UNIVERSITÀ DI PISA

Sant'Anna
Scuola Universitaria Superiore Pisa

Consiglio Nazionale delle Ricerche

Università degli Studi
di Cagliari

Book of Short Papers
SIS 2021

SIS2021
PISA, 21-25 GIUGNO 2021

SIS
Società
Italiana di
Statistica

Editors: Cira Perna, Nicola Salvati and Francesco Schirripa Spagnolo

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891927361

Contents

2.11 New developments in latent variable models

A Hidden Markov Model for Variable Selection with Missing Values

Un Modello Hidden Markov per la Selezione delle Variabili con Valori Mancanti

Fulvia Pennoni, Francesco Bartolucci, and Silvia Pandolfi

Abstract We propose a hidden Markov model for longitudinal multivariate continuous responses, accounting for missing data under the missing at random assumption. Maximum likelihood estimation of this model is carried out through the Expectation-Maximization algorithm. To address the problem of dimensionality reduction, we develop a greedy search algorithm based on the Bayesian Information Criterion. We illustrate the proposal through a dataset collected by the World Bank and UNESCO Institute for Statistics on the basis of which we dynamically cluster countries according to the selected variables observed during the period 2000-2017.
Abstract *Viene proposto un modello hidden Markov per risposte continue multivariate longitudinali e possibili dati mancanti sotto l'assunzione* missing at random. *Il metodo della massima verosimiglianza è utilizzato per la stima dei parametri attraverso l'algoritmo Expectation-Maximization. Si implementa anche un algoritmo per la selezione delle variabili e del modello basato sul Bayesian Information Criterion. La proposta è illustrata tramite dati raccolti dalla Banca Mondiale e dall'Istituto di Statistica dell'UNESCO nel periodo 2000-2017, sulla base dei quali i paesi vengono classificati in modo dinamico considerando le variabili selezionate.*

Key words: development changes, Gaussian distribution, longitudinal data, missing at random assumption

Fulvia Pennoni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca
e-mail: fulvia.pennoni@unimib.it

Francesco Bartolucci, Silvia Pandolfi
Department of Economics, University of Perugia
e-mail: francesco.bartolucci@unipg.it, e-mail: silvia.pandolfi@unipg.it

1 Introduction

We consider hidden (or latent) Markov models (HMMs) for the analysis of time-series and panel data [1, 2]. The main assumption is that the observed data depend on a latent process that follows a first-order Markov chain that may be time homogeneous or heterogeneous. In this way, we can account for the unobserved heterogeneity in a time-varying fashion, and we are able to cluster the units in the panel into homogeneous groups corresponding to comparable unobservable characteristics. Given each latent state, the continuous responses at the same time occasion are assumed to follow a multivariate Gaussian distribution with specific mean vector and variance-covariance matrix. We focus on the problem of non-monotone missing data patterns considering partially or totally missing responses at one or more time occasions, under the *missing at random* (MAR) assumption [3]. Following the idea proposed in [4], we implement a greedy forward-backward procedure based on an approximation of the Bayes factor so as to select the subset of the most useful responses for clustering and simultaneously choose the optimal number of latent states. A modified Expectation-Maximization (EM) algorithm [5] is employed to obtain maximum likelihood estimates of the model parameters.

To illustrate the proposal we consider data derived from the World Bank and UNESCO Institute for Statistics to study countries' economic conditions over the period 2000-2017. We use several variables, including GDP per capita, educational levels, life expectancy at birth, and others related to the human development index proposed by the United Nations Development Programme[1] for measuring the well-being at the country level. The proposed approach allows us to characterize disparities among countries in a dynamic fashion and to evaluate development changes.

In the following section we show the proposed HMM accounting for missing data. In Section 3, we outline the main features of the greedy search algorithm for variable and model selection, and in Section 4, we describe the application.

2 Model Formulation and Estimation

Let $\boldsymbol{Y}_{it} = (Y_{i1t}, \ldots, Y_{irt})'$ denote the vector of r continuous response variables measured at time t, $t = 1, \ldots, T_i$, where T_i denotes the number of occasions of observation for unit i, $i = 1, \ldots, n$. Also, let $\boldsymbol{Y}_i$ be the vector obtained by stacking $\boldsymbol{Y}_{it}$ for $t = 1, \ldots, T_i$. The latent process denoted as $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{iT_i})'$ is assumed to follow a first-order Markov chain with state-space ranging from 1 to k. This process is characterized by the initial probabilities

$$\pi_u = p(U_{i1} = u), \quad u = 1, \ldots, k,$$

and the transition probabilities

[1] Data are available at https://datacatalog.worldbank.org/dataset/world-development-indicators

$$\pi_{u|\bar{u}} = p(U_{it} = u | U_{i,t-1} = \bar{u}), \quad t = 2, \dots, T_i, \quad u, \bar{u} = 1, \dots, k,$$

where u denotes a realization of U_{it} and $\bar{u}$ a realization of $U_{i,t-1}$. Under the *local independence assumption*, the response vectors $\boldsymbol{Y}_{it}$ collected in $\boldsymbol{Y}_i$ are conditionally independent given the latent process $\boldsymbol{U}_i$. A conditional multivariate Gaussian distribution is assumed for the responses:

$$\boldsymbol{Y}_{it} | U_{it} = u \sim N(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u),$$

where $\boldsymbol{\mu}_u$ and $\boldsymbol{\Sigma}_u$ are latent state specific mean vectors and variance-covariance matrices. These matrices are constrained to be equal each other when homoscedasticity is assumed, as is usually done in the HMM and finite mixture context [6].

In presence of partially incomplete data, the response variables may be partitioned as $(\boldsymbol{Y}_{it}^o, \boldsymbol{Y}_{it}^m)'$, where $\boldsymbol{Y}_{it}^o$ corresponds to the observed variables and $\boldsymbol{Y}_{it}^m$ corresponds to the missing ones. Accordingly, the conditional mean vectors and variance-covariance matrices may be decomposed as follows

$$\boldsymbol{\mu}_u = \begin{pmatrix} \boldsymbol{\mu}_u^o \\ \boldsymbol{\mu}_u^m \end{pmatrix}, \quad \boldsymbol{\Sigma}_u = \begin{pmatrix} \boldsymbol{\Sigma}_u^{oo} & \boldsymbol{\Sigma}_u^{om} \\ \boldsymbol{\Sigma}_u^{mo} & \boldsymbol{\Sigma}_u^{mm} \end{pmatrix},$$

where the single blocks are identified by letters o and m when referred to observed and missing components, respectively.

Likelihood based inference with missing data is performed under the MAR assumption and independence between sample units. The log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_{it}^o) = \sum_{i=1}^{n} \log \sum_{\boldsymbol{u}_i} \left(\prod_{t=1}^{T_i} f(\boldsymbol{y}_{it}^o | u_{it}) \right) \left(\pi_{u_{i1}} \prod_{t=2}^{T_i} \pi_{u_{it}|u_{i,t-1}} \right),$$

where $\boldsymbol{\theta}$ is the vector of all the model parameters, $f(\boldsymbol{y}_{it}^o)$ is the manifest distribution of the observed responses $\boldsymbol{y}_{it}^o$, and $\boldsymbol{u}_i = (u_{i1}, \dots, u_{iT_i})'$.

The EM algorithm maximizes the above likelihood by alternating two steps until convergence. In particular, at the *E-step* we compute the posterior expected value of the *complete data log-likelihood*, $\ell^*(\boldsymbol{\theta})$, given the observed data and the current value of the parameters. With missing data, this step includes the computation of $\mathrm{E}(\boldsymbol{Y}_{it} \mid \boldsymbol{y}_{it}^o, u)$ and $\mathrm{E}\left[(\boldsymbol{Y}_{it} - \boldsymbol{\mu}_u)(\boldsymbol{Y}_{it} - \boldsymbol{\mu}_u)' \mid \boldsymbol{y}_{it}^o, u\right]$. At the *M-step* we update the estimate of $\boldsymbol{\theta}$ by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$ obtained at the *E-step*. We combine deterministic and random initializations of the EM algorithm to limit the problem of multimodality of the log-likelihood function.

3 Variable and Model Selection

We implement an algorithm to perform variable and model selection in line with the proposal in [4], which is based on assessing the importance of each variable, among those available, by comparing two suitably chosen models. In the first of these models, the candidate variable is assumed to provide additional information about clustering allocation beyond that contained in the already selected variables; in the second model, this variable is not used for clustering. The two models are compared through the Bayesian Information Criterion (BIC) [7], which is related to the Bayes factor and is based on the following index

$$BIC_k = -2\hat{\ell}_k + \log(n)\#par,$$

where $\hat{\ell}_k$ denotes the maximum of the log-likelihood of the HMM with k states and $\#par$ denotes the number of free parameters.

We propose a greedy forward-backward procedure that starts with an initial set of clustering variables, denoted by $\mathscr{Y}^{(0)}$, and a number of latent states, denoted by $k^{(0)}$. At the h-th iteration, the algorithm performs the following three steps:

- *Inclusion step*: each variable j in the remaining set of variables, is singly proposed for inclusion in $\mathscr{Y}^{(h)}$. The variable to be included is selected on the basis of the following difference between *BIC* indexes:

  $$BIC_{diff} = BIC_{k^{(h-1)}}(\mathscr{Y}^{(h-1)} \cup j) - \left[BIC_{k^{(h-1)}}(\mathscr{Y}^{(h-1)}) + BIC_{reg}(j \sim \mathscr{Y}^{(h-1)})\right],$$

  where BIC_k is the index computed under the proposed HMM with k states, and BIC_{reg} is the index related to the multivariate linear regression of the candidate variable on the currently selected set of variables. The variable with the smallest negative BIC_{diff} is included in $\mathscr{Y}^{(h-1)}$, and this set is updated.
- *Exclusion step*: each variable j in $\mathscr{Y}^{(h)}$ is singly proposed for the exclusion on the basis of the following index:

  $$BIC_{diff} = BIC_{k^{(h)}}(\mathscr{Y}^{(h)}) - \left[BIC_{k^{(h)}}(\mathscr{Y}^{(h)} \setminus j) + BIC_{reg}(j \sim \mathscr{Y}^{(h)} \setminus j)\right].$$

  The variable with the highest positive value of the BIC_{diff} is removed from $\mathscr{Y}^{(h)}$.
- *Model selection*: the current value of $k^{(h-1)}$ is updated by minimizing the BIC_k index of the HMM for the current set of clustering variables $\mathscr{Y}^{(h)}$ over k, from $(k^{(h-1)} - 1)$ to $(k^{(h-1)} + 1)$, so as to obtain the new value of $k^{(h)}$.

The algorithm ends when no variable is added to or is removed from $\mathscr{Y}^{(h)}$. It is worth mentioning that the proposed approach may be influenced by the choice of the initial set of responses, therefore some preliminary or sensitivity analyses at this aim are needed.

Once the variables and the number of states have been selected, the EM algorithm directly provides the estimated posterior probabilities of U_{it} used to obtain a prediction of the latent states of each unit i at every time occasion t. The code implemented to perform the estimation and the selection of the proposed HMM is developed by extending the functions included in the `R` package `LMest` [8].

4 Application

Data referred to $n = 217$ countries followed for $T = 18$ years over a set of $r = 25$ responses with missing values are used to illustrate the proposal, which is based on a model assuming a constant variance-covariance matrix across latent states. The greedy search algorithm is applied starting from a model with only one response variable and $k = 6$ latent states chosen on the basis of a preliminary analysis. In the end, this algorithm leads us to choose a model including $r = 15$ responses with $k = 9$ latent states and heterogeneous transition probabilities.

The selected responses are reported in Table 1 along with the estimated cluster conditional means. The latent states are ordered according to increasing values of the estimated means of the variables highlighted in bold and are able to discriminate between countries with different income levels. The estimated parameters of the latent model are reported in Table 2. We notice that the first group of countries (about 11% in 2000) is characterized mainly by low values of GDP, current health expenditure, and school enrollment in tertiary education. However, we estimate that in 2017 the 43% of countries moves to the 3rd cluster referred to countries having especially a higher coverage of social safety net programs in the poorest quintile. The 5th group of countries (about 13% in 2000) shows intermediate levels of development with a remarkable high rate of primary school enrollment. For these countries, we observe a probability of around 0.03 of moving towards the 6th state in 2017.

Table 1 Estimated conditional means of the HMM with $k = 9$ latent states (in bold variables with increasing means across states).

	1	2	3	4	5	6	7	8	9
Ele	13.25	14.45	34.64	54.54	77.15	96.84	99.64	100.00	99.99
GDP	1457.75	1717.51	3228.88	5689.88	6456.77	9816.22	25361.15	41879.28	79947.84
Hea	68.78	106.66	152.39	242.18	313.16	545.37	1493.95	3801.95	2673.04
Lex	52.72	57.01	59.59	60.55	67.64	72.12	76.00	80.58	78.92
Sav	10.72	8.01	19.76	21.84	21.44	21.81	20.51	24.74	42.97
Imp	34.29	51.55	44.93	39.58	51.10	46.64	56.47	38.80	96.42
Sch3	2.89	4.09	8.03	9.45	20.59	33.01	56.35	70.06	32.32
Rese	0.14	0.13	0.36	0.35	0.30	0.39	0.70	2.43	0.61
Trade	58.95	78.29	79.55	72.77	87.08	83.32	110.38	81.14	217.73
Edu	2.91	4.81	3.84	4.77	4.57	4.52	4.88	5.72	3.00
Sch1	71.69	117.63	98.24	95.58	107.49	105.20	101.75	102.28	102.14
Int	1.11	3.49	6.18	9.39	11.69	22.39	52.43	73.82	61.19
Sch2	19.16	31.39	44.12	44.93	73.25	83.69	97.30	112.36	94.59
Safe	7.72	19.99	22.50	20.25	58.25	51.51	68.94	24.08	29.87
Lit	36.21	65.86	59.02	63.89	82.11	91.60	95.19	83.60	96.64

Note: Ele: access to electricity; GDP: gross domestic product per capita; Hea: current health expenditure; Lex: life expectancy at birth; Sav: gross savings; Imp: import of goods, and services; Sch3: school enrollment, tertiary; Rese: research and development expenditure; Trade: exports and imports of goods, and services; Edu: government expenditure on education; Sch1: school enrollment, primary; Int: individuals using the Internet; Sch2: school enrollment, secondary; Safe: coverage of social safety net programs in poorest quintile; Lit: literacy rate.

Table 2 Estimated averaged initial and transition probabilities for the HMM with $k = 9$ states referred to the period 2016-2017.

	1	2	3	4	5	6	7	8	9
$\hat{\pi}_u$	0.11	0.06	0.06	0.05	0.13	0.34	0.11	0.08	0.06
$\hat{\pi}_{u\|1}$	0.57	0.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\pi}_{u\|2}$	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\pi}_{u\|3}$	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
$\hat{\pi}_{u\|4}$	0.00	0.00	0.00	0.94	0.06	0.00	0.00	0.00	0.00
$\hat{\pi}_{u\|5}$	0.00	0.00	0.00	0.00	0.97	0.03	0.00	0.00	0.00
$\hat{\pi}_{u\|6}$	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
$\hat{\pi}_{u\|7}$	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.02	0.00
$\hat{\pi}_{u\|8}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.03
$\hat{\pi}_{u\|9}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

The 6th group differs from the 5th mainly for higher values of GDP, electricity access, and health expenditure. Most countries (about 34%) are allocated to this cluster in 2000 with a persistence probability of around 1.00. The 8th cluster is that of high-income countries (about 8% in 2000), and we estimate a probability of 0.03 of moving towards the 9th cluster in 2017 that is characterized by the highest average values of GDP, trade, import of goods and services, and literacy rate.

Using *local decoding*, we identify development changes of each country over time. For example, the following countries are allocated in the 1st cluster in 2000: Afghanistan, Angola, Benin, Burkina Faso, Burundi, Central African Republic, Chad, Congo, Democratic Republic of Congo, Ethiopia, Guinea, Guinea-Bissau, Mali, Mauritania, Mozambique, Niger, Papua New Guinea, Sierra Leone, Solomon Islands, Somalia, South Sudan, Tanzania, Zambia. We estimate that only Chad and Niger remain in this cluster at the end of 2017, revealing that their economic and social conditions have not changed over time.

Acknowledgements We thank A. Serafini for the support in the preliminary data analysis.

References

1. Bartolucci, F., Farcomeni, A., Pennoni, F.: *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: Chapman & Hall/CRC press (2013)
2. Zucchini, W., MacDonald, I.L., Langrock, R.: *Hidden Markov Models for Time Series: An Introduction using* R. Boca Raton, FL: CRC press (2017)
3. Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*, 3rd Ed. Wiley Series in Probability and Statistics, Hoboken, NJ: John Wiley Sons (2019)
4. Raftery, A.E., Dean, N.: Variable selection for model-based clustering. *J. Am. Stat. Assoc.*, **101**, 168–178 (2006)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **39**, 1–38 (1977)
6. McLachlan G, Peel D. *Finite Mixture Models*. New York: Wiley (2000)
7. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464 (1978)
8. Bartolucci, F., Pandolfi, S., Pennoni, F.: LMest: An R package for latent Markov models for longitudinal categorical data. *J. Stat. Softw.*, **81**, 1–38 (2017)