions or opportune parameterizations. For instance one could parameterize

$$\xi_{jk_t} = \delta_{jk_t} + \beta' X_{it},$$

in order to take into account vectors of covariates X_{it} (which might include time). Additionally, the number of parameters can be reduced by making assumptions of Σ_{jk_t} (e.g., that it is diagonal or at least that its off-diagonal elements are equal).

3.1 Inference

The complete log-likelihood can be written as

$$\begin{aligned}\ell^*(\theta) &= \sum_c \sum_i w_{i1ck_1} \log[\pi_{ck_1}] + \\ &+ \sum_k \sum_l \sum_c \sum_d z_{cdlk} \log(\pi_{cdlk}) + \\ &+ \sum_i \sum_t \sum_c w_{itck_t} \log[p(y_{it}|\xi_{ck_t}, \Sigma_{ck_t})]\end{aligned}$$

where w_{itck_t} is a dummy variable equal to 1 if subject i is in latent state c at occasion t and there are k_t groups at that time; $z_{cdlk} = \sum_i \sum_{t>1} w_{i,t-1,ck_{t-1}} w_{itdk_t} I(k_{t-1} = l, k_t = k)$.

The usual expectation-maximization (EM) algorithm for LM models is directly generalized to rectangular LM models. The E-step amounts to performing a forward and a backward recursion to obtain posterior expectations of w_{itck_t} and z_{icdlk}, and the observed likelihood. The latter quantities are iteratively updated after an M-step in which parameters are obtained by maximization of the expected complete log-likelihood.

Formally, let

$$\alpha_{i1}(c) = \pi_{ck_1} p(y_{i1}|\xi_{ck_1}, \Sigma_{ck_1}),$$

and (if $T_i > 1$), for $t = 2, \ldots, T_i$

$$\alpha_{it}(c) = p(y_{it}|\xi_{ck_t}, \Sigma_{ck_t}) \sum_{h=1}^{k_{t-1}} \alpha_{i,t-1}(h) \pi_{hck_{t-1}k_t}.$$

Then, it is straighforward to check that the observed log-likelihood corresponds to $l(\theta) = \sum_{i=1}^n \sum_{c=1}^{k_{T_i}} \alpha_{iT_i}(c)$. Additionally, let $\beta_{iT_i}(c) = 1$ and (if $T_i > 1$), for $t = T_i - 1, \ldots, 1$

$$\beta_{it}(c) = \sum_{h=1}^{k_{t+1}} p(\underset{\sim}{y}_{i,t+1}|\xi_{hk_{t+1}}, \Sigma_{hk_{t+1}}) \beta_{t+1}(h) \pi_{chk_tk_{t+1}}.$$

After the backward recursion one can proceed with the E step by setting

$$E[w_{itck_t}] \propto \alpha_{it}(c)\beta_{it}(c),$$

$$E[z_{ichlk}] = \pi_{chlk} \sum_{t=1}^{T_i-1} I(k_t = l, k_{t+1} = k) \alpha_{it}(c) \beta_{i,t+1}(h) p(\underset{\sim}{y}_{i,t+1}|\xi_{hk_{t+1}}, \Sigma_{hk_{t+1}}) / \sum_c \alpha_{it}(c)\beta_{it}(c)$$

At the M-step, closed form expressions are available for updating model parameters. Formally,

$$\hat{\pi}_{ck_1} \propto \sum_{i=1}^n E[w_{i1ck_1}]$$

$$\hat{\pi}_{cdlk} \propto \sum_{i=1}^{n} \sum_{t=1}^{T_i} E[z_{icdlk}]$$

$$\hat{\xi}_{ck} = \frac{\sum_{it} E[w_{itck}] I(k_t = k) y_{it}}{\sum_{it} E[w_{itck}] I(k_t = k)}$$

$$\hat{\Sigma}_{ck} = \frac{\sum_{it} E[w_{itck}] I(k_t = k)(y_{it} - \hat{\xi}_{ck})(y_{it} - \hat{\xi}_{ck})'}{\sum_{it} E[w_{itck}] I(k_t = k)}.$$

The E and M step are iterated until convergence. As usual, a multi-start strategy shall be used to increase the likelihood of finding the global optimum.

4 Choice of the number of latent components

A simple possibility for the choice of the configuration for the number of latent components $k_1, \ldots, k_T$ is to pre-specify several configurations of interest and compare them through the usual information criteria, like AIC or BIC.

A more automatic and unified approach could be given by the maximization of a penalized likelihood form, similarly to the approaches of Figueiredo and Jain (2000), Yang *et al.* (2012), Chamroukhi (2016) for static mixtures (e.g., latent class models). These approaches are based on the principle that one can set $k_1 = k_2 = \cdot = k_t = k_{\max}$ for some large value of $k_{\max}$, and then penalize the observed likelihood so that some clusters are empty at convergence. In this section we exploit this idea for obtaining automatic and optimal simultaneous model choice and estimation procedures for three latent Markov model specifications.

The general principle is as follows: suppose that at each time point there are at least $k_{\min,t}$ latent states, where usually one can assume $k_{\min,t} = 1 \; \forall t$ and at most $k_{\max,t}$, where usually one can assume $k_{\max,t} = k_{\max} \; \forall t$. Let also p_{ctk} denote the proportion of subjects in cluster c at time t when there are k_t latent states at that occasion. Here $p_{c1} = \pi_{ck_1}$ and $p_{ct} = \pi_{k_1} \prod_{h=1}^{t-1} \Pi_{k_h,k_{h+1}}$, where with a slight abuse of notation the $\prod$ sign indicates a matrix product. Let also n_t denote the number of subjects with measurements available at time t (i.e., $T_i \geq t$). A sensible penalty is given by the total entropy for U_{it}, which is given by $-n_t \sum_c p_{ct} \log(p_{ct})$: the entropy induces more unbalanced groups, eventually emptying the least important ones. Consequently, one could maximize

$$l(\theta) + \lambda \sum_{t=1}^{T} n_t \sum_{c=1}^{k_t} p_{ct} \log(p_{ct}) \tag{1}$$

in θ and $k_{\min,t} \leq k_t \leq k_{\max,t}$ for a given penalty parameter λ. When $\lambda = 0$ exactly $k_{\max,t}$ groups are obtained at each time point, while as λ is increased groups are more and more unbalanced, eventually with $k_t = k_{\min,t} \; \forall t$ for $\lambda > \lambda_u$. For values of $\lambda \in (0, \lambda_u)$ the optimum of (1) corresponds to a situation where not all $k_t = k_{\min,t}$ and not all $k_t = k_{\max,t}$.

Finally, note that for the first time, to the best of the authors knowledge, it is possible to obtain an automatic selection procedure for the number of latent states

in classical (squared) latent Markov models simply by assuming $k_1 = k_2 = \cdots = k_T$. It is also straightforward to modify the procedure for *time-inhomogeneous* square or rectangular transition matrices, even with the presence of covariates for the latent distribution.

In order to maximize (1) it is possible to proceed by adapting the general framework of Chakraborty and Chaudhury (2008), which was proposed for the completely different context of robust estimation. The optimization procedure we propose is a generalized EM algorithm which we name EMMM algorithm. At the first iteration we start from a current solution $\hat{\theta}$, $\hat{k}_1, \ldots, \hat{k}_T$ and perform an EM update, where the E-step is as above and the M-step is slightly modified to take into account the penalty when estimating the latent distribution. Namely, as in Chamroukhi (2016) the complete likelihood is augmented with the penalty term. For simplicity we update the initial and transition distributions by numerical optimization of the expected complete likelihood. Then, a Markov-Metropolis (MM) algorithm is run by building a sequence of random proposals $k(1), \ldots, k(B)$, where $k(0) = \{k_1, \ldots, k_T\}$ is the current configuration. At each step, $k(b+1)$ is obtained by picking the t-th time-occasion with probability $1/T$, and then setting $k_t = k_t + 1$ with probability $1/2$ and $k_t = k_t - 1$ otherwise. If $k_t = k_{\min,t}$ then $k_t = k_t + 1$ with probability 1 and similarly for the case $k_t = k_{\max,t}$. A number c of *candidate* EM iterations is performed under the candidate configuration $k(b)$. A crucial point in the procedure is how the candidate EM is initialized. In fact, if the initialization is far from the current solution, the candidate solution might be rejected due to lack of convergence to a better one. Ideally, the initial solution should have at least the same value for the objective function as the current solution. This is achieved easily when k_t is increased, as one generates an initial solution by replicating one latent intercept chosen at random and updating the latent probabilities at random. The current value for the observed likelihood is unchanged. When k_t is decreased, two consecutive latent intercepts (chosen at random) are averaged and the latent distribution is updated accordingly. The tuning parameter c can be set as the minimum between $c_{\max}$ and the first exceedance of the current objective function by the candidate solution. Experience suggests as a good choice, used in the examples in this paper, $c_{\max} = 1$. Call $o(k(b))$ the resulting objective function computed after the final candidate M step. The random proposal $k(b)$ is accepted with probability

$$p_b = \min\left(e^{-\frac{\log(b+1)}{D}(o(k(b-1))-o(k(b)))}, 1\right). \tag{2}$$

If the random proposal is not accepted, $k(b) = k(b-1)$. After B iterations of the MM algorithm the new configuration $k(B)$ is retained, together with the (possibly new) current solution for the parameters.

If an implementation of the EMMM algorithm for the case of classical latent Markov models with equal number of latent states at each iteration is desired, then only the procedure of random proposal generation needs to be changed, where $k(b)$ is obtained by adding 1 to all entries of $k(b-1)$ with probability $1/2$, and subtracting 1 otherwise (unless $k_1 = k_2 = \cdots = k_T = k_{\min}$ or $k_1 = k_2 = \cdots = k_T = k_{max}$, of course).

There are two tuning parameters for EMMM algorithm, B and D. The latter is used in (2) to control the acceptance probability. A large D will lead to a greedy

stochastic optimizer, while a small D will allow some candidates leading to smaller likelihoods to be accepted, so to escape local optima for the configuration of latent states. It should be noted that Chakraborty and Chaudhury (2008) algorithm was not conceived for use within an EM algorithm. In order to do so, an increase-control step in which the final configuration $k(B)$ is rejected as soon as $l(k(B))$ is below the current largest observed likelihood must be included. It is generally recommended (Chakraborty and Chaudhury, 2008) that D is set approximately equal to the maximal change in the objective function that can be seen when performing a random update. Few pilot runs can be used to estimate the latter quantity and fine tune D. The other tuning parameter, B, should be set large enough so that the stochastic optimizer is allowed to improve the current k. We have actually found that often even as much as $B = 50$ is large enough, which is not surprising given that the stochastic optimizer is repeated at each iteration. It should be noted that a similar MM within EM algorithm was previously used by Farcomeni and Viviani (2011); Farcomeni (2014a,b) but in completely different contexts.

Finally, note that when $\prod_t (k_{\max,t} - k_{\min,t})$, the number of possible configurations, is low, then instead of an MM step one could perform an exhaustive search over all possible configurations of $k_1, \ldots, k_T$. This can be done by performing a candidate EM step for each configuration and then selecting the one leading to the largest increase in (1).

In order to choose the penalty parameter λ several strategies can be set forth. One possibility as in Chamroukhi (2016) is to devise a data driven expression; another one, as suggested in Dotto *et al.* (2017) in a different context, is to evaluate stability of the results via resampling for several values in a grid. In the current context however data driven approaches might increase the likelihood of incurring in local optima (at least in our dynamic latent state framework), and resampling might become very cumbersome from a computational perspective. The suggestion is to evaluate the likelihood for several values of λ in a grid, and (i) compare the optimal configuration at convergence for each λ, in a table, (ii) plot the penalized likelihood at convergence. Experience suggests that, when groups are somewhat well separated, for $\lambda > 0$ and up until a certain value of λ one will get the same (correct)solution consistently, making a precise choice of λ not crucial. Additionally, the penalized likelihood at convergence should be not increasing in λ, allowing for a heuristic choice looking for an "elbow" in the plot. Another possibility is to compute a standardized difference of penalized likelihoods, that is, called L_j the penalized likelihood corresponding to λ_j and assuming $\lambda_1 = 0 < \lambda_2 < \lambda_3 < \cdots < \lambda_g$, one can compute for $j > 1$

$$\frac{L_j - L_{j-1}}{\bar{L}(\lambda_j - \lambda_{j-1})}, \tag{3}$$

where $\bar{L}$ is the average of $L_1, \ldots, L_g$; and pick λ_j as the minimum value for which the quantity above is below a pre-specified threshold, say 5%. An illustration will be given in Section 6.

5 Simulation studies

The following simulation study was performed in order to assess the ability of the penalized likelihood approach of recovering the true underlying configuration of the number of latent groups; and the MSE performance in general for estimation of the centroids.

Data was generated from our model with $r = 3$ Gaussian outcomes, $T = 6$ occasions, $n = 100, 200$, and four sequences of true configuration (which can be seen in Tables 1 and 2). An initial uniform distribution (with mass $1/k_1$) was fixed and transition matrices with entries proportional to the unity for the probability of transitions to different states, proportional to $8 + k_{t-1}$ for the probability of staying in the same state label (even when the number of groups changes). As a result, for instance, we fix Π_{22} as

$$\begin{bmatrix} 0.909 & 0.091 \\ 0.091 & 0.909 \end{bmatrix}$$

and Π_{23}

$$\begin{bmatrix} 0.833 & 0.083 & 0.083 \\ 0.083 & 0.833 & 0.083. \end{bmatrix}$$

Other transitions used include Π_{33}, Π_{32}, etc. For the centroids we set unit variances; and means for the l-th variable as equally spaced between 1 and sk_t, with two values for s leading to medium/low separation ($s = 2$) and to high separation ($s = 3$).

At each iteration, after generating data as above, four different models were estimated: an oracle rectangular latent Markov model with known configuration $k_1, \ldots, k_6$; a classical latent Markov model with fixed number of groups $\max_t k_t$, a classical latent Markov model with fixed number of groups $\min_t k_t$; and a rectangular latent Markov model with unspecified configuration of latent states and maximization of the penalized likelihood (1). A fixed $\lambda = 0.1$ was used for the latter. Note that the resulting performance of the penalized likelihood approach is therefore conservative, being based on a fixed and non-optimized λ. Data generation and model estimation was repeated $B = 1000$ times for each of the 16 scenarios.

The results are reported in Table 1, where the proportion of times that maximization of the penalized likelihood lead to the correct configuration of latent states is shown; and in Table 2, the average MSE for estimation of the centroid when $k = 2$ and when $k = 3$ for each of the four estimation methods is reported.

First of all, from Table 1 it can be seen that the penalized likelihood approach is able to recover the true underlying configuration of the number of latent states with a satisfactory high probability.

Secondly, from Table 2 it can be seen that when the number of latent groups is not fixed, the naive approach based on usual square latent Markov models leads to increased MSE (mostly due to bias), even when $k = \max_t k_t$. Notably, the penalized approach leads to comparable MSE with respect to the oracle one (and even lower in few scenarios).

Table 1: *Simulation study. Proportion of iterations in which maximization of (1) leads to the correct configuration of latent groups (p_C), for different n, separation (s), configuration (k_t: number of latent states at time t). Results are based on $B = 1000$ replicates.*

n	s	k_1	k_2	k_3	k_4	k_5	k_6	p_C
100	2	2	2	2	3	3	3	.99
100	2	2	3	2	3	2	3	.97
100	2	3	2	3	4	2	2	.95
100	2	3	3	3	3	3	3	1.00
200	2	2	2	2	3	3	3	1.00
200	2	2	3	2	3	2	3	1.00
200	2	3	2	3	4	2	2	.99
200	2	3	3	3	3	3	3	1.00
100	3	2	2	2	3	3	3	1.00
100	3	2	3	2	3	2	3	.99
100	3	3	2	3	4	2	2	.99
100	3	3	3	3	3	3	3	.99
200	3	2	2	2	3	3	3	.99
200	3	2	3	2	3	2	3	1.00
200	3	3	2	3	4	2	2	1.00
200	3	3	3	3	3	3	3	1.00

6 Human Development Index and nation clubs

The common practice of measuring the well-being of a society with purely economic variables, like the GDP per capita, has been recently challenged by alternative multi-dimensional measures that policy makers and international organizations should consider in the design and implementation of policies. The Human Development Index (HDI), proposed by the United Nations Development Programme (UNDP, 1990), is by far the most prominent attempt at expanding the dimensions of well-being. Common concerns with this index are questions regarding aspects of well-being that should be included in the analysis, how they should be aggregated, and how to establish the cut-off points for the categories of the index to classify the nations. The UNDP classifies nations in four categories (low/medium/high/very high development) based on fixed cut-off points derived from the quartiles of the development index distribution. A first problem with this approach is that it prevents analysis of club convergence and mobility between classes. More generally, the specification of class cut-offs is usually arbitrary and has met with criticisms (Anderson *et al.*, 2016).

The HDI extends the simple GDP or GNI per capita as an indicator of well-being by incorporating information on health and education. Currently, the index builds on three indicators (UNDP, 2016): gross national income (GNI) per capita (income index), life expectancy (health index) and a weighted average of expected years of schooling and mean years of schooling (education index). Here the analysis is carried out on a balanced panel of $n = 164$ countries over a period spanning from 1990 to 2014. Data

Table 2: *Simulation study. MSE in estimation of the centroid when $k = 2$ and $k = 3$ (ξ_2, ξ_3) by maximizing (1) (pen), with $k = 2$ at all times (min), with $k = 3$ at all times (max), with known configuration $k_1, \ldots, k_6$ (oracle). Results are shown for different n, separation (s), configuration (k_t: number of latent states at time t), and are based on $B = 1000$ replicates.*

								MSE ξ_2			MSE ξ_3			
n	s	k_1	k_2	k_3	k_4	k_5	k_6	pen	min	oracle	pen	min	max	oracle
100	2	2	2	2	3	3	3	.04	.04	.04	.13		.40	.16
100	2	2	3	2	3	2	3	.04	.03	.04	.20		1.32	.19
100	2	3	2	3	4	2	2	.04	.15	.04	.19			.16
100	2	3	3	3	3	3	3				.05	.05	.05	.05
200	2	2	2	2	3	3	3	.02	.03	.02	.06		.28	.06
200	2	2	3	2	3	2	3	.02	.02	.02	.10		.40	.10
200	2	3	2	3	4	2	2	.02	.14	.02	.13			.08
200	2	3	3	3	3	3	3				.02	.02	.02	.02
100	3	2	2	2	3	3	3	.04	.05	.04	.12		1.14	.20
100	3	2	3	2	3	2	3	.04	.15	.04	.49		4.58	.66
100	3	3	2	3	4	2	2	.04	.16	.04	.16			.14
100	3	3	3	3	3	3	3				.04	.04	.04	.04
200	3	2	2	2	3	3	3	.02	.03	.02	.06		.91	.06
200	3	2	3	2	3	2	3	.02	.14	.02	.08		1.79	.23
200	3	3	2	3	4	2	2	.02	.14	.02	.10			.07
200	3	3	3	3	3	3	3				.02	.02	.02	.02

are taken from the Human Development Reports web-site (`hdr.undp.org/en/data`) and have been collected every five years. Per capita GNI are reported in 2011 purchasing power parity. The expected years of schooling is used for the education index as several mean years of schooling are missing.

In estimation the model has been estimated with penalty for a few values of λ. The final estimated configuration of the number of latent groups for each value of λ is given in Table 3, where the max and min values have been set at 4 and 1 respectively.

It can be seen that for $\lambda = 0$, obviously, $k_t = k_{\max}$; but as soon as $\lambda > 0$ a fairly stable solution based on $k = (4\ 4\ 4\ 3\ 3\ 3)$ is obtained. Of course, for large penalty values a different solution is obtained with smaller numbers of groups, and eventually for $\lambda \to \infty$ the optimal solution is $k_t = 1$. Figure 1 reports the penalized log-likelihood at convergence as a function of λ, where an elbow may be observed at $\lambda = 0.25$. An alternative strategy of choosing λ in a forward manner by computing (3) as discussed above leads to the same conclusions. Note that these results provide evidence of club convergence since the number of clubs is shrinking over time with 4 groups in the period 1990–2000 and 3 groups in the period 2005–2014.

The centroids when $k = 3$ and when $k = 4$, together with the estimated standard deviations, are reported in Table 4. As evident from the table, the groups are well separated in all cases. The four groups are labelled according to the value of their respective means as "low", "middle", "high" and "very high" development.

Figure 1: *HDI data. Penalized loglikelihood at convergence for different values of* λ.

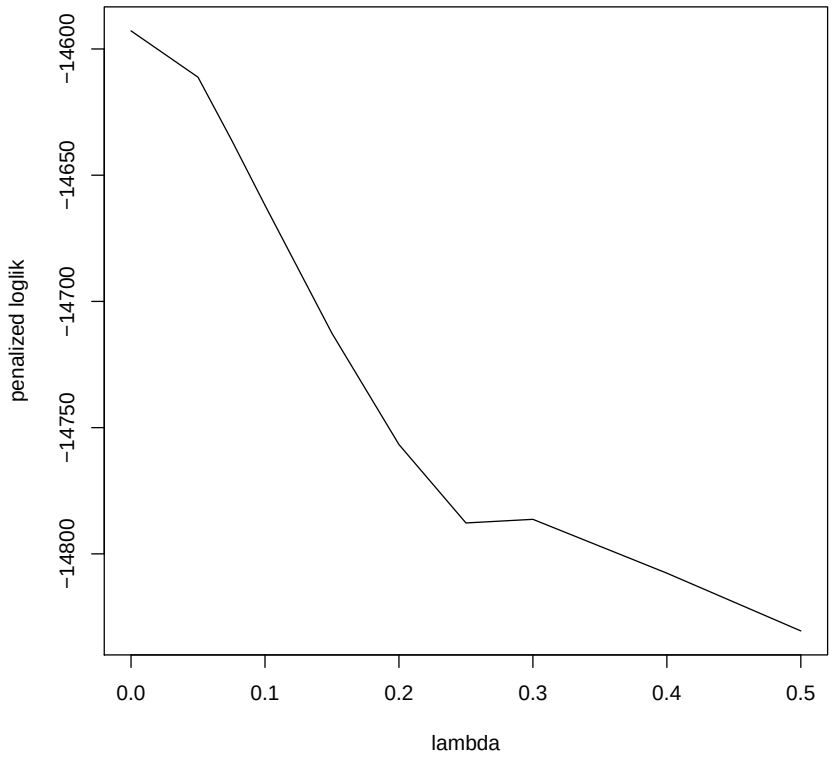

Table 3: *HDI data. Estimated configuration for the number of latent states at convergence of (1) for different values of* λ.

λ	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_3$	$\hat{k}_4$	$\hat{k}_5$	$\hat{k}_6$
0.00	4	4	4	4	4	4
0.02	4	4	4	3	3	3
0.05	4	4	4	3	3	3
0.07	4	4	4	3	3	3
0.10	4	4	4	3	3	3
0.15	4	4	4	3	3	3
0.20	4	4	4	3	3	3
0.25	4	4	4	3	3	3
0.30	4	4	4	3	3	3
0.35	4	4	4	3	3	3
0.40	4	4	4	3	3	3
0.50	4	4	4	3	3	3

Table 4: *HDI data. Estimated centroids when there are three groups (upper panel) and when there are four (lower panel). Here* ξ_1, ξ_2 *and* ξ_3 *refer to the estimated group-specific means for GDP, life expectancy and years of education, respectively; while* σ_1, σ_2 *and* σ_3 *refer to the estimated group-specific standard deviations.*

group	ξ_1	ξ_2	ξ_3	σ_1	σ_2	σ_3
g_1	2358.06	58.37	9.07	1364.44	6.51	1.79
g_2	12065.35	72.20	13.30	6037.87	4.09	1.52
g_3	44873.41	79.69	15.77	19525.94	2.38	1.77
g_1	1658.31	52.20	6.12	760.97	6.53	2.05
g_2	5841.71	66.81	10.61	3240.31	5.45	1.42
g_3	11361.46	70.64	12.49	4536.14	2.01	1.14
g_4	38708.36	76.34	14.03	20352.22	2.42	2.21

The reduction in the number of groups represents a clear signal that a process of convergence of well-being occurs after 2000. To better evaluate this process, we report the estimated transition matrix $\hat{\Pi}_{4,3}$:

$$\begin{bmatrix} 1.000 & 0.000 & 0.000 \\ 0.093 & 0.907 & 0.000 \\ 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 1.000 \end{bmatrix}$$

which implies that the process of convergence involves essentially the two central groups, since $\pi_{2,2} \approx 1$ and $\pi_{3,2} \approx 1$. We could name this new central group as "middle-high". A proportion of 9% of countries belonging to the "middle" group in the first period is "left behind" and moves to the new "low developed" group. The transition matrix reveals no upward mobility from the "low" class as well as no movements in

and out the "very high" developed group, indicating a persistent gap between the two groups. Overall, we there is no indication that low developed countries are catching-up with the "middle-high" and "very high" developed groups.

To corroborate the substantial lack of mobility between groups, the square transitions $\hat{\Pi}_{4,4}$ and $\hat{\Pi}_{3,3}$, which are almost diagonal, are reported:

$$\begin{bmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.998 & 0.002 & 0.000 \\ 0.000 & 0.000 & 0.961 & 0.039 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{bmatrix}$$

and

$$\begin{bmatrix} 0.987 & 0.013 & 0.000 \\ 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 1.000 \end{bmatrix}$$

.

As a last remark, note that this approach offers merely a perspective on the data, and that different results can be obtained considering different pre-processing of the data, or more generally different indicators of well-being altogether. Using a regressive version of the proposed model, in a Bayesian framework, Anderson *et al.* (2017) have argued in favor of a constant 3-clubs world. This in the light of a different pre-processing of the data and the use of population-weights.

7 Conclusions

In line with the discussion of Bartolucci *et al.* (2014), there are two main reasons for using latent Markov models. First, the hidden structure can be used to capture as much unobserved heterogeneity as possible, in order to estimate parameters for the manifest distribution after removing time-varying sources of bias. In this case, a fixed (and slightly overestimated) number of latent states might be a satisfactory choice. Secondly, the hidden structure can be itself of interest in order to cluster units in a time-varying fashion. In this case, a time-varying number of groups (as proposed in this paper for the first time) might be often needed and justifiable in our opinion; and treating the true time-varying configuration k_t in a classical time-fixed framework $k_t = k$ might lead to bias and wrong conclusions. In our experience, it often happens with latent Markov models that one latent group is almost empty at one or more time occasions. A more parsimonious and less biased approach would in all those cases be the use of our rectangular latent Markov models.

Inference can be carried out at the same computational price as usual latent Markov models, and is less biased when in general there are processes of inflation or deflation of the unobserved heterogeneity, as testified by the simulation study. A penalized likelihood form can be used to efficiently estimate the underlying configuration of the number of latent states (or, after assuming that these are not time-varying, of the number of latent states). Notably, a novel algorithm for optimization of a penalized

form of the likelihood has been introduced. Our novel EMMM algorithm has little computational overhead with respect to the EM algorithm used for fixed configuration of latent states, despite being slightly slower due to the use of numerical maximization for update of the initial and transition distributions. Notably, due to the properties of the MM algorithm, conditionally on the other final estimated parameters the estimated configuration of latent states attains the global optimum for the penalized likelihood (Chakraborty and Chaudhury, 2008; Farcomeni, 2014a,b). Use of penalized likelihood forms for latent Markov models is a new area: to the best of our knowledge, it has been considered previously only in Farcomeni (2017) in a completely different context. In our opinion is promising for solving several open issues with latent Markov models, and selection of the number (or configuration) of latent states is one of those as testified by the present work.

This work has focused, as motivated by the real word application, on clustering. As mentioned, there are several extensions possible for rectangular latent Markov models. Most of them are rather straightforward while, for instance, parameterizing the number of clusters as a function of covariates (e.g., time) might be slightly more cumbersome.

References

R. M. Altman (2007). Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, **102**, 201–210.

G. Anderson, A. Farcomeni, M. G. Pittau, and R. Zelli (2017). Nation wellbeing and the HDI, more equality yet less similarity: A multidimensional mixture distribution analysis. *submitted*.

G. Anderson, M.G. Pittau, and R. Zelli (2016). Assessing the convergence and the mobility of nations without artificially specified class boundaries. *Journal of Economic Growth*, **21**, 283–304.

F. Bartolucci (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society, series B*, **68**, 155–178.

F. Bartolucci and A. Farcomeni (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, **104**, 816–831.

F. Bartolucci, A. Farcomeni, and F. Pennoni (2013). *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC Press, Boca Raton, FL.

F. Bartolucci, A. Farcomeni, and F. Pennoni (2014). Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates (with discussion). *TEST*, **23**, 433–486.

B. Chakraborty and P. Chaudhury (2008). On an optimization problem in robust statistics. *Journal of Computational and Graphical Statistics*, **17**, 683–702.

F. Chamroukhi (2016). Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation*, **86**, 2308–2334.

F. Dotto, A. Farcomeni, L. A. Garcia-Escudero, and A. Mayo-Iscar (2017). A fuzzy approach to robust regression clustering. *Advances in Data Analysis and Classification*, to appear.

S. Durlauf, P.A. Johnson, and J. Temple (2005). Growth econometrics. In: P. Aghion and S. Durlauf, eds., *Handbook of Economic Growth*, vol. 1B, Chapter 8. North Holland.

A. Farcomeni (2014a). Robust constrained clustering in presence of entry-wise outliers. *Technometrics*, **56**, 102–111.

A. Farcomeni (2014b). Snipping for robust k-means clustering under component-wise contamination. *Statistics and Computing*, **24**, 909–917.

A. Farcomeni (2015). Generalized linear mixed models based on latent Markov heterogeneity structures. *Scandinavian Journal of Statistics*, **42**, 1127–1135.

A. Farcomeni (2017). Penalized estimation in latent Markov models, with application to monitoring serum Calcium levels in end-stage kidney insufficiency. *Biometrical Journal*, available online.

A. Farcomeni and S. Viviani (2011). Robust estimation for the Cox regression model based on trimming. *Biometrical Journal*, **53**, 956–973.

M. A. T. Figueiredo and A. K. Jain (2000). Unsupervised learning of finite mixture models. *IEEE Transactions Pattern Analysis and Machine Intelligence*, **24**, 381–396.

I. L. MacDonald (2014). Numerical maximisation of likelihood: A neglected alternative to EM. *International Statistical Review*, **82**, 296–308.

A. Maruotti (2011). Mixed hidden markov models for longitudinal data: An overview. *International Statistical Review*, **79**, 427–454.

D. Quah (1997). Empirics for growth and distribution: Polarization, stratification, and convergence clubs. *Journal of Economic Growth*, **2**, 27–59.

R. Turner (2008). Direct maximization of the likelihood of a hidden Markov model. *Computational Statistics and Data Analysis*, **52**, 4147–4160.

United Nations Development Programme UNDP (1990). *Human Development Report 1990*. Oxford University Press.

United Nations Development Programme UNDP (2016). *Human Development Report 2016*. UNDP, New York.

M-S. Yang, C-Y. Lai, and C-Y. Lin (2012). A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognition*, **45**, 3950–3961.

W. Zucchini and I. L. MacDonald (2009). *Hidden Markov Models for time series: an introduction using R*. Springer-Verlag, New York.