Journal of Classification
https://doi.org/10.1007/s00357-023-09457-9

Variable Selection for Hidden Markov Models with Continuous Variables and Missing Data

Fulvia Pennoni[1] · **Francesco Bartolucci**[2] · **Silvia Pandofi**[2]

Accepted: 20 November 2023

Abstract
We propose a variable selection method for multivariate hidden Markov models with continuous responses that are partially or completely missing at a given time occasion. Through this procedure, we achieve a dimensionality reduction by selecting the subset of the most informative responses for clustering individuals and simultaneously choosing the optimal number of these clusters corresponding to latent states. The approach is based on comparing different model specifications in terms of the subset of responses assumed to be dependent on the latent states, and it relies on a greedy search algorithm based on the Bayesian information criterion seen as an approximation of the Bayes factor. A suitable expectation-maximization algorithm is employed to obtain maximum likelihood estimates of the model parameters under the missing-at-random assumption. The proposal is illustrated via Monte Carlo simulation and an application where development indicators collected over eighteen years are selected, and countries are clustered into groups to evaluate their growth over time.

Keywords Expectation-maximization algorithm · Greedy search algorithm · Missing-at-random assumption · Model-based variables selection · Sustainable development

1 Introduction

We propose a variable selection method for a hidden (or latent) Markov (HM) model with continuous responses for the analysis of time series (Zucchini et al., 2016) and longitudinal data (Bartolucci et al., 2013), possibly affected by non-monotone missing data patterns. The HM approach is of particular interest in different fields as it represents a practical model-based dynamic clustering method for identifying latent group structures and individual trajectories (Bouveyron et al., 2019). This is possible because a sequence of discrete latent variables following a Markov process, generally of first order, is assumed to represent the behavior

✉ Fulvia Pennoni
fulvia.pennoni@unimib.it

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20128 Milan, Italy

[2] Department of Economics, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy

Published online: 23 January 2024

of every unit. The Markov chain may be homogeneous or heterogeneous over time, and the states correspond to latent clusters or subpopulations of homogeneous individuals that may be visited at each time occasion (Bartolucci et al., 2022).

In the present paper, we deal with the case of continuous responses assumed to follow a Gaussian distribution in which, as in finite mixture models (Banfield and Raftery, 1993; McLachlan and Peel, 2000), it is natural to assume that the responses at the same time occasion are correlated, according to a specific variance-covariance matrix, even conditionally on the underlying latent state. In this setting, completely missing outcomes may arise, also referred to as intermittent patterns, when individuals do not provide responses at one or more time occasions; moreover, partially missing outcomes at a given time occasion may occur. We account for these two missing patterns under the missing-at-random (MAR) assumption (Rubin, 1976; Little and Rubin, 2020). In particular, we develop a method of dimensionality reduction and variable selection to reduce model complexity. The proposed model-based approach simultaneously performs variable and model selection in order to choose the optimal number of variables and clusters (or states). The proposal represents a relevant advancement in the literature on model-based clustering (Frühwirth-Schnatter, 2011; McNicholas, 2016) as the variable selection approaches currently adopted, such as the wrapped methods, are not tailored to deal with different missing data patterns; see, among others, Gales (1999) and Adams and Beling (2019).

HM models are often employed to analyze complex data sets with a wide range of variables (Celeux and Durand, 2008). One of the benefits of performing the simultaneous selection of the number of states and the informative variables is that of obtaining a more parsimonious model that provides more stable parameter estimates and enhances interpretability, especially for high-dimensional data. This may lead to more meaningful clusters, prevents overfitting, and, more importantly, may reduce the computational burden required for estimation and related inference, once the irrelevant variables are discarded. Additionally, this clustering becomes more accurate. We recall that in this setting the clusters correspond to the latent (or hidden) states.

The proposed variable selection procedure takes inspiration from the previous proposals of Raftery and Dean (2006), Maugis et al. (2009a), and Flynt and Dean (2019), developed for model-based clustering using finite mixture models; see Gormley et al. (2023) for a recent review on these models. In particular, we develop a greedy forward-backward search algorithm aimed at selecting the subset of the most informative variables for clustering by means of a series of inclusion and exclusion steps until a suitable stopping rule is satisfied, also allowing the selection of the optimal number of latent states. The decision on whether to include a candidate variable to the clustering set is based on comparing two models according to the Bayesian Information Criterion (BIC; Schwarz, 1978), and this rule may be seen as an approximation of that based on the Bayes factor (Kass and Raftery, 1995).

Maximum likelihood estimation of the model parameters is carried out by an extended expectation-maximization (EM) algorithm (Dempster et al., 1977) based on suitable recursions, developed in Pandolfi et al. (2023). Once the variables and the number of states have been selected, the EM algorithm directly provides the estimated posterior probabilities, which are employed for dynamic model-based clustering, that is, allocation of the sample units to the latent components at each time occasion. This is performed according to the estimated parameters on the basis of the maximum-a-posteriori (MAP) rule. A sort of multiple imputation is also performed in order to predict the missing responses conditionally or unconditionally to the estimated model components.

In order to illustrate the proposed method, we rely on a series of simulations that allow us to assess the performance of the greedy search procedure under different scenarios. We also show an application to macroeconomic data by considering socioeconomic variables selected among the world development indicators (WDI; The World Bank Group, 2018), which allows us to illustrate the results of our proposal in selecting the most relevant variables to achieve a dynamic clustering of countries. In this way, considering that the progress of a country is a multidimensional phenomenon, we also provide a contribution to the literature of data-driven methodology on how to cluster countries; on this topic see Nielsen (2013) among others.

The functions used to perform the search procedure and estimation of the HM model are implemented by extending the functions of the package `LMest` (Bartolucci et al., 2017) developed for the R environment (R Core Team, 2023), and are available in the GitHub repository at the following link https://github.com/penful/HM_varSel.

The remainder of the paper is organized as follows. Section 2 shows the HM model formulation, and Sect. 3 illustrates the inferential approach outlined for estimating model parameters and some other aspects related to the computation of the standard errors and decoding. Section 4 details the main features of the greedy search algorithm and model selection. Section 5 presents the simulation design and the results. Section 6 illustrates the data used for the applicative example, and reports the results obtained with the proposed method, whereas Sect. 7 provides some final conclusions. The Supplementary Information (SI) reports additional information on the data and results of the application.

2 Hidden Markov Model for Clustering Longitudinal Data

We consider the multivariate longitudinal case in which we observe more response variables at each time occasion. Let $\boldsymbol{Y}_{it} = (Y_{i1t}, \ldots, Y_{irt})'$ denote the vector of r continuous response variables measured at time t, with $t = 1, \ldots, T$, where T denotes the number of time occasions, and $i = 1, \ldots, n$, where n denotes the number of sample units. Also let $\boldsymbol{Y}_i$ be the vector obtained by stacking $\boldsymbol{Y}_{it}$ for $t = 1, \ldots, T$.

The general HM model assumes, for every unit i, the existence of a discrete latent process, denoted by $\boldsymbol{U}_i = (U_{i1}, \ldots, U_{iT})'$, affecting the distribution of the response variables and assumed to follow a first-order Markov chain with state-space $\{1, \ldots, k\}$, where k is the number of latent (or hidden) states. Under the local independence assumption, the response vectors $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT}$ are conditionally independent given the latent process $\boldsymbol{U}_i$.

This model is typically employed in the presence of multivariate categorical data to account for unobserved heterogeneity (Bartolucci et al., 2014). The HM model is made of two components: the *measurement model*, concerning the conditional distribution of the response variables given the latent process, and the *latent model*, concerning the distribution of the latent process. When dealing with continuous outcomes, for the first component we assume a conditional Gaussian distribution, that is,

$$\boldsymbol{Y}_{it} | U_{it} = u \sim N(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}), \quad u = 1, \ldots, k.$$

The parameters of the measurement model are the conditional means $\boldsymbol{\mu}_u$, $u = 1, \ldots, k$, which are state-specific, and the variance-covariance matrix $\boldsymbol{\Sigma}$, which is assumed constant across states under the assumption of homoschedasticity. This assumption may avoid certain estimation problems (see McLachlan and Peel, 2000, with reference to finite mixture models). However, it may be suitably relaxed on the basis of proper matrix decompositions (Banfield and Raftery, 1993; Celeux and Govaert, 1995).

The parameters of the latent model are the initial probabilities

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

and the transition probabilities

$$\pi^{(t)}_{u|\bar{u}} = p(U_{it} = u | U_{i,t-1} = \bar{u}), \quad t = 2, \ldots, T, \ \bar{u}, u = 1, \ldots, k,$$

where u denotes a realization of U_{it} and $\bar{u}$ a realization of $U_{i,t-1}$. Note that the transition probabilities are time heterogeneous, but according to the applicative context, they can be assumed as time-homogeneous in order to make the model more parsimonious; see Pennoni and Bal-Domńska (2022) for an illustration on this approach. On the basis of the above parameters, the distribution of $\boldsymbol{U}_i$ is given by

$$P(\boldsymbol{U}_i = \boldsymbol{u}_i) = \pi_{u_{i1}} \prod_{t=2}^{T} \pi^{(t)}_{u_{it}|u_{i,t-1}},$$

where $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iT})'$. Accordingly, the conditional distribution of the observed responses $\boldsymbol{Y}_i$ given $\boldsymbol{U}_i$ may be expressed as

$$f(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{U}_i = \boldsymbol{u}_i) = \prod_{t=1}^{T} f(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} | U_{it} = u_{it}),$$

where $\boldsymbol{y}_i$ is a realization of $\boldsymbol{Y}_i$ made by the subvectors $\boldsymbol{y}_{it} = (y_{i1t}, \ldots, y_{irt})'$.

With possible missing responses, we partition $\boldsymbol{Y}_{it}$ as $\boldsymbol{Y}_{it} = (\boldsymbol{Y}^o_{it}, \boldsymbol{Y}^m_{it})'$, where $\boldsymbol{Y}^o_{it}$ is the vector of observed responses and $\boldsymbol{Y}^m_{it}$ is the vector corresponding to the missing data. Using a straightforward notation, the conditional mean vectors and variance-covariance matrix may be decomposed as

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

where, for instance, $\boldsymbol{\Sigma}^{om}$ is the block of $\boldsymbol{\Sigma}$ containing the covariances between each observed and missing response. As also illustrated in Pandolfi et al. (2023), we note that

$$\boldsymbol{Y}^o_{it} | U_{it} = u \sim N(\boldsymbol{\mu}^o_u, \boldsymbol{\Sigma}^{oo}), \quad u = 1, \ldots, k. \tag{1}$$

The manifest distribution is expressed with reference to the observed data, that is,

$$f(\boldsymbol{y}^o_i) = \sum_{\boldsymbol{u}_i} \left[\prod_{t=1}^{T} f(\boldsymbol{y}^o_{it} | u_{it}) \right] \left(\pi_{u_{i1}} \prod_{t=2}^{T} \pi^{(t)}_{u_{it}|u_{i,t-1}} \right), \tag{2}$$

where $\boldsymbol{y}^o_{it}$ denotes a realization of $\boldsymbol{Y}^o_{it}$ and $f(\boldsymbol{y}^o_{it}|u_{it})$ is the multivariate Gaussian probability density function corresponding to Eq. 1. As usual, when dealing with HM models, to efficiently compute this distribution, we rely on a forward recursion (Baum et al., 1970; Welch, 2003).

3 Model Inference

In the following, we illustrate the adopted likelihood inferential approach and discuss issues related to the initialization of the estimation algorithm. Then, we describe the strategies employed for the computation of the standard errors, and for the *a posteriori* allocation of the units to the estimated states.

3.1 Maximum Likelihood Estimation with Missing Responses

Assuming independence between units, the log-likelihood referred to the observed data can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i^o).$$

In the above expression, $\boldsymbol{\theta}$ is the vector of all model parameters and $f(\boldsymbol{y}_i^o)$ is the manifest distribution of the observed response variables defined in Eq. 2. In order to estimate these parameters, we maximize $\ell(\boldsymbol{\theta})$ by an EM algorithm (Baum et al., 1970; Dempster et al., 1977), which is based on the *complete-data log-likelihood* expressed as

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\sum_{t=1}^{T} \sum_{u=1}^{k} z_{itu} \log f(\boldsymbol{y}_{it}|u) + \sum_{u=1}^{k} z_{i1u} \log \pi_u + \sum_{t=2}^{T} \sum_{\bar{u}=1}^{k} \sum_{u=1}^{k} z_{it\bar{u}u} \log \pi_{u|\bar{u}}^{(t)} \right].$$

In the previous formulation, $z_{itu} = I(u_{it} = u)$ is an indicator variable equal to 1 if individual i is in latent state u at time t and $z_{it\bar{u}u} = z_{i,t-1,\bar{u}}\, z_{itu}$ is the indicator variable for the transition from state $\bar{u}$ to state u of individual i at time occasion t. Also, note that $\ell^*(\boldsymbol{\theta})$ is the sum of three components that may be maximized separately. Regarding the first component, we have to consider that the model assumptions imply that

$$\log f(\boldsymbol{y}_{it}|u) = -\frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\boldsymbol{y}_{it} - \boldsymbol{\mu}_u)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_{it} - \boldsymbol{\mu}_u)$$

and, therefore, this component simplifies as

$$\begin{aligned} &\sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{u=1}^{k} z_{itu} \log f(\boldsymbol{y}_{it}|u) \\ &\qquad = \sum_{i=1}^{n} \left\{ -\frac{T}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2} \sum_{t=1}^{T} \sum_{u=1}^{k} z_{itu} \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_{it} - \boldsymbol{\mu}_u)(\boldsymbol{y}_{it} - \boldsymbol{\mu}_u)' \right] \right\}. \end{aligned}$$

In practice, the *E-step* of the EM algorithm computes the posterior expected value of the dummy variables involved in $\ell^*(\boldsymbol{\theta})$ given the observed data and the current value of the parameters. In particular, these expected values correspond to the following quantities:

$$\hat{z}_{itu} = P(U_{it} = u|\boldsymbol{y}_i^o), \quad t = 1, \ldots, T, \ u = 1, \ldots, k, \tag{3}$$

$$\hat{z}_{it\bar{u}u} = P(U_{i,t-1} = \bar{u}, U_{it} = u|\boldsymbol{y}_i^o), \quad t = 2, \ldots, T, \ \bar{u}, u = 1, \ldots, k. \tag{4}$$

They are computed by means of forward-backward recursions of Baum et al. (1970) and Welch (2003); for an illustration see Bartolucci et al. (2013, Chapter 3). With missing observations and under the MAR assumption, the E-step also includes the computation of the following expected values

$$\begin{aligned} \mathrm{E}[(\boldsymbol{Y}_{it} - \boldsymbol{\mu}_u)(\boldsymbol{Y}_{it} - \boldsymbol{\mu}_u)'|\boldsymbol{y}_{it}^o, u] &= \mathrm{Var}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o) + \mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u)\mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u)' \\ &\quad - \boldsymbol{\mu}_u \mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u)' - \mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u)\boldsymbol{\mu}_u' + \boldsymbol{\mu}_u \boldsymbol{\mu}_u' \\ &= \mathrm{Var}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o) + [E(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u) - \boldsymbol{\mu}_u][E(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u) - \boldsymbol{\mu}_u]', \end{aligned}$$

where

$$\mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u) = \begin{pmatrix} \boldsymbol{y}_{it}^o \\ \boldsymbol{\mu}_u^m + \boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}(\boldsymbol{y}_{it}^o - \boldsymbol{\mu}_u^o) \end{pmatrix} \tag{5}$$

and

$$\text{Var}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^{o}) = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}^{mm} - \boldsymbol{\Sigma}^{mo}(\boldsymbol{\Sigma}^{oo})^{-1}\boldsymbol{\Sigma}^{om} \end{pmatrix}.$$

At the *M-step* of the EM algorithm, we update the model parameters by considering the following closed form for the means:

$$\boldsymbol{\mu}_u = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{z}_{itu}\text{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^{o}, u)}{\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{z}_{itu}}, \quad u = 1, \ldots, k;$$

moreover, $\boldsymbol{\Sigma}$ is updated as follows:

$$\boldsymbol{\Sigma} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{u=1}^{k}\hat{z}_{itu}\left\{\text{Var}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^{o}) + \left[\text{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^{o}, u) - \boldsymbol{\mu}_u\right]\left[\text{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^{o}, u) - \boldsymbol{\mu}_u\right]'\right\}.$$

Finally, the initial and transition probabilities may be obtained on the basis of the usual formulation, that is,

$$\begin{aligned} \pi_u &= \frac{\sum_{i=1}^{n}\hat{z}_{i1u}}{n}, \quad u = 1, \ldots, k, \\ \pi_{u|\bar{u}}^{(t)} &= \frac{\sum_{i=1}^{n}\hat{z}_{it\bar{u}u}}{\sum_{i=1}^{n}\hat{z}_{i,t-1,\bar{u}}}, \quad t = 2, \ldots, T, \ u, \bar{u} = 1, \ldots, k. \end{aligned}$$

The E- and M-step are repeated until convergence, which is checked on the basis of the relative log-likelihood difference, that is

$$\frac{\ell(\boldsymbol{\theta}^{(s)}) - \ell(\boldsymbol{\theta}^{(s-1)})}{|\ \ell(\boldsymbol{\theta}^{(s)})\ |} \leq \epsilon,$$

where $\boldsymbol{\theta}^{(s)}$ is the vector of parameter estimates obtained at the end of the s-th iteration of the M-step and ϵ is a suitable tolerance level (e.g., 10^{-8}).

A crucial aspect of the estimation process is the initialization of the EM algorithm as the model log-likelihood is typically multimodal. In such a situation, a multi-start strategy, based both on a deterministic and a random starting rule, is required. Overall, for a given k, the inference is based on the solution corresponding to the largest value of the log-likelihood at convergence, which typically corresponds to the global maximum. The parameter estimates obtained in this way are collected in the vector $\hat{\boldsymbol{\theta}}$. To obtain the corresponding standard errors, we can rely on a non-parametric bootstrap procedure (Davison and Hinkley, 1997), which is performed by drawing, with replacement, a number of bootstrap samples from the original one, and estimating the proposed HM model with the selected number of states on these samples. This procedure is easily implemented and it has the advantage of the robustness of the results.

With reference to a given number of states, k, the dynamic assignment of units to the latent states is of particular interest. As usual, the estimation algorithm directly provides the estimated posterior probabilities of U_{it}, as defined in Eqs. 3 and 4. These probabilities can be directly used to perform *local decoding* so as to obtain a prediction of the latent states of every unit i at each time occasion t. In particular, the predicted latent state at occasion t for a sample unit i, denoted by $\hat{u}_{it}$, is obtained as the value of u corresponding to the maximum of $\hat{z}_{itu}$. The entire sequence of predicted latent states resulted by the local decoding is denoted as $\hat{u}_{i1}, \ldots, \hat{u}_{iT}$. The so-called

global decoding, which is based on an adaptation of the Viterbi algorithm (Viterbi, 1967), may also be adopted to obtain the prediction of the latent trajectories of a unit across time, that is, the *a posteriori* most likely sequence of latent states.

Finally, note that it is also possible to perform multiple imputations of the missing responses conditionally or unconditionally to the predicted latent state. In more detail, in the conditional case the predicted value is simply $\hat{\boldsymbol{y}}_{it} = \mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, \hat{u}_{it})$, where $\hat{u}_{it}$ is the predicted latent state and the expected value is defined in Eq. 5. The unconditional prediction of the missing responses is instead obtained as

$$\tilde{\boldsymbol{y}}_{it} = \sum_{u=1}^{k} \hat{z}_{itu} \mathrm{E}(\boldsymbol{Y}_{it}|\boldsymbol{y}_{it}^o, u), \quad i = 1, \ldots, n, \ t = 1, \ldots, T.$$

4 Proposed Algorithm for Variable Selection

In this section, we describe the proposed procedure for variable selection. In particular, we extend the method developed by Raftery and Dean (2006), Fop and Murphy (2018), and Bartolucci et al. (2016) to the case of longitudinal data analyzed by a multivariate Gaussian HM model with missing data of the type formulated above. The proposal relies on a greedy search algorithm based on alternating an inclusion and an exclusion step to select the subset of variables that are useful for clustering, also allowing the simultaneous selection of the optimal number of hidden states. First, the variables are divided into three subsets: (i) the subset of the current clustering variables; (ii) that of the single variable proposed to be included in the previous set; and (iii) the subset with all the remaining variables. The decision to include a candidate variable to the clustering set is based on comparing two models. In the first model, the proposed variable is assumed to provide additional information about the clustering allocation beyond the current set of clustering variables. In the second model, the same variable is considered as not useful for clustering.

4.1 Model Comparison

Starting from an initial set of variables, the proposed method is aimed at finding the set of items that provides the best value of the BIC index (Schwarz, 1978), which may be seen as an approximation of the Bayes factor (Kass and Raftery, 1995). In particular, for a given model, the index may be defined as

$$\mathrm{BIC} = -2\hat{\ell} + \log(nT)\#\mathrm{par}, \tag{6}$$

where $\hat{\ell}$ denotes the maximum log-likelihood of the model of interest and #par denotes the corresponding number of free parameters. The same criterion is used to perform the model selection step as illustrated in the following. Note that, with respect to other implementations of the BIC available in the literature about HM models (for a review see Bacci et al., 2014), we rely on a stronger penalization based on the overall number of observations equal to nT rather than on n only. We prefer this choice that leads to a more parsimonious selection of the relevant variables, providing also better results within the simulation study and easiness of comparisons with alternative procedures.

Let $\mathcal{Y}$ be the set of clustering variables and $\overline{\mathcal{Y}}$ the complement of $\mathcal{Y}$ with respect to the full set, that is, the set of remaining variables. The variable selection procedure is based on finding the set of relevant variables, that is, the set of the most useful variables for clustering, which minimizes the following index

$$\mathrm{BIC}_{\mathrm{tot}}(\mathcal{Y}, k) = \mathrm{BIC}_k(\mathcal{Y}) + \mathrm{BIC}_{\mathrm{reg}}(\overline{\mathcal{Y}} \sim \mathcal{Y}).$$

In the previous expression, $\mathrm{BIC}_k(\cdot)$ is the BIC index computed as in Eq. 6 under the proposed HM model with k latent states, fitted on the set of clustering variables indicated in parentheses,

and $\text{BIC}_{\text{reg}}(\cdot)$ is the BIC index referred to the multivariate linear regression for the irrelevant or noise variables, $\overline{\mathcal{Y}}$, on the set $\mathcal{Y}$, which are assumed to be independent of the cluster memberships and thus not providing additional information about clustering.

In applying this procedure, attention must be paid to the estimation of the above linear regression models when missing data are present in the set of the independent variables. In this case, imputation is necessary and, as discussed in Sect. 3.1, in particular we adopt a sort of multiple imputation based on the posterior expected values in Eq. 5 obtained at convergence of the EM algorithm used to estimate the proposed HM model on the full set of variables. In this preliminary stage, a model selection procedure is also performed, aimed at selecting the proper number of states based on the BIC_k index, for increasing values of k. When missing data are present in the subset of variables considered as dependent variables in the linear regression model, maximum likelihood estimation of model parameters is performed under the MAR assumption, and the corresponding BIC_{reg} is computed.

4.2 Inclusion-Exclusion Algorithm

We propose a greedy inclusion-exclusion procedure that starts with an initial set of clustering variables, denoted by $\mathcal{Y}^{(0)}$, an initial number of latent states, denoted by $k^{(0)}$, and the corresponding $\text{BIC}_{\text{tot}}(\mathcal{Y}^{(0)}, k^{(0)})$ index. At the h-th iteration, the algorithm performs the following two steps:

- *Inclusion step*: each variable j in the set of irrelevant variables, $\overline{\mathcal{Y}}^{(h-1)}$, is singly proposed for inclusion in $\mathcal{Y}^{(h)}$. The variable to be included is selected on the basis of the following difference between BIC_{tot} indices:

$$\text{BIC}_{\text{diff}} = \min_{k_0^{(h-1)} \le k \le k^{(h-1)}+1} \text{BIC}_{\text{tot}}(\mathcal{Y}^{(h-1)} \cup j, k) - \text{BIC}_{\text{tot}}(\mathcal{Y}^{(h-1)}, k^{(h-1)}),$$

  where $k_0^{(h-1)} = \max(1, k^{(h-1)} - 1)$. The variable with the smallest negative BIC_{diff} is included in $\mathcal{Y}^{(h-1)}$, and this set is updated as $\mathcal{Y}^{(h)} = \mathcal{Y}^{(h-1)} \cup j$. If no item yields a negative BIC_{diff}, then we set $\mathcal{Y}^{(h)} = \mathcal{Y}^{(h-1)}$.
- *Exclusion step*: each variable j in $\mathcal{Y}^{(h)}$ is singly proposed for the exclusion on the basis of the following index:

$$\text{BIC}_{\text{diff}} = \min_{k_0^{(h-1)} \le k \le k^{(h-1)}+1} \text{BIC}_{\text{tot}}(\mathcal{Y}^{(h-1)} \setminus j, k) - \text{BIC}_{\text{tot}}(\mathcal{Y}^{(h-1)}, k^{(h-1)}).$$

  The variable with the smallest negative value of the BIC_{diff} is removed from the set of relevant variables $\mathcal{Y}^{(h)}$. If no variable is found with a negative BIC_{diff}, we set $\mathcal{Y}^{(h)} = \mathcal{Y}^{(h-1)}$.

The algorithm ends when no variable is added to or is removed from $\mathcal{Y}^{(h)}$. It is worth mentioning that the proposed approach may be influenced by the choice of the initial set of responses, $\mathcal{Y}^{(0)}$; therefore, some preliminary or sensitivity analyses at this aim are required. In general, starting with the complete set of variables is also possible, so relying on a backward scheme. Otherwise, it is possible to start with an empty set and follow a forward procedure. The latter should be preferred, especially when there are many response variables. In the simulation study presented in the next section and in the applicative example illustrated in Sect. 6, we opt for a forward procedure.

Taking inspiration from Maugis et al. (2009b), we also propose an improved version of the above algorithm in which the variables not included in the HM model are regressed to a subset of those included in this model, and this subset is selected by a suitable forward procedure. Obviously, the resulting algorithm is slower than the initial version but, as we experimented, it leads to a more reasonable choice of the relevant variables. The results of the simulation study and the application reported in the next sections are obtained with this improved version, while those obtained with the initial version are available upon request to the authors.

5 Simulation Study

In the following, we illustrate the simulation design carried out to assess the performance of the proposed approach. Here we aim to evaluate the ability of the greedy search algorithm to simultaneously select the correct set of clustering variables and the right number of latent states under different simulated settings.

5.1 Simulation Design

Within the simulation study, we randomly drew $B = 100$ samples of size $n = 250, 500, 1{,}000$ from an HM model of the type formulated in Sect. 2, with a number of hidden states equal to $k = 2, 3$. We assumed $r = 2, 4$ continuous outcomes and $T = 5, 10$ time occasions. Regarding the measurement model, the following values for the conditional means are considered: $\boldsymbol{\mu}_1 = (0, 0)'$ and $\boldsymbol{\mu}_2 = (4, 0)'$, with $k = 2$ states and $r = 2$ response variables, whereas with $r = 4$ variables we set $\boldsymbol{\mu}_1 = (0, 0, 0, 0)'$ and $\boldsymbol{\mu}_2 = (4, 0, 4, 0)'$. Moreover, with $k = 3$ latent states and $r = 2$ outcomes we set $\boldsymbol{\mu}_1 = (0, 0)'$, $\boldsymbol{\mu}_2 = (4, 0)'$, and $\boldsymbol{\mu}_3 = (4, 2)'$, by duplicating the means of the variables when $r = 4$ as before. When $k = 2$ and $r = 2$, an additional scenario characterized by less separated states is also considered, by letting $\boldsymbol{\mu}_1 = (0, 0)'$ and $\boldsymbol{\mu}_2 = (2, 0)'$. We also assumed a variance-covariance matrix constant across states, with all variances equal to 1 and covariances equal to 0.5. Equally likely states are considered at the first time occasion, that is, $\pi_u = 1/k$, $u = 1, \ldots, k$, whereas time-homogeneous transition probabilities, with persistence probabilities equal to $\pi_{u|\bar{u}} = 0.8$, when $u = \bar{u}$, and off-diagonal probabilities equal to $\pi_{u|\bar{u}} = 0.2/(k - 1)$, when $u \neq \bar{u}$, $u = 1, \ldots, k$, are assumed.

In order to evaluate the performance of the variable selection procedure, a total of $J = 15$ variables are included, with the r clustering variables generated according to the HM model defined above. The $J - r$ irrelevant variables are instead simulated according to a multiple regression model depending on the relevant clustering variables, with the intercept varying on an equally spaced grid of values from -2 to 2, and the regression coefficients fixed to $(-1, 1)$ for all $J - r$ variables when $r = 2$. An alternative setting is also considered, by simulating the $J - r$ noise variables from a standard multivariate Gaussian distribution, so that they are assumed to be independent of the relevant clustering variables.

A varying proportion of intermittent missing responses, $p_{miss} = 0,\ 0.05,\ 0.1,\ 0.25$, is also assumed for the full set of variables. Finally, in order to evaluate the robustness of the method in the presence of mild skewness in the data, an alternative scenario in which the relevant variables are simulated according to a Chi-squared distribution with 10 degrees of freedom is also implemented.

Overall, a total of 194 different scenarios are considered. In the following, we report the results obtained under the most informative scenarios in order to assess the effect of the different design factors on the selection of the proper set of clustering variables and of a valid number of latent states.

5.2 Results

The results reported in the following are based on a series of different scenarios:

- The 1st scenario, considered as a benchmark, is based on $n = 250$ as sample size, $k = 2$ latent states, $r = 2$ clustering variables, and a low proportion of intermittent missing values, that is, $p_{miss} = 0.05$. The irrelevant variables are assumed as independent of the clustering variables so that they are simulated according to a standard Gaussian distribution.
- The 2nd scenario, which is aimed at evaluating the performance of the proposed approach when the sample size increases, is based on $n = 1{,}000$, whereas the other parameters of the simulation study are left unchanged with respect to the 1st scenario.

- The 3rd scenario is aimed at evaluating the effect of an increase in the number of latent states, by assuming $k = 3$, while letting the other parameters as in the benchmark.
- The 4th scenario is aimed at assessing the results when the number of time occasions increases, with $T = 10$.
- The 5th scenario differs from the benchmark with respect to the number of relevant clustering variables, that is, $r = 4$.
- The 6th, 7th, and 8th scenarios evaluate how the presence of missing values affects the variable selection procedure, by letting $p_{miss} = 0$, $p_{miss} = 0.1$, and $p_{miss} = 0.25$, respectively.
- The 9th scenario differs from the benchmark with respect to the assumption about the noise variables, which are assumed to depend on the relevant clustering variables through a regression model.
- The 10th scenario evaluates the effect of less separated states, by assuming $\boldsymbol{\mu}_1 = (0, 0)'$ and $\boldsymbol{\mu}_2 = (2, 0)'$.
- The 11th scenario investigates how the presence of mild skewness in the relevant variables may affect the variable and model selection process.

In Table 1, for each scenario described above, we report the frequency of samples in which the correct variable partition is chosen together with the frequency of samples in which the true number of states is selected. We also report the average computing time (in seconds) required by the greedy algorithm to reach the convergence and the average adjusted Rand index (ARI, Hubert and Arabie, 1985), which is aimed at evaluating the agreement between the estimated and the true latent structure. In particular, this index attains its upper bound equal to 1 when there is a perfect agreement between the true and estimated clustering structure and to 0 otherwise.

From the results, we observe that the true clustering variables are always correctly selected under all scenarios, except for scenario 8, that is, when the proportion of missing responses increases. In such a situation, the relevant variables are selected in just over half of the samples. However, we notice that in all samples the true clustering variables are always selected, in addition to some irrelevant variables. This worsening in the results is likely due to the imputed explanatory variables in the regression model estimated on the noise variables.

Referring to the performance of the simultaneous model selection procedure, we observe that the number of latent states is correctly chosen in all samples, regardless of the simulated scenario, apart from scenario 11, where the clustering variables are generated from a Chi-squared distribution. Here, the true number of states is always overestimated, due to the presence of skewness in

Table 1 Results of the variable selection procedure under different simulated scenarios

Scenarios	Frequency of correct variable partition	Frequency of correct number of states	Average ARI	Average computing time
1 - Benchmark	1.00	1.00	0.933	253.94
2 - $n = 1{,}000$	1.00	1.00	0.931	924.36
3 - $k = 3$	1.00	1.00	0.813	245.53
4 - $T = 10$	1.00	1.00	0.934	503.15
5 - $r = 4$	1.00	1.00	0.993	490.71
6 - $p_{miss} = 0$	1.00	1.00	0.972	249.88
7 - $p_{miss} = 0.1$	1.00	1.00	0.891	283.35
8 - $p_{miss} = 0.25$	0.55	1.00	0.761	622.49
9 - Dependence	1.00	1.00	0.933	349.42
10 - Less separation	1.00	1.00	0.632	575.82
11 - Chi-squared	1.00	0.00	0.672	186.10

the data, leading to selecting $k = 3$ in 3 samples out of 100, $k = 4$ in 73 samples out of 100, and $k = 5$ in 24 samples.

Looking at the results in terms of clustering performance, measured by the average ARI over the simulated samples, it is evident that the proposed approach performs well under all scenarios. In particular, the clustering quality worsens as the number of states increases and with lower separation of the states due to the increasing complexity of the model associated with a higher uncertainty on the latent structure. Moreover, as expected, the average ARI across simulations reduces when there is an increasing proportion of missing responses or when the simulated data deviate from a Gaussian distribution. On the other hand, increasing the number of relevant variables and the number of time occasions allows us to recover in a better manner the latent structure of the model, thanks to a higher amount of available information.

In evaluating the performance of the proposed approach, it is also relevant to take into account the computational cost of the corresponding greedy search algorithm. All scenarios are run on an Apple M1 Pro with 16 Gb for a fair comparison. As mentioned, when considering the improved version of the algorithm that includes an additional forward step, the computing time necessary to reach convergence increases. In particular, it varies across simulated scenarios, on average, from a minimum of about 245 s to a maximum of about 622 s under the most complex scenario characterized by a high proportion of missing variables.

It is also interesting to compare our proposal with a simplified version of the model, which does not take into account the longitudinal structure of the data and avoids missing values. In particular, we first imputed the missing responses before starting the variable selection algorithm by using function `imp.mix` of the R package `mix` (Schafer, 2022). Then, we run the `clustvarsel` algorithm that performs variable selection for Gaussian model-based clustering according to similar greedy search algorithm (Scrucca and Raftery, 2018), following the classical approach of Raftery and Dean (2006). In our longitudinal context, the responses of the same unit to the different time occasions are considered independent, and a finite mixture model of Gaussian distribution with a variance-covariance matrix common to all components is estimated. We observe that, under the first six scenarios reported above, the simplified procedure performs well in selecting the clustering variables and the true number of states but it leads to a reduction of the average ARI. Obviously, the computational time is reduced with respect to our proposal due to the simplified version of the model and the variable selection algorithm. On the other hand, in more complex scenarios, characterized by a large proportion of missing values, less separation of the hidden states, and in the presence of mild skewness in the data, the performance of the `clustvarsel` algorithm gets worse, especially concerning the correct estimation of the number of clusters. Consequently, this approach also attains poor results when considering the clustering quality, substantially reducing the average ARI. Results are available upon request by the authors.

6 Application

In order to illustrate the proposed approach, a set of 25 socioeconomic indicators of the WDI[1] collected by the World Bank[2] and the UNESCO Institute for Statistics[3] is considered. The 25 indicators are listed Table 2, and their detailed definition is provided in Tables 1, 2, 3, 4, 5, and 6 of the SI. They are collected yearly for $n = 217$ countries listed in Table 7 of the SI.

[1] For a detailed description see webpage: https://databank.worldbank.org/data/source/world-development-indicators, Last Updated: 03/21/2019.

[2] See the webpage: https://www.worldbank.org/en/home

[3] See the webpage: http://uis.unesco.org/

Table 2 List of the 25 variables of the World Development Indicators and their acronyms used in the paper

Number	Abbreviation	Indicators
1	Life	*Life expectancy at birth*
2	Pop	*Population ages 0–14*
3	Infa	*Infant mortality rate*
4	Sch1	*School enrollment, primary*
5	Sch2	*School enrollment, secondary*
6	Sch3	*School enrollment, tertiary*
7	Edu	*Government expenditure on education*
8	Gedu	*Gross national expenditure*
9	Rese	*Research and development expenditure*
10	GDP	*GDP per capita*
11	Une	*Unemployment*
12	Gsav	*Gross savings*
13	Ele	*Access to electricity*
14	Int	*Individuals using the Internet*
15	Ren	*Renewable electricity output*
16	Gini	*GINI index*
17	Trade	*Trade*
18	Saf	*Coverage of social safety net programs in poorest quintile*
19	Lit	*Literacy rate*
20	Hea	*Current health expenditure*
21	Hyd	*Electricity production from hydroelectric sources*
22	Imp	*Imports of goods and services*
23	Comb	*Combustible renewables and waste*
24	Lab	*Labor force participation rate*
25	Fert	*Fertility rate*

Figures 1, 2, 3, and 4 of the SI show observed values for each country at every time occasion with missing observations depicted in black. A substantial heterogeneity across countries in all the macroeconomic indicators can be noticed; there are intermittent missing observations for some countries at certain time occasions and complete missing values of some indicators in certain years for all countries, such as for Comb. Moreover, some countries do not provide at all values of Edu. We refer the reader to the SI for more details and descriptive statistics on these data. We applied a logit transformation for the variables expressed in a percentage scale and a Box-Cox transformation to the other variables (Box and Cox, 1964) to improve the applicability of the model to the available dataset, and we suitably re-scaled all variables.

The time heterogeneous HM model illustrated in Sect. 2 is useful for analyzing these data. In particular, we recall that the model is based on a first-order Markov process on which the literature on the analysis of longitudinal data mainly relies essentially for two reasons: (i) it is perfectly justifiable that today's latent variable depends, first of all, on the yesterday's latent variable; (ii) relaxing this assumption would complicate the model considerably. We also must consider that, marginally with respect to the latent variables, the distribution of the outcomes is characterized by a more sophisticated dependence structure that is not simply of the first order.

6.1 Results

In the analysis, using the approach proposed in the paper we selected the best subset of indicators able to highlight the patterns of disparities among countries and to provide a picture of how countries are performing over time. Thus we shed light on the way to classify or cluster countries based on their level of socioeconomic development. As illustrated at the end of Sect. 4.2, the greedy search algorithm is implemented by performing a preliminary inclusion step, in which each variable is singly proposed in turn for inclusion in the initial set. At the same time, for each candidate variable, a model selection step is performed to choose the optimal number of states.

The procedure led us to select a model with $k = 6$ latent states including $r = 12$ indicators, showing a log-likelihood value at convergence equal to -20,998.52 with 665 parameters and a BIC value of 47,496.78. Table 3 reports the estimated cluster conditional means $\boldsymbol{\mu}_u$, $u = 1, \ldots, 6$, referring to the selected set of indicators.

For a more straightforward interpretation, the latent states are increasingly ordered according to the values of the estimated means of the variables Lit and Ele, and decreasingly ordered according to values of Fert. For a visual comprehension of the different features among selected clusters, Fig. 1 shows the resulting heatmap. We note that:

- the 1st and 2nd groups differ from the other groups for high values of Fert, Comb, and Hyd, and they also show high values of Gini. Therefore, natural resources appear to be important as countries in 1st group especially lack all the other indicators. The 1st and 2nd mainly differ from each other by Edu and Sch1. Higher values of these two variables characterize the 2nd group;
- the 2nd and the 3rd groups also differ in Saf, Ele, and Gsav, the 3rd group having higher values of these indicators;
- the 4th group has much higher values of Edu, Hyd, and Lit and lower values of Saf compared to the 3rd;
- the 5th group is characterized by high levels of Gsav, and especially by a lower value of Gini with respect to the 4th, and slightly less for Edu;

Table 3 Estimated conditional means of the selected $r = 12$ indicators (reported in their original scale) under the HM model with $k = 6$ latent states (variables in bold are those with almost increasing or decreasing ordered values)

Indicators	u					
	1	2	3	4	5	6
Sch1	81.292	115.131	108.075	107.964	101.386	101.785
Sch2	4.636	7.758	23.295	33.317	27.883	58.775
Edu	3.392	4.474	4.449	4.760	3.842	4.838
Rese	0.248	0.246	0.226	0.401	0.225	1.263
Gsav	13.491	19.997	23.680	21.466	31.492	22.427
Ele	30.005	37.785	84.145	97.411	99.805	99.954
Gini	59.675	61.050	61.398	61.515	57.244	58.206
Saf	11.976	18.037	69.534	58.655	47.320	46.892
Lit	47.322	70.127	87.715	92.249	95.462	97.296
Hyd	36.803	56.426	23.504	35.465	8.477	18.049
Comb	55.986	51.857	19.101	9.732	0.176	4.881
Fert	5.560	4.196	2.881	2.272	2.591	1.621

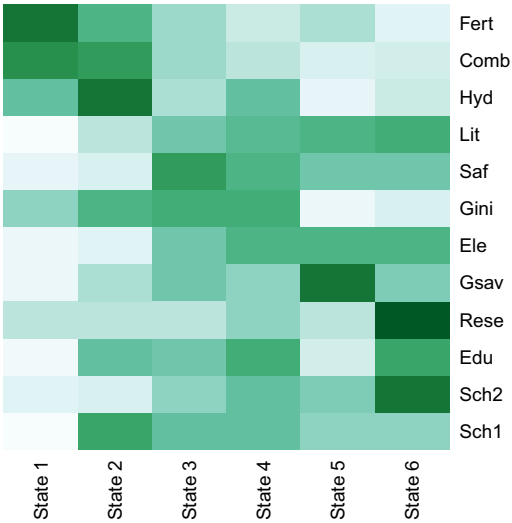


Fig. 1 Heatmap of scaled cluster means under the HM model with $k = 6$ hidden states according to the $r = 12$ selected indicators

– countries in the 6th group present the highest values of Rese, compared to all the other countries, and they show higher values of Edu and Sch2 than countries in the 5th group, and slightly less for Gsav.

Table 4 shows the estimated covariances (lower part), variances (diagonal), and the partial correlations (upper part) among the selected variables. Looking at the partial correlations we observe that Rese and Sch2 have a quite high correlation, whereas Saf has a negative correlation with Rese, and a positive correlation with Gsav; Lit is negatively associated with Fert given all the remaining indicators.

The estimated parameters of the latent model referred to the initial probabilities in 2000 and transition probabilities from 2000 to 2001 are reported in Table 5. The statistical significance of the coefficients is established according to the estimated standard errors obtained by the non-parametric bootstrap procedure described in Sect. 3.1, based on 300 samples. Tables with the estimated standard errors are reported in the SI (see Tables 11 to 14).

At the beginning of the study, some countries belong to the 6th group (about 28%), while around 18% of countries belong to the 1st cluster. From 2000 to 2001, we mainly observed mobility of the countries of the 3rd cluster to the 4th (4.5%) and the 6th (3.2%) and from the 5th to 3rd (2.9%) and 4th (3.1%). Countries in the 1st group have a probability of around 0.01 of moving to the 2nd group. We note a very high persistence probability for the 4th and the 6th groups.

The estimated transition probabilities referred to the period just before the global financial crisis (2005 to 2006) are reported in Table 6. Two upward transitions are visible for countries in clusters 3 and 4: (i) a probability of 0.07 of moving from the 3rd to the 4th state and of 0.05 to the 5th; (ii) a probability of 0.06 of moving from the 4th to the 5th cluster.

Probabilities referred to the transitions in years 2010 to 2011 are reported in Table 7. Three years after the global economic downturn, we notice a probability of: (i) 0.1 of moving from the 2nd to the 3rd cluster; (ii) 0.04 from the 3rd to the 6th cluster; (iii) 0.10 from the 4th to the 6th. Transitions estimated for the last period of observation from 2016 to 2017 are reported in Table 8; 30% of countries in the 4th cluster are moving to the 6th cluster, thus showing a general growth for emerging economies.

Table 4 Estimated covariances (lower part), variances (diagonal in italics), and partial correlations (upper part) under the HM model with $k = 6$ hidden states; figures in bold are the partial correlations greater than |0.3|

	Sch1	Sch2	Edu	Rese	Gsav	Ele	Gini	Saf	Lit	Hyd	Comb	Fert
Sch1	*76.243*	−0.053	−0.068	0.118	−0.002	0.058	0.114	0.067	−0.015	0.135	0.162	−0.009
Sch2	−4.147	*143.050*	0.065	**0.494**	−0.286	0.146	−0.152	0.612	0.052	0.215	−0.145	0.183
Edu	−1.198	3.854	*1.668*	0.283	−0.276	−0.028	−0.062	−0.003	0.284	−0.122	−0.074	0.159
Rese	0.125	2.298	0.224	*0.268*	**0.366**	−0.060	0.027	**−0.436**	−0.054	−0.140	0.087	−0.053
Gsav	−5.123	4.481	−2.484	0.880	*99.472*	0.165	−0.326	**0.425**	−0.010	0.108	−0.150	0.095
Ele	−10.772	14.646	0.918	0.462	24.782	*100.092*	0.005	−0.229	0.223	−0.116	**−0.326**	−0.236
Gini	1.927	−1.404	−0.137	−0.081	−3.405	−2.779	*0.857*	0.245	−0.177	0.386	−0.107	0.164
Saf	−2.183	68.553	−1.657	−1.140	42.813	0.897	−1.127	*192.882*	0.081	**−0.399**	0.219	**−0.324**
Lit	−1.404	14.324	1.928	0.094	3.341	21.655	−1.142	8.846	*40.792*	**0.419**	−0.026	**−0.350**
Hyd	63.619	−20.886	−2.398	−0.974	−65.931	−66.958	10.586	−111.413	28.281	*666.151*	**0.349**	0.093
Comb	29.508	−14.978	−1.881	−0.535	−25.821	−60.208	3.056	−2.402	−8.898	138.450	*137.307*	0.106
Fert	0.640	−0.642	0.050	0.025	−1.667	−2.998	0.218	−3.520	−1.851	5.362	2.523	*0.503*

Table 5 Estimated average initial and transition probabilities under the HM model with $k = 6$ hidden states referred to the period 2001–2002; figures in italics are those in the main diagonal (significant **at 1%, *at 10%)

	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 5$	$u = 6$
$\hat{\pi}_u$	0.181**	0.106**	0.152**	0.114**	0.163**	0.284**
$\hat{\pi}_{u\|1}$	*0.991***	0.009*	0.000	0.000	0.000	0.000
$\hat{\pi}_{u\|2}$	0.040**	*0.960***	0.000	0.000	0.000	0.000
$\hat{\pi}_{u\|3}$	0.000	0.000	*0.923***	0.045	0.000	0.032**
$\hat{\pi}_{u\|4}$	0.000	0.000	0.000	*1.000***	0.000	0.000
$\hat{\pi}_{u\|5}$	0.000	0.000	0.029	0.031**	*0.939***	0.000
$\hat{\pi}_{u\|6}$	0.000	0.000	0.000	0.000	0.000	*1.000***

Table 6 Estimated average transition probabilities under the HM model with $k = 6$ hidden states referred to the period 2005–2006; figures in italics are those in the main diagonal (significant at **at 1%)

	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 5$	$u = 6$
$\hat{\pi}_{u\|1}$	*1.000***	0.000	0.000	0.000	0.000	0.000
$\hat{\pi}_{u\|2}$	0.000	*1.000***	0.000	0.000	0.000	0.000
$\hat{\pi}_{u\|3}$	0.000	0.000	*0.885***	0.069**	0.046**	0.000
$\hat{\pi}_{u\|4}$	0.000	0.000	0.000	*0.940***	0.059	0.000
$\hat{\pi}_{u\|5}$	0.000	0.000	0.000	0.000	*1.000***	0.000
$\hat{\pi}_{u\|6}$	0.000	0.000	0.000	0.000	0.000	*1.000***

Table 7 Estimated average transition probabilities under the HM model with $k = 6$ hidden states referred to the period 2010–2011; figures in italics are those in the main diagonal (significant at **at 1%, *at 10%)

	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 5$	$u = 6$
$\hat{\pi}_{u\|1}$	*1.000***	0.000	0.000	0.000	0.000	0.000
$\hat{\pi}_{u\|2}$	0.032	*0.873***	0.096*	0.000	0.000	0.000
$\hat{\pi}_{u\|3}$	0.000	0.000	*0.915***	0.048**	0.000	0.037**
$\hat{\pi}_{u\|4}$	0.000	0.000	0.000	*0.902***	0.000	0.098**
$\hat{\pi}_{u\|5}$	0.000	0.000	0.000	0.000	*1.000***	0.000
$\hat{\pi}_{u\|5}$	0.000	0.000	0.000	0.000	0.000	*1.000***

Table 8 Estimated average transition probabilities under the HM model with $k = 6$ hidden states referred to the period 2016–2017; figures in italics are those in the main diagonal (significant at **at 1%)

	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 5$	$u = 6$
$\hat{\pi}_{u\|1}$	*1.000***	0.000	0.000	0.000	0.000	0.000
$\hat{\pi}_{u\|2}$	0.000	*1.000***	0.000	0.000	0.000	0.000
$\hat{\pi}_{u\|3}$	0.000	0.000	*1.000***	0.000	0.000	0.000
$\hat{\pi}_{u\|4}$	0.000	0.000	0.000	*0.697***	0.000	0.303*
$\hat{\pi}_{u\|5}$	0.000	0.000	0.000	0.028	*0.972***	0.000
$\hat{\pi}_{u\|6}$	0.000	0.000	0.000	0.000	0.000	*1.000***

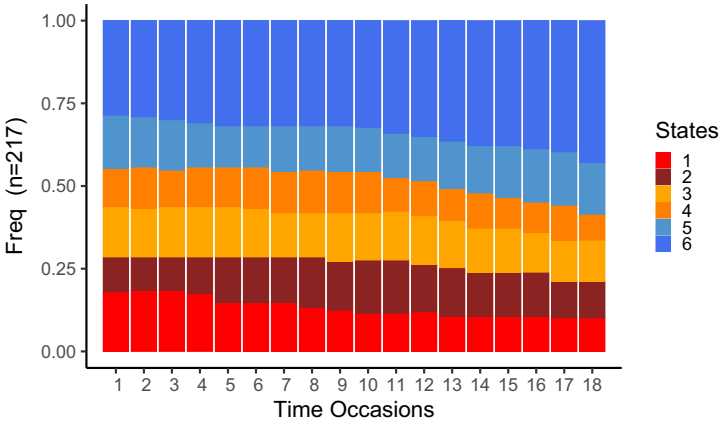


Fig. 2 Proportions of countries assigned to each latent state from 2000 to 2017 (eighteen years) under the HM model with $k = 6$ hidden states

Using *global decoding*, we can inspect the most dynamic countries in terms of changing clusters over time. Figure 2 shows the proportion of countries predicted in each cluster at every time occasion. It results that the frequencies of countries predicted in the 1st, 3rd, 4th, and 5th clusters are lower at the end of the period, whereas those of the 6th and 2nd clusters are higher. A better understanding of the predicted dynamics can be gained by the estimated relative frequencies of each cluster reported in Table 9 for years 2000, 2006, 2011, and 2017.

Looking at the predicted sequence of latent states, we also observe that twenty countries illustrated in Fig. 3 are predicted to have three transitions across states during the whole period except Bhutan (Thunder Dragon Kingdom), which results in the unique country to show four transitions. In particular, we notice that Afghanistan moves from the 1st up to the 4th cluster. El Salvador and Dominican Republic transit from the 4th to the 6th during the last period. Also, Ghana and Timor-Leste move from the 1st up to the 3rd cluster, during 2007–2008 and 2009–2010, respectively. These results are in line with the fact that, for example, Ghana is often considered one of the fast-developing countries in Africa due to a government that promoted a stable political environment and effective management of the significant natural resources present in the country. Timor-Leste is recognized as a fast-developing country in Southeast Asia due to natural resource revenue and poverty reduction after its independence in 2002; see the reports of The World Bank Group (2015, 2022), and Zallé (2019), for further details.

Figure 4 shows the maps of the countries according to the predicted states both in 2000 and 2017. At the beginning of the period, thirty-nine countries are allocated in the 1st cluster, and

Table 9 Proportions of countries assigned to each latent state under the selected HM model at time occasions corresponding to years 2000, 2006, 2011, and 2017

Year	$u = 1$	$u = 2$	$u = 3$	$u = 4$	$u = 5$	$u = 6$
2000	0.180	0.106	0.152	0.115	0.161	0.286
2006	0.147	0.138	0.129	0.129	0.138	0.318
2011	0.120	0.143	0.147	0.106	0.134	0.350
2017	0.101	0.111	0.124	0.074	0.157	0.433

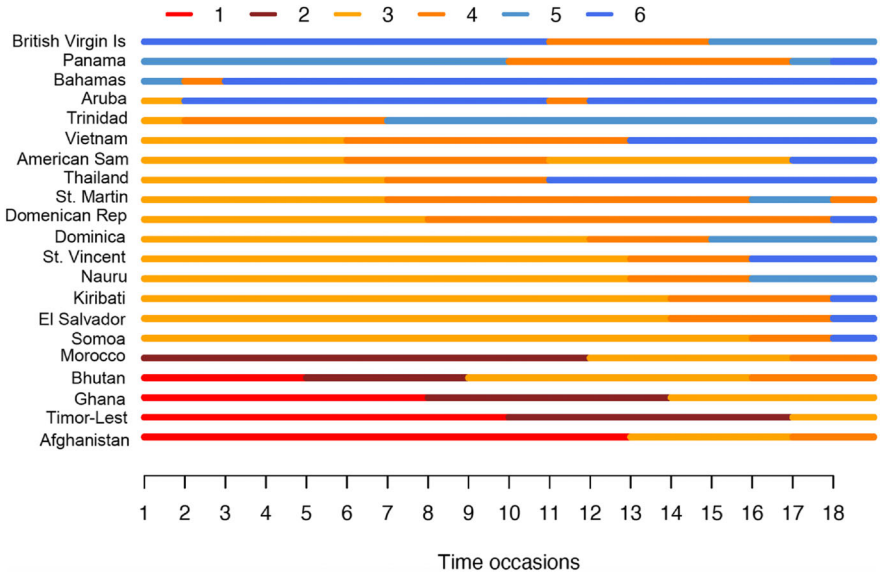


Fig. 3 Predicted sequence of latent states across years (from 2000 to 2017, eighteen years) under the HM model with $k = 6$ hidden states for countries estimated to switch between latent states more than three times over the period

only twenty of them remained in the same cluster at the end of the period,[4] thus confirming as the worst countries in terms of socioeconomic development. From the maps, we notice that many Central and Latin American countries made significant progress as well as India. In particular, the Republic of Ecuador has made progress in social and human development and Vietnam was the fastest growing economy in Asia.[5] On the other hand, Libya changed from the 5th to the 3rd cluster probably due to the civil war that started in 2014.[6]

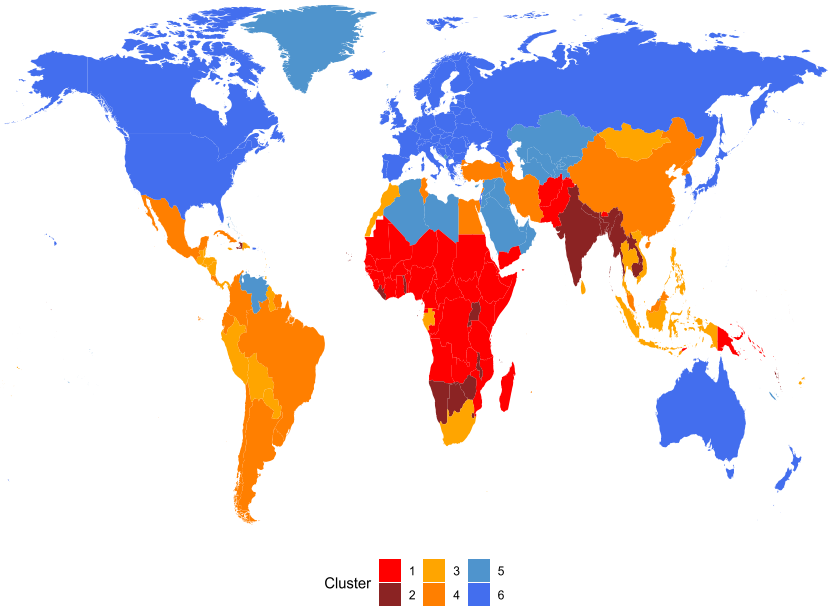
We assessed the clustering quality by the Q statistic proposed in Hennig and Coretto (2022). It measures non-parametrically how close the within-cluster distributions are to the elliptical unimodal distributions with the only mode on the mean. The realized values of the statistics are near zero (best value) for all the variables and clusters at every time occasion. The highest values of around 0.21 are observed only in some clusters and time occasions for the variable Ele.

The classification uncertainty is measured by the entropy calculated according to the posterior probabilities, which is 158.38 and, compared with its maximum 6,998.61, indicates that the model provides a suitable clustering structure while accounting for parsimony.

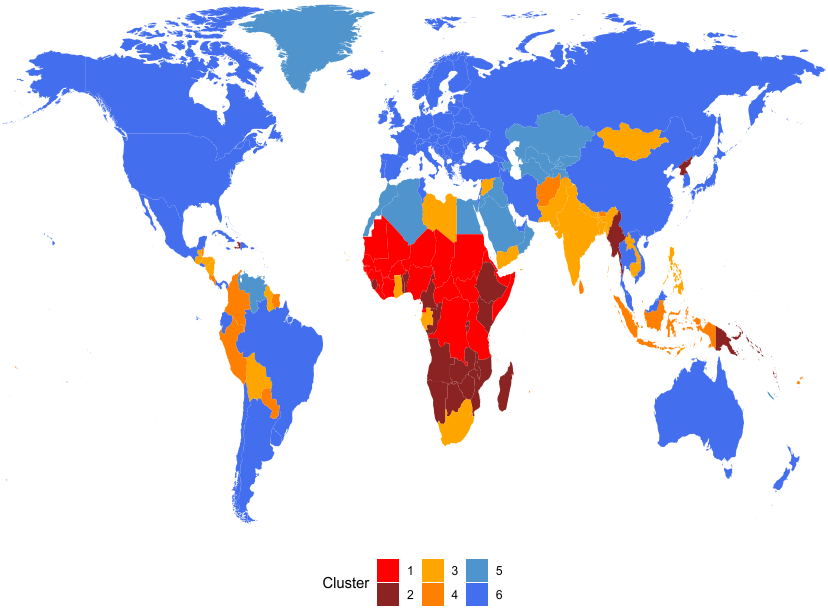
[4] The countries which are predicted in the 1st cluster for the whole period are: Burkina Faso, Central African Republic, Chad, Congo, Dem. Rep., Cote d'Ivoire, Djibouti, Equatorial Guinea, Eritrea, Gambia, The, Guinea, Guinea-Bissau, Mali, Mauritania, Niger, Nigeria, Senegal, Somalia, South Sudan, Sudan, Tanzania.

[5] For more details about Vietnam development see https://www.ft.com/content/fa1db5ce-8f65-4b28-ab6d-b78730f98195.

[6] For more details see https://www.worldbank.org/en/country/libya/publication/economic-outlook-april-2017



(a) 2000



(b) 2017

Fig. 4 Maps of the countries according to the predicted states under the HM model with $k = 6$ hidden states upper panel (a) in 2000, and bottom panel (b) in 2017

7 Conclusions

We introduce a novel approach for model and variable selection in the multivariate hidden Markov (HM) model with continuous variables for the analysis of time series and panel data while accounting for missing responses. Thus, to the best of our knowledge, we are the first to provide a general framework to: (i) account for complete or intermittent missing responses under the missing-at-random assumption while selecting the relevant set of variables useful for clustering purposes; (ii) cluster in a parsimonious way units showing similarities when the number of groups is unknown; (iii) provide a dynamic clustering allowing units to change groups over time.

We focus on the maximum likelihood estimation of model parameters through a modified version of the expectation-maximization algorithm. The algorithm also performs multiple imputations useful to predict the missing responses, conditionally or unconditionally to the assigned latent state. We develop a greedy search algorithm to achieve a dimensionality reduction of the complete set of variables to a smaller subset, by relying on an inclusion-exclusion procedure based on the Bayesian information criterion (Schwarz, 1978) to compare models. This criterion is seen as an approximation of that based on the Bayes factor through which we also select the optimal number of latent states. It has been proven reliable for estimating the number of mixture components for the HM model (Bacci et al., 2014) and, in general, for mixture models (Keribin, 2000).

The simulation study shows promising results; for all scenarios under comparison, the set of clustering variables is correctly identified in almost all the simulated samples and the true number of states is always correctly estimated. The different simulated scenarios allow us to conclude that the performance of the greedy search algorithm is slightly worse in the presence of missing values and when there is poor separation of the hidden states.

The proposal is applied to explore the socioeconomic growth of 217 countries for the period 2000–2017 using data collected by the World Bank and the UNESCO Institute for Statistics. The data are characterized by intermitting or complete missing responses at certain time occasions since countries do not always provide all data. Out of the 25 selected indicators, the proposed greedy search algorithm leads us to select 12 relevant indicators. The results of this application reveal the advantages of the proposed multivariate model-based approach in selecting the most relevant indicators to measure the growth and transformation of countries while accounting for the heterogeneity in development outcomes. In this way, countries are clustered into a suitable number of groups having straightforward interpretability. Moreover, countries are dynamically assigned to each cluster according to the estimated posterior probability, and it is possible to rank countries according to their improvements over time performing global decoding. Results of these analyses can also be suitable to evaluate if and how countries are reaching some of the sustainable development goals adopted by the United Nations in 2015 to ensure prosperity by 2030.[7] Our proposal can also be useful for policymakers because, through the predicted states, they can dispose of measures to monitor countries that are instead at risk of not further developing or remaining stagnant in poverty conditions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00357-023-09457-9.

Acknowledgements We acknowledge the financial support from the grant "Hidden Markov Models for Early Warning Systems" of Ministero dell'Università e della Ricerca (PRIN 2022TZEXKF) funded by European Union - Next Generation EU.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement.

[7] For more details see the webpage: https://www.undp.org/sustainable-development-goals

Data Availibility The dataset used in the applicative example is available at the website https://databank.worldbank.org/data/source/world-development-indicators.

Code Availability The code used to perform the proposed variable selection method for the hidden Markov model and the illustrative example is implemented in R by extending the functions of the `LMest` package (Bartolucci et al., 2017); the extended functions are available at the GitHub repository with link https://github.com/penful/HM_varSel.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Adams, S., & Beling, P. A. (2019). A survey of feature selection methods for Gaussian mixture models and hidden Markov models. *Artificial Intelligence Review, 52*, 1739–1779.

Bacci, S., Pandolfi, S., & Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification, 8*, 125–145.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49*, 803–821.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2013). *Latent Markov models for longitudinal data*. Boca Raton, FL: Chapman & Hall/CRC Press.

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.

Bartolucci, F., Montanari, G. E., & Pandolfi, S. (2016). Item selection by latent class-based methods: An application to nursing home evaluation. *Advances in Data Analysis and Classification, 10*, 245–262.

Bartolucci, F., Pandolfi, S., & Pennoni, F. (2017). LMest: An R package for latent Markov models for longitudinal categorical data. *Journal of Statistical Software, 81*, 1–38.

Bartolucci, F., Pandolfi, S., & Pennoni, F. (2022). Discrete latent variable models. *Annual Review of Statistics and its Application, 9*, 425–452.

Baum, L., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics, 41*, 164–171.

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science: With applications in R*. Cambridge, UK: Cambridge University Press.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26*, 211–243.

Celeux, G., & Durand, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics, 23*, 541–564.

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition, 28*, 781–793.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, MA: Cambridge University Press.

Dempster, A. P., Laird, N. M., & 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.

Flynt, A., & Dean, N. (2019). Growth mixture modeling with measurement selection. *Journal of Classification, 36*, 3–25.

Fop, M., & Murphy, T. B. (2018). Variable selection methods for model-based clustering. *Statistics Surveys, 12*, 18–65.

Frühwirth-Schnatter, S. (2011). Panel data analysis: A survey on model-based clustering of time series. *Advances in Data Analysis and Classification, 5*, 251–280.
Gales, M. J. (1999). Semi-tied covariance matrices for hidden Markov models. *IEEE Trans Speech Audio Process, 7*, 272–281.
Gormley, I., Murphy, T., & Raftery, A. (2023). Model-based clustering. *Annual Review of Statistics and its Application, 10*, 573–595.
Hennig, C., & Coretto, P. (2022). An adequacy approach for deciding the number of clusters for OTRIMLE robust Gaussian mixture-based clustering. *Australian & New Zealand Journal of Statistics, 64*, 230–254.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.
Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statitical Association, 90*, 773–795.
Keribin, C. (2000). Consistent estimation of the order of mixture models. Sankhya: The Indian Journal of Statistics. *Series A, 62*, 49–66.
Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data*. New York: John Wiley & Sons.
Maugis, C., Celeux, G., & Martin-Magniette, M.-L. (2009). Variable selection for clustering with Gaussian mixture models. *Biometrics, 65*, 701–709.
Maugis, C., Celeux, G., & Martin-Magniette, M.-L. (2009). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis, 53*, 3872–3882.
McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification, 33*, 331–373.
Nielsen, L. (2013). How to classify countries based on their level of development. *Social Indicators Research, 114*, 1087–1107.
Pandolfi, S., Bartolucci, F., & Pennoni, F. (2023). A hidden Markov model for continuous longitudinal data with missing responses and dropout. *Biometrical Journal, 65*, 1–28.
Pennoni, F., & Bal-Domńska, B. (2022). NEETs and youth unemployment: A longitudinal comparison across European countries. *Social Indicator Research, 162*, 739–761.
R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association, 101*, 168–178.
Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.
Schafer, J. L. (2022). *mix: Estimation/multiple imputation for mixed categorical and continuous data*. R package version 1.0-11.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.
Scrucca, L., & Raftery, A. E. (2018). clustvarsel: A package implementing variable selection for Gaussian model-based clustering in R. *Journal of Statistical Software, 84*, 1–28.
The World Bank Group (2015). Bhutan - macroeconomic and public finance policy note: Hydropower impact and public finance reforms towards economic self-reliance. *Bhutan - Macroeconomic and Public Finance Policy Note*. Available from: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/211211467995430678/bhutan-macroeconomic-and-public-finance-policy-note-hydropower-impact-and-public-finance-reforms-towards-economic-self-reliance.
The World Bank Group (2018). Data catalog: World development indicators. Available from: https://datacatalog.worldbank.org/dataset/world-development-indicators.
The World Bank Group (2022). Honoring the past, securing the future. *Timor-Leste Economic Report*. Available from https://thedocs.worldbank.org/en/doc/89b675c65dab346ea6d01ba0e536f0bc-0070012022/original/December-2022-Timor-Leste-Economic-Report.pdf.
Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory, 13*, 260–269.
Welch, L. R. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter, 53*, 1–13.
Zallé, O. (2019). Natural resources and economic growth in Africa: The role of institutional quality and human capital. *Resources Policy, 62*, 616–624.
Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov models for time series: An introduction using R*. Boca Raton, FL: CRC Press.