Annals of Operations Research
https://doi.org/10.1007/s10479-025-06986-x

ORIGINAL RESEARCH

# Parsimonious parametrizations of transition matrices of Markov chain and hidden Markov models

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

Received: 31 December 2024 / Accepted: 3 December 2025


**Abstract**
We introduce a set of constraints on the multinomial logit (sub)model for the transition probabilities of Markov chain and Hidden Markov models with covariates. These constraints have a straightforward interpretation and make the model more parsimonious with respect to the standard formulation. Estimation based on the maximum likelihood approach is developed under different constraints. The proposal is validated by a series of simulations and illustrated by an application about the evaluation of differences in general self-assessed health according to the available covariates, using longitudinal data from the Health and Retirement Study.

**Keywords** Discrete latent variable models · Expectation-Maximization algorithm · Longitudinal data · Multinomial logit models · Self-rated health

**Mathematics Subject Classification** 60J05 · 60J10 · 60J20 · 62F10 · 62M05

## 1 Introduction

In the analysis of time-series and longitudinal data, Markov Chain (MC; Anderson, 1954; Meyn and Tweedie, 2012) and Hidden Markov (HM) models, also known as latent Markov models (Bartolucci et al., 2013; Zucchini, MacDonald, and Langrock, 2016), are fundamen-

Francesco Bartolucci, Silvia Pandolfi, and Fulvia Pennoni have contributed equally to this work.

✉ Francesco Bartolucci
francesco.bartolucci@unipg.it

Silvia Pandolfi
silvia.pandolfi@unipg.it

Fulvia Pennoni
fulvia.pennoni@unimib.it

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

[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, Milan 20126, Italy

Published online: 18 March 2026

tal tools, especially when the analysis is focused on transitions, or the need to dynamically cluster individuals arises. In particular, with only one categorical response, an MC model allows studying the initial distribution of this response variable and its conditional distribution given the previous category. HM models are more flexible as they can be used with response variables of any nature, even when more responses are available at each time occasion. In fact, the basic assumption of HM models is that, underlying every time-specific response, there is a latent variable and that the sequence of these latent variables follows a Markov chain with a given number of latent states. Therefore, HM models are among the most important discrete latent variable models (Bartolucci et al., 2022).

Several applications of MC and HM models are nowadays available in many fields; see Mor et al. (2021) and Visser and Speekenbrink (2022), among others. In particular, the inclusion of latent variables in an HM model may be motivated in different ways (see Bartolucci et al., 2013; 2014). When a single categorical response is observed, an HM model with a number of latent states equal to the number of categories may be seen as an extension of an MC model for measurement errors. Another motivation is based on the assumption that the responses, possibly multivariate, depend on a discrete latent variable, the values of which correspond to separate clusters; in this way it is possible to perform model-based clustering in a dynamic fashion.

Especially with longitudinal data, which are of main interest for the present proposal, individual time-varying covariates are typically available. These covariates can be included in an MC or HM model by a suitable parametrization (Bartolucci et al., 2014), so that the model is time non-homogenous. Typically, both (sub)models for the initial and transition probabilities of the (possibly latent) Markov chain are based on multinomial logits, conditional on the previous state in the case of the transition probabilities; see Azzalini (1994) for alternative parametrizations for MC models. Regarding the application of HM models based on a multinomial logit parametrization of the transition probabilities, it is worth mentioning Spezia (2006), Meligkotsidou and Dellaportas (2011), Maruotti and Rocci (2012), Di Mari et al. (2016), Koki et al. (2020), and Wang et al. (2023).

With several latent states, say at least three, the multinomial logit parametrization of the transition probabilities makes the model very complex as it includes many parameters, which may be challenging to interpret, and could lead to numerical instabilities in performing Maximum Likelihood (ML) estimation, especially when transitions between certain pairs of states are unlikely. To clarify this point, consider that with $k$ states there are $k-1$ logits for the initial probabilities and $k(k-1)$ logits for the transition probabilities and, for each logit, the model uses $s+1$ parameters, being $s$ the number of covariates.

In order to face the parsimony issue of the MC and HM models with individual covariates, we introduce a series of meaningful constraints on the multinomial logit parameters for the transition probabilities. Paying attention to the simplicity of the proposed approach, all constraints that we consider are linear. Obviously, without covariates, these constraints may be expressed on the intercepts only. The proposed constraints also improve interpretability of the model, but we must be aware that, obviously, this comes at the cost of reduced flexibility.

Alternative approaches may be adopted to increase the interpretability and the parsimony of an MC or HM model. One of the main of these approaches was introduced by Bartolucci (2006), where linear constraints are directly expressed on the transition probabilities. The latter approach, while allowing constraints of interest such as that certain probabilities are

equal to zero, may be more complex to be used in practice. It is also obvious that reducing the number of latent states in an HM model makes the model more parsimonious. However, the point of view here adopted is that, in certain applications, it may be more interesting to employ a larger number of latent states with interpretable constraints on the transition probabilities, rather than using a reduced number of states without such constraints.

In the present paper, we also show how to perform ML estimation of the constrained model by a Newton–Raphson (NR) algorithm for MC models and an Expectation-Maximization (EM) algorithm for HM models. To assess the precision of the ML parameter estimates, asymptotic standard errors may be obtained in the usual way through the observed information matrix, which is directly provided by the NR algorithm. On the other hand, the EM algorithm does not use the information (neither expected nor observed) matrix. In such a situation, an alternative procedure to obtain standard errors for parameter estimates is to rely on a (parametric or nonparametric) bootstrap procedure (Davison & Hinkley, 1997). Obviously, apart from full ML estimation based on the EM algorithm, other methods may be applied to estimate the models developed in the present paper. In this regard, the three-step estimation method proposed in Bartolucci et al. (2015) and improved by Di Mari et al. (2016) is of particular interest for certain advantages in terms of interpretation and applicability; see also Vermunt (2025).

The proposed approach is validated by a simulation study aimed at showing the effectiveness of the proposed constrained models in recovering the true data generating process in terms of higher accuracy of the parameter estimates and lower numerical instability of the estimation process. We also present an application based on the Health and Retirement Study (HRS) that explores how the perceived individual health is associated with the available covariates. This study, conducted by the University of Michigan (Juster & Suzman, 1995), concerns aspects related to retirement and health among elderly individuals in the United States; see Bartolucci et al. (2014) for more details on the data and other model formulations. This application also demonstrates how to select a suitable constraint on the basis of the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978), and offers an interesting comparison between MC and HM models.

The remainder of the paper is organized as follows. After an outline of the MC and HM models of interest, which is provided in the following section, in Sect. 3 we illustrate the proposed constraints on the parameters of the transition probabilities of both models and we provide a guidance in the choice of the constraints for a specific application. Likelihood inference on the basis of algorithms of NR and EM type is illustrated in Sect. 4. In Sect. 5 we report the results of the simulation study. Results of the illustrative application based on the HRS data are reported in Sect. 6, while Sect. 7 contains main conclusions. Additional details on the applications are presented in the Supplementary Information (SI) file.

The algorithms used for model formulation and estimation have been implemented in a series of functions in `R` (R Core Team, 2025), which represent an extension of the `LMest` package (Bartolucci et al., 2017; Pennoni et al., 2025) and are available to readers upon request. These functions also rely on `Fortran` routines in order to enhance scalability with large datasets.

## 2 Preliminaries

Consider longitudinal data in which a vector of $r$ response variables $\boldsymbol{Y}_{it}$ is observed together with a corresponding vector of covariates $\boldsymbol{x}_{it}$, considered as fixed, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. In the following, we outline the MC and HM approaches to model these data. Note that we refer to the case of balanced longitudinal data, where we have the same number of observations $T$ for all sample units, but all the following methods can be easily adapted to the case of unbalanced longitudinal data. In the latter case a specific number of time occasions $T_i$ is available for each individual $i$.

### 2.1 Markov chain approach

When $\boldsymbol{Y}_{it}$ corresponds to a single categorical response variable denoted by $Y_{it}$ and with $k$ categories labeled from 1 to $k$, an MC model with $k$ states can be directly formulated. In particular, the first-order version of this model assumes that, for $t = 3, \ldots, T$ and given $Y_{i,t-1}$, $Y_{it}$ is conditionally independent of $Y_{i1}, \ldots, Y_{i,t-2}$. The model is characterized by initial and transition probabilities that, for every $i = 1, \ldots, n$, are denoted by

$$\begin{aligned} \lambda_{i,u} &= p(Y_{i1} = u), \quad u = 1, \ldots, k, \\ \pi_{it,uv} &= p(Y_{it} = v | Y_{i,t-1} = u), \quad t = 2, \ldots, T, \; u, v = 1, \ldots, k, \end{aligned}$$

respectively. These probabilities are collected in the column vectors $\boldsymbol{\lambda}_i$, $i = 1, \ldots, n$, and in the transition matrices $\boldsymbol{\Pi}_{it}$, $i = 1, \ldots, n$, $t = 2, \ldots, T$. The dependence of the initial probabilities on the vector of covariates in $\boldsymbol{x}_{i1}$, as well as the dependence of the transition probabilities at time $t$ on $\boldsymbol{x}_{it}$, is not explicitly indicated.

In modeling the dependence of the above probabilities on the covariates, the natural parametrization is of multinomial logit type for all $n$ units (Agresti, 2013; Bartolucci et al., 2014). For the initial probabilities this parametrization assumes that

$$\log \frac{\lambda_{i,u}}{\lambda_{i,1}} = \alpha_u + \boldsymbol{x}_{i1}' \boldsymbol{\beta}_u, \quad u = 2, \ldots, k. \tag{1}$$

In the following developments, it is convenient to express this *initial (sub)model* parametrization using the matrix notation. In particular, we have that

$$\boldsymbol{\lambda}_i = \frac{1}{\mathbf{1}_k' \exp(\boldsymbol{G} \boldsymbol{X}_{i1} \boldsymbol{\theta}_1)} \exp(\boldsymbol{G} \boldsymbol{X}_{i1} \boldsymbol{\theta}_1),$$

where $\mathbf{1}_k$ denotes a column vector of $k$ ones, $\boldsymbol{G} = \bar{\boldsymbol{I}}_{k,1}$ and, in general, $\boldsymbol{X}_{it} = \boldsymbol{I}_{k-1} \otimes (1, \boldsymbol{x}_{it}')$. In the previous expressions, $\boldsymbol{I}_k$ denotes an identity matrix of dimension $k$, $\bar{\boldsymbol{I}}_{k,u}$ denotes an identity matrix of dimension $k$ without the $u$-th column, and $\otimes$ denotes the Kronecker product. The parameter vector $\boldsymbol{\theta}_1$ has elements $(\alpha_2, \boldsymbol{\beta}_2', \ldots, \alpha_k, \boldsymbol{\beta}_k')'$ and then is of length $(k-1)(s+1)$, where $s$ is the number of covariates in each vector $\boldsymbol{x}_{it}$. However, the main focus is usually on the transition probabilities. In this regard, we formulate the *transition (sub)model* by assuming that

$$l_{it,uv} = \log \frac{\pi_{it,uv}}{\pi_{it,uu}} = \gamma_{uv} + \boldsymbol{x}'_{it}\boldsymbol{\delta}_{uv}, \quad u, v = 1, \ldots, k,\ v \neq u, \tag{2}$$

where the logits $l_{it,uv}$ have, as reference state, that corresponding to the row of the transition matrix. An alternative parametrization may be based on using the first state as a reference category; however, we consider the latter one as less interpretable. Denoting by $\boldsymbol{\pi}_{it,u}$ the $u$-th row of the transition matrix $\boldsymbol{\Pi}_{it}$, parametrization (2) may be expressed as

$$\boldsymbol{\pi}_{it,u} = \frac{1}{\mathbf{1}'_k \exp(\boldsymbol{H}_u \boldsymbol{X}_{it} \boldsymbol{\eta}_u)} \exp(\boldsymbol{H}_u \boldsymbol{X}_{it} \boldsymbol{\eta}_u),$$

where $\boldsymbol{H}_u = \bar{\boldsymbol{I}}_{k,u}$, $\boldsymbol{X}_{it}$ is a design matrix defined as above, and $\boldsymbol{\eta}_u$ has elements $(\gamma_{u2}, \boldsymbol{\delta}'_{u2}, \ldots, \gamma_{uk}, \boldsymbol{\delta}'_{uk})'$ for $u = 1$, $(\gamma_{u1}, \boldsymbol{\delta}'_{u1}, \gamma_{u3}, \boldsymbol{\delta}'_{u3}, \ldots, \gamma_{uk}, \boldsymbol{\delta}'_{uk})'$ for $u = 2$, and so on for $u = 3, \ldots, k$. These parameter vectors are collected in the unique vector $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k)'$. As we show in the following, we will assume that

$$\boldsymbol{\eta} = \boldsymbol{Z}\boldsymbol{\theta}_2, \tag{3}$$

for a suitably defined design matrix $\boldsymbol{Z}$ and parameter vector $\boldsymbol{\theta}_2$, in order to make the model more parsimonious, as we propose in Sect. 3.

### 2.2 Hidden Markov approach

In the general case, $\boldsymbol{Y}_{it}$ may contain more response variables of any type, and the HM model assumes that, given a discrete latent process $U_{i1}, \ldots, U_{iT}$ with $k$ states, the response vectors $\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{iT}$ are conditionally independent, for $i = 1, \ldots, n$. The model is then based on a *measurement (sub)model*, corresponding to the conditional distribution of every $\boldsymbol{Y}_{it}$ given $U_{it}$, and initial and transition (sub)models (in the same sense as in Sect. 2.1) regarding the distribution of every unit-specific latent process.

To provide some examples, when $\boldsymbol{Y}_{it}$ is a vector of continuous variables, the measurement (sub)model may assume that

$$\boldsymbol{Y}_{it} | U_{it} = u \sim N_r(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}), \quad u = 1, \ldots, k, \tag{4}$$

where $N_r(\cdot)$ denotes the multivariate normal distribution of order $r$, equal to the number of response variables, $\boldsymbol{\mu}_u$ is the mean vector that is specific of latent state $u$, and $\boldsymbol{\Sigma}$ is the common variance-covariance matrix under the constraint of homoskedasticity. Note that the variance-covariance matrix may also be allowed to be heteroskedastic, that is, $\boldsymbol{\Sigma}_u$ can be assumed to vary across latent states to account for more complex data structures. However, in this case the number of parameters increases and the model becomes less parsimonious, although it is more flexible. Moreover, as for finite mixture models, the likelihood can be unbounded (McLachlan and Peel, 2000, Sec.2.5). With categorical response variables, the main formulation is based on the parameters

$$\phi_{jy,u} = p(Y_{ijt} = y | U_{it} = u), \quad j = 1, \ldots, r,\ u = 1, \ldots, k,\ y = 1, \ldots, c_j, \tag{5}$$

where $Y_{ijt}$, for every $t$, denote the $j$-variable in $\boldsymbol{Y}_{it}$ and $c_j$ is the number of categories of this variable. More sophisticated parametrizations may be based on suitable constraints that have a specific interpretation; for instance, with binary variables coded as 1 for a correct response and 2 for a wrong response, we may assume an HM Rasch model (Bartolucci et al., 2008) based on the assumption

$$\log \frac{\phi_{j1,u}}{\phi_{j2,u}} = \xi_u - \nu_j, \quad j = 1, \ldots, r,\ u = 1, \ldots, k,$$

where $\xi_u$ is an ability parameter and $\nu_j$ is a difficulty parameter. This parametrization may be suitably extended to the case of response variables with more than two categories using common approaches in the item response theory literature (Hambleton et al., 1991; Cai et al., 2016).

A constraint that in certain applications may be adopted is that the latent states are ordered (Bartolucci et al., 2009). With continuous variables and when assumption (4) is formulated, ordered latent states may be imposed by requiring that

$$\boldsymbol{\mu}_1 \leq \cdots \leq \boldsymbol{\mu}_k, \tag{6}$$

where the inequalities act element-wise. With categorical response variables, ordinal latent states are formulated by requiring that

$$\boldsymbol{\psi}_{j,1} \geq \cdots \geq \boldsymbol{\psi}_{j,k}, \quad j = 1, \ldots, r, \tag{7}$$

where $\boldsymbol{\psi}_{j,u}$ is the vector of the cumulative probabilities $\psi_{jy,u} = \phi_{j1,u} + \ldots + \phi_{jy,u}$, $y = 1, \ldots, c_j$, so that the distributions are stochastically ordered with the latent state. For any parametrization assumed on the conditional distribution of the response variables given the latent states, the parameters of the corresponding (sub)model will be collected in vector $\boldsymbol{\theta}_3$.

Regarding the latent structure, the typical assumption is that the Markov chain is of first order and, for the initial and transition probabilities, we can use the same notation previously adopted for the MC model. More precisely, we introduce the initial and transition probabilities

$$\begin{aligned} \lambda_{i,u} &= p(U_{i1} = u), \quad u = 1, \ldots, k \\ \pi_{it,uv} &= p(U_{it} = v | U_{i,t-1} = u), \quad t = 2, \ldots, T,\ u, v = 1, \ldots, k, \end{aligned}$$

which may depend on the individual covariates as already described in Sect. 2.1; see in particular equations (1) and (2). The corresponding vectors of parameters will be again denoted by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and have the same structure illustrated above.

## 3 Proposed models

In the following, we illustrate a set of constraints on the transition (sub)model for either an MC or HM model through the multinomial parametrization (2). We recall that these constrained models will be expressed by the design matrix $\boldsymbol{Z}$ according to the linear form (3). This matrix has the structure

$$\boldsymbol{Z} = \left(\boldsymbol{Z}_1 \otimes \bar{\mathbf{1}}_{s+1} \quad \boldsymbol{Z}_2 \otimes \bar{\boldsymbol{I}}_{s+1,1}\right), \tag{8}$$

where $\bar{\mathbf{1}}_{s+1}$ denotes the first column of the identity matrix $\boldsymbol{I}_{s+1}$. In this way, recalling the definition of $\boldsymbol{\eta}$ given at the end of Sect. 2.1, the first block of columns based on $\boldsymbol{Z}_1$ is used to impose constraints on the intercepts $\gamma_{uv}$, while the second block based on $\boldsymbol{Z}_2$ is used to impose constraints on $\boldsymbol{\delta}_{uv}$. Both matrices $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ have $k(k-1)$ rows and they are simply equal to $\boldsymbol{I}_{k(k-1)}$ when no constraints are assumed on the parameters at issue. For the constrained cases, the structure of these matrices is clarified in Appendix A.

In the following, we first illustrate general constraints on the parameters at issue and then constraints that make sense when the latent states are ordered. This corresponds to the case of ordinal response variables when an MC model is assumed or to the case of ordinal latent states when a constraint of type (6) or (7) is adopted with HM models.

### 3.1 General constraints

We initially propose some constraints on the intercepts $\gamma_{uv}$ that do not necessarily require the states to have an ordinal nature. According to the first of these constraints, the parameters $\gamma_{uv}$ are simply *constant*:

$$\text{CONST-INT} : \gamma_{uv} = \bar{\gamma},$$

where, as for the following constraints, $u, v = 1, \ldots, k,\ v \neq u$. To clarify, we report below, for $k = 4$ states and without the component associated to the covariates, the structure of the matrix $\boldsymbol{L}_{it}$ of the logits $l_{it,uv}$ defined in (2) and that of $\boldsymbol{\Pi}_{it}$ for a numerical example with $\bar{\gamma} = -1$:

$$\boldsymbol{L}_{it} = \begin{pmatrix} 0 & \bar{\gamma} & \bar{\gamma} & \bar{\gamma} \\ \bar{\gamma} & 0 & \bar{\gamma} & \bar{\gamma} \\ \bar{\gamma} & \bar{\gamma} & 0 & \bar{\gamma} \\ \bar{\gamma} & \bar{\gamma} & \bar{\gamma} & 0 \end{pmatrix}, \quad \boldsymbol{\Pi}_{it} = \begin{pmatrix} 0.475 & 0.175 & 0.175 & 0.175 \\ 0.175 & 0.475 & 0.175 & 0.175 \\ 0.175 & 0.175 & 0.475 & 0.175 \\ 0.175 & 0.175 & 0.175 & 0.475 \end{pmatrix}.$$

We interpret each logit $l_{it,uv}$ as the *differential attraction* of state *v* with respect to state *u* and then, under the constraint at issue, we note that this attraction is always the same, and the off-diagonal elements of $\boldsymbol{L}_{it}$ and $\boldsymbol{\Pi}_{it}$ are constant. Moreover, as $\bar{\gamma}$ increases these elements obviously rise, meaning that the tendency to move away from the current state, expressed as transition probability, increases, while the persistence probability decreases. To provide a clearer understanding of this constraint, a graphical representation of the resulting transition probability matrix is shown in the top-left panel of Fig. 1.

Other constraints are those of *symmetry* or *reverse symmetry*, which may be expressed as

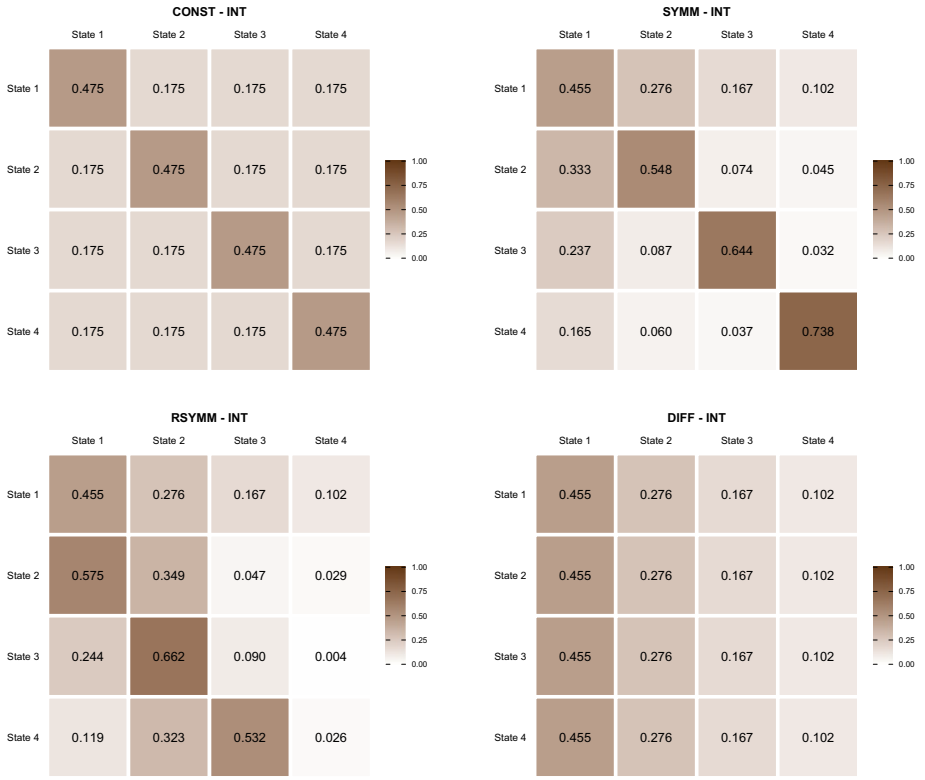


**Fig. 1** Graphical representation of the transition probability matrices, $\mathbf{\Pi}_{it}$, resulting from different constraints: CONST-INT (top-left), SYMM-INT (top-right), RSYMM-INT (bottom-left), DIFF-INT (bottom-right). Darker shades indicate higher probabilities

$$\text{SYMM-INT} : \gamma_{uv} = \begin{cases} \bar{\gamma}_{uv} & \text{if } v > u, \\ \bar{\gamma}_{vu} & \text{if } v < u, \end{cases}$$

$$\text{RSYMM-INT} : \gamma_{uv} = \begin{cases} \bar{\gamma}_{uv} & \text{if } v > u, \\ -\bar{\gamma}_{vu} & \text{if } v < u. \end{cases}$$

As examples, for the symmetric constraint (SYMM-INT), we consider $k = 4$, $\bar{\gamma}_{12} = -0.5$, $\bar{\gamma}_{13} = -1$, $\bar{\gamma}_{14} = -1.5$, $\bar{\gamma}_{23} = -2$, $\bar{\gamma}_{24} = -2.5$, and $\bar{\gamma}_{34} = -3$. Therefore, we have the following structure for $\boldsymbol{L}_{it}$ and $\mathbf{\Pi}_{it}$:

$$\boldsymbol{L}_{it} = \begin{pmatrix} 0 & \bar{\gamma}_{12} & \bar{\gamma}_{13} & \bar{\gamma}_{14} \\ \bar{\gamma}_{12} & 0 & \bar{\gamma}_{23} & \bar{\gamma}_{24} \\ \bar{\gamma}_{13} & \bar{\gamma}_{23} & 0 & \bar{\gamma}_{34} \\ \bar{\gamma}_{14} & \bar{\gamma}_{24} & \bar{\gamma}_{34} & 0 \end{pmatrix}, \quad \mathbf{\Pi}_{it} = \begin{pmatrix} 0.455 & 0.276 & 0.167 & 0.102 \\ 0.333 & 0.548 & 0.074 & 0.045 \\ 0.237 & 0.087 & 0.644 & 0.032 \\ 0.165 & 0.060 & 0.037 & 0.738 \end{pmatrix};$$

moreover, the following matrices result for the second case (RSYMM-INT) with the same parameter values:

$$\boldsymbol{L}_{it} = \begin{pmatrix} 0 & \bar{\gamma}_{12} & \bar{\gamma}_{13} & \bar{\gamma}_{14} \\ -\bar{\gamma}_{12} & 0 & \bar{\gamma}_{23} & \bar{\gamma}_{24} \\ -\bar{\gamma}_{13} & -\bar{\gamma}_{23} & 0 & \bar{\gamma}_{34} \\ -\bar{\gamma}_{14} & -\bar{\gamma}_{24} & -\bar{\gamma}_{34} & 0 \end{pmatrix}, \quad \boldsymbol{\Pi}_{it} = \begin{pmatrix} 0.455 & 0.276 & 0.167 & 0.102 \\ 0.575 & 0.349 & 0.047 & 0.029 \\ 0.244 & 0.662 & 0.090 & 0.004 \\ 0.119 & 0.323 & 0.532 & 0.026 \end{pmatrix}.$$

Both transition probability matrices, $\boldsymbol{\Pi}_{it}$, are graphically displayed in Fig. 1, in the top-right and bottom-left panels, respectively.

In the first case, the attraction of state *v* with respect to *u* is the same as the attraction of *u* with respect to *v*, with both attractions increasing with $\bar{\gamma}_{uv}$, and then $\boldsymbol{L}_{it}$ is symmetric. Obviously, this symmetry does not hold for the transition matrix $\boldsymbol{\Pi}_{it}$ in general, although as $\bar{\gamma}_{uv}$ increases, with $\bar{\gamma}_{u'v'}$ remaining constant for $(u', v') \neq (u, v)$, both $\pi_{uv}$ and $\pi_{vu}$ rise. Under the RSYMM-INT constraint, the attraction of state *v* with respect to *u* is the opposite of the attraction of *u* with respect to *v* with rather obvious consequences on the elements of $\boldsymbol{L}_{it}$ and $\boldsymbol{\Pi}_{it}$ as $\bar{\gamma}_{uv}$ increases while keeping the other parameters constant.

Finally, we consider the constraint

$$\text{DIFF-INT} : \gamma_{uv} = \bar{\gamma}_v - \bar{\gamma}_u,$$

under the identifiability constraint $\bar{\gamma}_1 = 0$. For example, we have the following matrices, with $\boldsymbol{\Pi}_{it}$ computed letting $\bar{\gamma}_2 = -0.5$, $\bar{\gamma}_3 = -1$, and $\bar{\gamma}_4 = -1.5$:

$$\boldsymbol{L}_{it} = \begin{pmatrix} 0 & \bar{\gamma}_2 & \bar{\gamma}_3 & \bar{\gamma}_4 \\ -\bar{\gamma}_2 & 0 & \bar{\gamma}_3 - \bar{\gamma}_2 & \bar{\gamma}_4 - \bar{\gamma}_2 \\ -\bar{\gamma}_3 & \bar{\gamma}_2 - \bar{\gamma}_3 & 0 & \bar{\gamma}_4 - \bar{\gamma}_3 \\ -\bar{\gamma}_4 & \bar{\gamma}_2 - \bar{\gamma}_4 & \bar{\gamma}_3 - \bar{\gamma}_4 & 0 \end{pmatrix}, \quad \boldsymbol{\Pi}_{it} = \begin{pmatrix} 0.455 & 0.276 & 0.167 & 0.102 \\ 0.455 & 0.276 & 0.167 & 0.102 \\ 0.455 & 0.276 & 0.167 & 0.102 \\ 0.455 & 0.276 & 0.167 & 0.102 \end{pmatrix};$$

see Fig. 1 (bottom-right panel) for the matrix plot of $\boldsymbol{\Pi}_{it}$. Note that parameter $\bar{\gamma}_v$ is a general measure of attraction of state *v*; see Bartolucci et al. (2015) for an example of application of this constraint. It is also worth noting that the same constraint may be formulated using logits for the transition probabilities having the first as reference state and assuming that each of these logits only depends on the destination state. In particular, in absence of covariates we have

$$\log \frac{\pi_{it,uv}}{\pi_{it,u1}} = \bar{\gamma}_v, \quad u = 1, \ldots, k,\ v = 2, \ldots, k,$$

and then this is the case of independence between consecutive latent states because, as shown in the previous example, the rows of the transition matrix are identical. Obviously, this does not generally happen when a component depending on the covariates is included. As $\bar{\gamma}_v$ increases while $\bar{\gamma}_u$ remains constant ($u \neq v$), the elements of $\boldsymbol{L}_{it}$ in the corresponding columns increase, as do those of $\boldsymbol{\Pi}_{it}$.

The same constraints as above may be expressed on the vector of regression coefficients $\boldsymbol{\delta}_{uv}$; in particular, we have:

$$
\begin{aligned}
\text{CONST-COV} &: \boldsymbol{\delta}_{uv} = \bar{\boldsymbol{\delta}}, \\
\text{SYMM-COV} &: \boldsymbol{\delta}_{uv} = \begin{cases} \bar{\boldsymbol{\delta}}_{uv} & \text{if } v > u, \\ \bar{\boldsymbol{\delta}}_{vu} & \text{if } v < u, \end{cases} \\
\text{RSYMM-COV} &: \boldsymbol{\delta}_{uv} = \begin{cases} \bar{\boldsymbol{\delta}}_{uv} & \text{if } v > u, \\ -\bar{\boldsymbol{\delta}}_{vu} & \text{if } v < u, \end{cases} \\
\text{DIFF-COV} &: \boldsymbol{\delta}_{uv} = \bar{\boldsymbol{\delta}}_{v} - \bar{\boldsymbol{\delta}}_{u},
\end{aligned}
$$

where in the last case we assume $\bar{\boldsymbol{\delta}}_1 = \mathbf{0}$.

The interpretation of the parameter vectors $\bar{\boldsymbol{\delta}}$, $\bar{\boldsymbol{\delta}}_{uv}$, and $\bar{\boldsymbol{\delta}}_u$ may be derived on the basis of that provided for the corresponding intercept parameters $\bar{\gamma}$, $\bar{\gamma}_{uv}$, and $\bar{\gamma}_u$. For instance, in the case of CONST-COV, a positive value of an element of $\bar{\boldsymbol{\delta}}$ means that the corresponding covariate has an increasing effect on the attraction of $v$ with respect to $u$, and this effect is constant for $v \neq u$; this implies that the probability of moving away from the current state also rises.

Note that in our approach it is possible to combine a specific constraint on the intercepts with a different one on the regression coefficients, as for instance SYMM-INT with CONST-COV, and the structure of the matrix of the logits $\boldsymbol{L}_{it}$ and the transition matrix $\boldsymbol{\Pi}_{it}$ will depend on both constraints and the specific value of the covariates. See Sect. 3.3 for further comments.

### 3.2 Constrains with ordinal states

With ordinal states, we can formulate specific constraints for this case. The first is that only the distance between the states matters, that is,

$$
\text{DIST1-INT} : \gamma_{uv} = \bar{\gamma}_{v-u},
$$

where, again, $u, v = 1, \ldots, k$, $v \neq u$. As an example, for the case of $k = 4$ and without covariates, we have the following matrices, where $\boldsymbol{\Pi}_{it}$ is computed with $\bar{\gamma}_{-3} = -3$, $\bar{\gamma}_{-2} = -2.5$, $\bar{\gamma}_{-1} = -2$, $\bar{\gamma}_1 = -0.5$, $\bar{\gamma}_2 = -1$, and $\bar{\gamma}_3 = -1.5$:

$$
\boldsymbol{L}_{it} = \begin{pmatrix} 0 & \bar{\gamma}_1 & \bar{\gamma}_2 & \bar{\gamma}_3 \\ \bar{\gamma}_{-1} & 0 & \bar{\gamma}_1 & \bar{\gamma}_2 \\ \bar{\gamma}_{-2} & \bar{\gamma}_{-1} & 0 & \bar{\gamma}_1 \\ \bar{\gamma}_{-3} & \bar{\gamma}_{-2} & \bar{\gamma}_{-1} & 0 \end{pmatrix}, \quad \boldsymbol{\Pi}_{it} = \begin{pmatrix} 0.455 & 0.276 & 0.167 & 0.102 \\ 0.064 & 0.474 & 0.287 & 0.174 \\ 0.045 & 0.074 & 0.548 & 0.333 \\ 0.039 & 0.065 & 0.107 & 0.789 \end{pmatrix}.
$$

The graphical representation of $\boldsymbol{\Pi}_{it}$ is shown in Fig. 2 (top-left panel).

In this case, the elements in each diagonal of $\boldsymbol{L}_{it}$ are constant, although the same does not generally hold for $\boldsymbol{\Pi}_{it}$. However, we can establish that an increase of $\bar{\gamma}_w$, with the other parameters kept fixed, implies that the elements in the $w$-th diagonal of $\boldsymbol{\Pi}_{it}$, where $w$ corresponds to $v - u$, rise and then the probabilities of persistence decrease. It is important to note that the above constraint implicitly assumes the existence of a set of equally spaced scores for the different states so that the parameter $\bar{\gamma}_1$ affects the transition from state 1 to state 2, from state 2 to state 3, and so on. Obviously, when there are substantial reasons to assume a different system of scores, an alternative version of the constraint DIST1-INT may be adopted. For instance, we may assume that the logit for the transition from state 2 to state

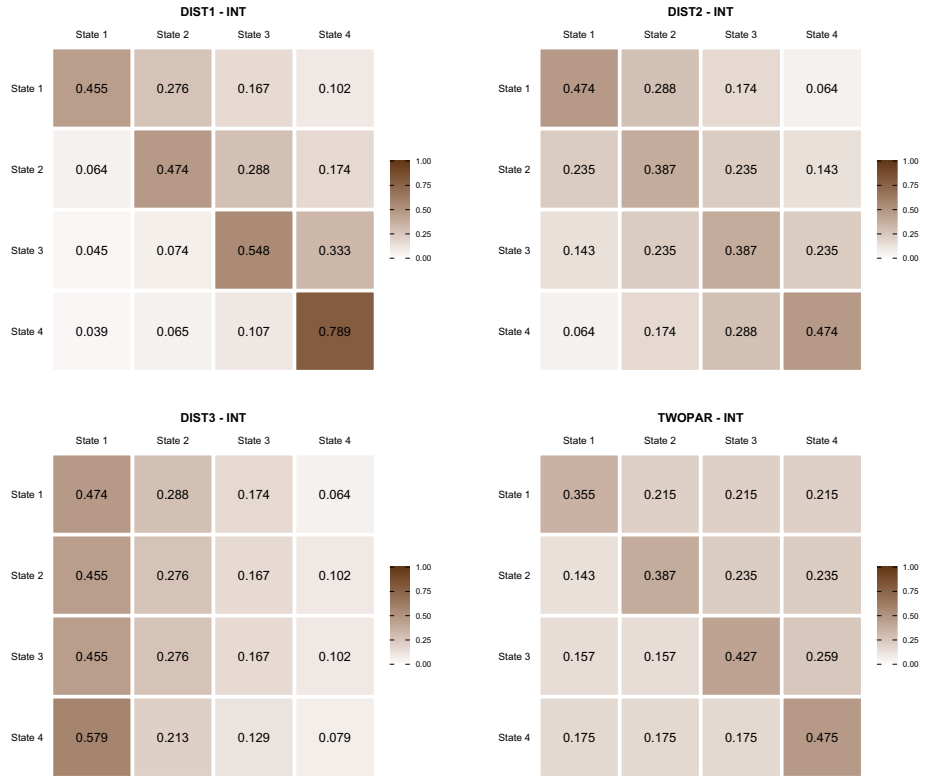


**Fig. 2** Graphical representation of the transition probability matrices, $\mathbf{\Pi}_{it}$, resulting from different constraints: DIST1-INT (top-left), DIST2-INT (top-right), DIST3-INT (bottom-left), TWOPAR-INT (bottom-right). Darker shades indicate higher probabilities

3 is equal to $\bar{\gamma}_2$ instead of $\bar{\gamma}_1$ when we know that the second and third states are far apart the double than the first and the second states. An assumption of this type may be formulated, in particular, for MC models when the states directly correspond to observable response categories, whereas assuming a system of non-equally spaced scores is more difficult to justify in HM models. In this regard, the literature on categorical data based upon the row-column model (Goodman, 1979; Bartolucci & Forcina, 2002) may provide useful insights. The same considerations hold for the other constraints illustrated in this section.

Nested cases in the previous example are obtained when the distance is taken in absolute value, while preserving the sign, so that

$$\text{DIST2-INT} : \gamma_{uv} = \begin{cases} \bar{\gamma}_{v-u} & \text{if } v > u, \\ \bar{\gamma}_{u-v} & \text{if } v < u, \end{cases}$$

$$\text{DIST3-INT} : \gamma_{uv} = \begin{cases} \bar{\gamma}_{v-u} & \text{if } v > u, \\ -\bar{\gamma}_{u-v} & \text{if } v < u. \end{cases}$$

As a possible example we have the following matrices, where $\mathbf{\Pi}_{it}$ is computed with $\bar{\gamma}_1 = -0.5, \bar{\gamma}_2 = -1, \bar{\gamma}_3 = -2$

$$\boldsymbol{L}_{it} = \begin{pmatrix} 0 & \bar{\gamma}_1 & \bar{\gamma}_2 & \bar{\gamma}_3 \\ \bar{\gamma}_1 & 0 & \bar{\gamma}_1 & \bar{\gamma}_2 \\ \bar{\gamma}_2 & \bar{\gamma}_1 & 0 & \bar{\gamma}_1 \\ \bar{\gamma}_3 & \bar{\gamma}_2 & \bar{\gamma}_1 & 0 \end{pmatrix}, \quad \boldsymbol{\Pi}_{it} = \begin{pmatrix} 0.474 & 0.287 & 0.174 & 0.064 \\ 0.235 & 0.387 & 0.235 & 0.143 \\ 0.143 & 0.235 & 0.387 & 0.235 \\ 0.064 & 0.174 & 0.287 & 0.474 \end{pmatrix};$$

for the second case and with the same parameter values, we have:

$$\boldsymbol{L}_{it} = \begin{pmatrix} 0 & \bar{\gamma}_1 & \bar{\gamma}_2 & \bar{\gamma}_3 \\ -\bar{\gamma}_1 & 0 & \bar{\gamma}_1 & \bar{\gamma}_2 \\ -\bar{\gamma}_2 & -\bar{\gamma}_1 & 0 & \bar{\gamma}_1 \\ -\bar{\gamma}_3 & -\bar{\gamma}_2 & -\bar{\gamma}_1 & 0 \end{pmatrix}, \quad \boldsymbol{\Pi}_{it} = \begin{pmatrix} 0.474 & 0.287 & 0.174 & 0.064 \\ 0.455 & 0.276 & 0.167 & 0.102 \\ 0.455 & 0.276 & 0.167 & 0.102 \\ 0.579 & 0.213 & 0.129 & 0.078 \end{pmatrix}.$$

The transition matrices obtained under constraints DIST2-INT and DIST3-INT are depicted in Fig. 2 (top-right and bottom-left panels, respectively). In these two cases we are combining constraint DIST1-INT with SYMM-INT or RSYMM-INT and, considering this aspect, we can easily interpret the involved parameters $\bar{\gamma}_w$.

Finally, we can use a different parameter uniformly for $v > u$ and for $v < u$, so that

$$\text{TWOPAR-INT} : \gamma_{uv} = \begin{cases} \bar{\gamma}_1 & \text{if } v > u, \\ \bar{\gamma}_{-1} & \text{if } v < u. \end{cases}$$

The following is an example where in computing $\boldsymbol{\Pi}_{it}$ we use $\bar{\gamma}_{-1} = -1$ and $\bar{\gamma}_1 = -0.5$:

$$\boldsymbol{L}_{it} = \begin{pmatrix} 0 & \bar{\gamma}_1 & \bar{\gamma}_1 & \bar{\gamma}_1 \\ \bar{\gamma}_{-1} & 0 & \bar{\gamma}_1 & \bar{\gamma}_1 \\ \bar{\gamma}_{-1} & \bar{\gamma}_{-1} & 0 & \bar{\gamma}_1 \\ \bar{\gamma}_{-1} & \bar{\gamma}_{-1} & \bar{\gamma}_{-1} & 0 \end{pmatrix}, \quad \boldsymbol{\Pi}_{it} = \begin{pmatrix} 0.355 & 0.215 & 0.215 & 0.215 \\ 0.143 & 0.387 & 0.235 & 0.235 \\ 0.157 & 0.157 & 0.427 & 0.259 \\ 0.175 & 0.175 & 0.175 & 0.475 \end{pmatrix}.$$

Therefore, all upper diagonal elements and all lower diagonal elements of $\boldsymbol{L}_{it}$ are constant. Though the same does not hold for $\boldsymbol{\Pi}_{it}$, we can easily check that as $\bar{\gamma}_1$ increases, with $\bar{\gamma}_{-1}$ kept fixed, the upper diagonal elements of this matrix increase and then the persistence probabilities decrease. A similar consequence is observed as $\bar{\gamma}_{-1}$ increases with $\bar{\gamma}_1$ kept fixed. The bottom-right panel of Fig. 2 shows the resulting transition matrix. Expressed for the vector of regression coefficients, these constraints become

$$\begin{aligned} \text{DIST1-COV} &: \boldsymbol{\delta}_{uv} = \bar{\boldsymbol{\delta}}_{v-u}, \\ \text{DIST2-COV} &: \boldsymbol{\delta}_{uv} = \begin{cases} \bar{\boldsymbol{\delta}}_{v-u} & \text{if } v > u, \\ \bar{\boldsymbol{\delta}}_{u-v} & \text{if } v < u, \end{cases} \\ \text{DIST3-COV} &: \boldsymbol{\delta}_{uv} = \begin{cases} \bar{\boldsymbol{\delta}}_{v-u} & \text{if } v > u, \\ -\bar{\boldsymbol{\delta}}_{u-v} & \text{if } v < u, \end{cases} \\ \text{TWOPAR-COV} &: \boldsymbol{\delta}_{uv} = \begin{cases} \bar{\boldsymbol{\delta}}_1 & \text{if } v > u, \\ \bar{\boldsymbol{\delta}}_{-1} & \text{if } v < u. \end{cases} \end{aligned}$$

In this case, we can also interpret the regression coefficients by recalling the corresponding interpretation for the intercept parameters. We further note that a constraint of a certain type for the intercepts may be combined with a different constraint on the regression coefficients.

### 3.3 Hierarchy among the constraints

A crucial point for model selection is the hierarchy among the constraints introduced above. In this regard, Fig. 3 shows the set of these constraints with the general ones on the left and those applicable with ordinal states on the right. Additionally, when a constraint is nested in another constraint, an arrow goes from the latter to the former. This scheme applies to both constraints formulated on the intercepts and on the regression coefficients of the transition (sub)model.

It is important to note that there is no complete nested structure, or in other words, no full ordering of these constraints. For instance, CONST is nested within any other constraint except DIFF and DIST3; SYMM, RSYMM, and DIST1 are not nested within any other constraint, but they include only some of the remaining ones; DIST2 and TWOPAR are nested within other constraints and, at the same time, include CONST.

The above conclusions lead us to adopt model selection criteria that can be applied even with nonnested constraints; see Sect. 4.3. For interpretability reasons, we adopt the convention that the constraint applied to the regression coefficients for the covariates must be nested within, or at most equal to, the constraint used for the intercepts. For instance, the CONST-COV constraint is not compatible with DIFF-INT and DIST3-INT, and so on.

## 4 Likelihood-based inference

Parameter estimation is based on maximization of the model log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}),$$

which involves the probability or density of the observed vectors of response variables, $f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT})$, and $\boldsymbol{\theta}$ is the vector of all model parameters that, for an MC model, includes

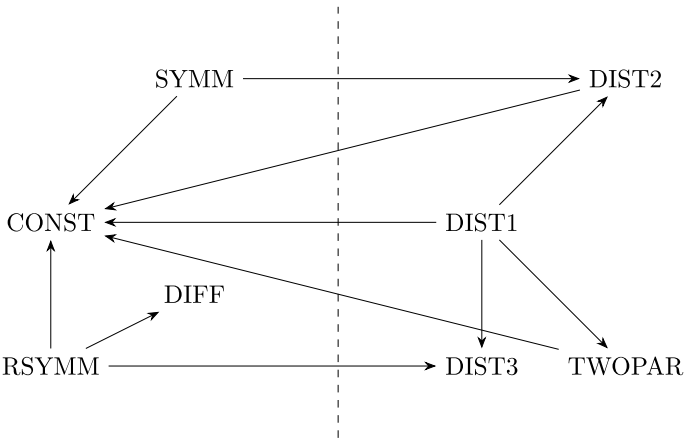


**Fig. 3** Graphical representation of the nested constraints with ordered ones on the right of the dashed line

parameters $\boldsymbol{\theta}_1$ for the initial probabilities and $\boldsymbol{\theta}_2$ for the transition probabilities, as defined in Sect. 2.1. This overall parameter vector will also include parameters $\boldsymbol{\theta}_3$ involved in the conditional distribution of the responses given the latent states for an HM model; see Sect. 2.2.

In the following, we discuss the maximization of $\ell(\boldsymbol{\theta})$ separately for the two classes of models of interest by the NR and the EM algorithm, respectively.

### 4.1 Markov chain model

Under the assumptions formulated in Sect. 2.1, we have that

$$f(\boldsymbol{y}_{i1},\ldots,\boldsymbol{y}_{iT}) = f(y_{i1},\ldots,y_{iT}) = \lambda_{i,y_{i1}} \prod_{t=2}^{T} \pi_{it,y_{i,t-1}y_{it}}.$$

Consequently, we can rewrite $\ell(\boldsymbol{\theta})$ as $\ell_1(\boldsymbol{\theta}_1) + \ell_2(\boldsymbol{\theta}_2)$, where the first component only involves the initial probabilities and the second involves the transition probabilities, while $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ contain the corresponding parameters. These two components may be written as

$$\begin{aligned}
\ell_1(\boldsymbol{\theta}_1) &= \sum_{i=1}^{n} \log \lambda_{i,y_{i1}} = \sum_{i=1}^{n} \boldsymbol{d}_{i1}' \log \boldsymbol{\lambda}_i, \\
\ell_2(\boldsymbol{\theta}_2) &= \sum_{i=1}^{n} \sum_{t=2}^{T} \log \pi_{it,y_{i,t-1}y_{it}} = \sum_{i=1}^{n} \sum_{t=2}^{T} \boldsymbol{d}_{it}' \log \boldsymbol{\pi}_{it,y_{i,t-1}},
\end{aligned}$$

where $\boldsymbol{d}_{it}$ is the indicator vector for $y_{it}$, that is, a column vector with $k$ elements equal to 0 with the exception of the $y_{it}$-th element equal to 1.

The corresponding score vectors have the following expressions:

$$\begin{aligned}
\boldsymbol{s}_1(\boldsymbol{\theta}_1) &= \sum_{i=1}^{n} \boldsymbol{X}_{i1}' \boldsymbol{G}' (\boldsymbol{d}_{i1} - \boldsymbol{\lambda}_i), \\
\boldsymbol{s}_2(\boldsymbol{\theta}_2) &= \boldsymbol{Z}' \begin{pmatrix} \dfrac{\partial \ell_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}_1'} & \cdots & \dfrac{\partial \ell_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}_k'} \end{pmatrix}',
\end{aligned}$$

where

$$\frac{\partial \ell_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}_u} = \sum_{i=1}^{n} \sum_{t=2}^{T} \boldsymbol{X}_{it}' \boldsymbol{H}_u' (\boldsymbol{d}_{it} - \boldsymbol{\pi}_{it,y_{i,t-1}}), \quad u = 1,\ldots,k.$$

Moreover, the corresponding information matrices have expression

$$\begin{aligned}
\boldsymbol{J}_1(\boldsymbol{\theta}_1) &= \sum_{i=1}^{n} \boldsymbol{X}_{i1}' \boldsymbol{G}' \boldsymbol{\Omega}_i \boldsymbol{G} \boldsymbol{X}_{i1}, \\
\boldsymbol{J}_2(\boldsymbol{\theta}_2) &= \boldsymbol{Z}' \mathrm{diag} \begin{pmatrix} \dfrac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}_1 \partial \boldsymbol{\eta}_1'} & \cdots & \dfrac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\eta}_k'} \end{pmatrix} \boldsymbol{Z},
\end{aligned}$$

where $\boldsymbol{\Omega}_i = \text{diag}(\boldsymbol{\lambda}_i) - \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'$,

$$\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}_u \partial \boldsymbol{\eta}_u'} = \sum_{i=1}^{n}\sum_{t=2}^{T} \boldsymbol{X}_{it}' \boldsymbol{H}_u' \boldsymbol{\Omega}_{it,u} \boldsymbol{H}_u \boldsymbol{X}_{it}, \quad u = 1, \ldots, k,$$

and $\boldsymbol{\Omega}_{it,u} = \text{diag}(\boldsymbol{\pi}_{it,u}) - \boldsymbol{\pi}_{it,u}\boldsymbol{\pi}_{it,u}'$.

On the basis of the score vectors $\boldsymbol{s}_1(\boldsymbol{\theta}_1)$ and $\boldsymbol{s}_2(\boldsymbol{\theta}_2)$ and the information matrices $\boldsymbol{J}_1(\boldsymbol{\theta}_1)$ and $\boldsymbol{J}_2(\boldsymbol{\theta}_2)$ it is possible to implement two NR algorithms that, in the usual way, maximize $\ell_1(\boldsymbol{\theta}_1)$ and $\ell_2(\boldsymbol{\theta}_2)$ separately. We just recall that each step of the algorithm consists in adding to the current parameter vector the inverse of the information matrix multiplied by the score vector; these steps are performed until a suitable convergence criterion is satisfied. From this iterative algorithm we obtain the ML estimates of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ denoted by $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$, respectively. The corresponding asymptotic standard errors can be obtained in the usual way from the diagonal elements of $\boldsymbol{J}_1(\hat{\boldsymbol{\theta}}_1)^{-1}$ and $\boldsymbol{J}_2(\hat{\boldsymbol{\theta}}_2)^{-1}$.

Regarding the ML likelihood estimator, the typical properties of consistency, asymptotic efficiency, and normality hold, provided that the model is identifiable as the number of sample units grows to infinity, while the number of time occasions remains fixed. The results can be proved in a rather standard way as we are essentially dealing with extended logistic models. Identifiability can be simply checked on the basis of the rank of the above information matrix for the model parameters.

### 4.2 Hidden Markov model

When an HM model for longitudinal data is assumed, using the notation formulated in Sect. 2.2, for the distribution of the response variables we have

$$f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}) = \sum_{u_1=1}^{k} \cdots \sum_{u_T=1}^{k} \lambda_{i,u_1} \prod_{t=2}^{T} \pi_{it,u_{t-1}u_t} \prod_{t=1}^{T} f(\boldsymbol{y}_{it}|u_t), \tag{9}$$

where $f(\boldsymbol{y}_{it}|u)$ is the conditional density or probability of $\boldsymbol{Y}_{it}$ given $U_{it} = u$. In order to compute $f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT})$ avoiding the sum in (9), which has a computational cost that exponentially increases in $T$, we rely on the Baum and Welch recursion (Baum & Petrie, 1966) that is described in Appendix B.

The model log-likelihood may be maximized by an EM algorithm (Dempster et al., 1977) that, as usual, is based on the complete data log-likelihood. The latter requires the knowledge of the latent state for each individual and time occasion in addition to the observed data, also called incomplete data, and may be expressed as the sum of the following three components:

$$\ell_1^*(\boldsymbol{\theta}_1) = \sum_{i=1}^{n}\sum_{u=1}^{k} a_{i1,u} \log \lambda_{i,u}, \tag{10}$$

$$\ell_2^*(\boldsymbol{\theta}_2) = \sum_{i=1}^{n}\sum_{t=2}^{T}\sum_{u=1}^{k}\sum_{v=1}^{k} b_{it,uv}\log \pi_{it,uv}, \tag{11}$$

$$\ell_3^*(\boldsymbol{\theta}_3) = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{u=1}^{k} a_{it,u}\log f(\boldsymbol{y}_{it}|u), \tag{12}$$

where $a_{it,u}$ is an indicator variable equal to 1 if $u_{it} = u$ and 0 otherwise, while $b_{it,uv}$ is an indicator variable equal to 1 if $u_{i,t-1} = u$ and $u_{it} = v$. The three functions above depend on distinct sub-vectors of parameters and this simplifies the estimation.

The EM algorithm is based on alternating two steps until convergence in the incomplete data log-likelihood $\ell(\boldsymbol{\theta})$. The E-step is based on computing the posterior expected value of the indicator variables in (10), (11), and (12) given the current value of the parameters and the observed data. In particular, these expected values correspond to posterior probabilities that may be obtained by the backward recursion reported in Appendix C and that are denoted as

$$\begin{aligned}
\hat{a}_{it,u} &= p(U_{it} = u|\boldsymbol{y}_{i1},\ldots,\boldsymbol{y}_{iT}), \quad t = 1,\ldots,T,\ u = 1,\ldots,k,\\
\hat{b}_{it,uv} &= p(U_{i,t-1} = u, U_{it} = v|\boldsymbol{y}_{i1},\ldots,\boldsymbol{y}_{iT}), \quad t = 2,\ldots,T,\ u,v = 1,\ldots,k,
\end{aligned}$$

for $i = 1,\ldots,n$. Once these expected values are substituted for the indicator variables in expressions (10), (11), and (12), the resulting expected values of the components of the complete data log-likelihood are denoted by $\hat{\ell}_1^*(\boldsymbol{\theta}_1)$, $\hat{\ell}_2^*(\boldsymbol{\theta}_2)$, and $\hat{\ell}_3^*(\boldsymbol{\theta}_3)$, respectively.

The M-step maximizes $\hat{\ell}_1^*(\boldsymbol{\theta}_1)$, $\hat{\ell}_2^*(\boldsymbol{\theta}_2)$, and $\hat{\ell}_3^*(\boldsymbol{\theta}_3)$ so as to update the three blocks of parameters. In particular, to maximize the first two complete log-likelihood functions, we rely on NR algorithms based on the corresponding score vectors and information matrices reported in Appendix D. Maximization of $\hat{\ell}_3^*(\boldsymbol{\theta}_3)$ depends on the specific measurement (sub)model that is assumed and, for certain models, it is based on explicit solutions. For instance, under assumption (4) for continuous responses, the means are directly updated as

$$\boldsymbol{\mu}_u = \frac{1}{\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{a}_{it,u}}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{a}_{it,u}\boldsymbol{y}_{it}, \quad u = 1,\ldots,k,$$

while the variance-covariance matrix is updated as

$$\boldsymbol{\Sigma} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{u=1}^{k}\hat{a}_{it,u}(\boldsymbol{y}_{it} - \boldsymbol{\mu}_u)(\boldsymbol{y}_{it} - \boldsymbol{\mu}_u)'.$$

For categorical data, the conditional response probabilities defined in (5) are updated as

$$\phi_{jy,u} = \frac{1}{\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{a}_{it,u}}\sum_{i=1}^{n}\sum_{t=1}^{T}\hat{a}_{it,u}I(y_{ijt} = y), \quad u = 1,\ldots,k,\ y = 1,\ldots,c_j,$$

for each response $j$, with $j = 1, \ldots, r$, where $I(\cdot)$ denotes the indicator function.

The two steps (E and M) are iterated until a suitable convergence criterion is satisfied, obtaining the ML estimates $\hat{\boldsymbol{\theta}}_1$, $\hat{\boldsymbol{\theta}}_2$, and $\hat{\boldsymbol{\theta}}_3$. One of the most used criteria relies on the change in log-likelihood between successive iterations. Specifically, letting $\boldsymbol{\theta}^{(m)}$ denote the parameter vector obtained after the $m$-th M-step, the algorithm stops when the relative change in the log-likelihood is not greater than a tolerance level $\epsilon > 0$, typically set to $10^{-10}$, that is, when

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

Differently from MC models, initialization of the estimation algorithm for an HM model is crucial as the log-likelihood $\ell(\boldsymbol{\theta})$ is usually multimodal. The typical solution to this problem consists in trying different starting values for the EM algorithm based on deterministic and random rules. Overall, for a given number of states, inference is based on the solution corresponding to the largest value of the log-likelihood at convergence, which is expect to correspond to the global optimum, although this is not guaranteed; see Maruotti and Punzo (2021) for an illustration of the alternative initialization strategies for HM models.

Once parameter estimates are computed for a given number of states, the corresponding standard errors may be obtained on the basis of different methods. As known, the EM algorithm does not directly provide either the observed or the expected information matrix. However, among the available approaches to obtain this matrix, one of the simplest is based on the numerical method proposed in Bartolucci and Farcomeni (2009), that is, as minus the numerical derivative of the score vector at convergence. The score vector is obtained, in turn, as the first derivative of the expected value of the complete data log-likelihood, as suggested by Oakes (1999).

Another method, which is typically preferred in the context of HM models, is based on a parametric or nonparametric bootstrap procedure (Davison & Hinkley, 1997) that consists in repeatedly drawing samples from the observed sample (nonparametric version), or from the estimated model (parametric version), and computing the maximum likelihood estimate for every bootstrap sample. Then, the standard errors are obtained by computing, in a suitable way, the standard deviation of the empirical distribution so obtained.

From the numerical point of view, it is well known that the EM algorithm may require a number of iterations larger than that required by an NR algorithm. However, the latter may be unstable and more difficult to implement as it does not use the log-likelihood referred to complete data. Moreover, as already mentioned, we implemented the EM algorithm also relying on `Fortran` language in order to improve its speed. Regarding the properties of the resulting ML estimates, under suitable conditions, consistency and asymptotic efficiency and normality hold. However, the theoretical framework is more complex than for MC models due to the presence of latent variables; for a precise description of this framework we refer to Bartolucci (2006). The main condition for these properties to hold is that of local identifiability, which can be again checked on the basis of the rank of the information matrix; for more details we refer to McHugh (1956), Goodman (1974), and Rothenberg (1971).

### 4.3 Model selection

When the number of states is given, model selection mainly concerns the constraints on the parameters of the transition (sub)model, as illustrated in Sect. 3. This situation clearly happens with univariate categorical responses when an MC model, or an HM model interpreted as an extension of an MC model accounting for measurement errors (Bartolucci et al., 2013), is adopted. In these cases the number of states is equal to the number of response categories. Apart from this specific situation, we base the selection of the optimal number of latent states of an HM model on common information criteria, such as AIC or BIC (Bacci et al., 2014). Given the number of states of the MC or HM model, we suggest a stepwise procedure for selecting the constraints on the parameters of the transition (sub)model. First, we select the constraint on the regression coefficients for the covariates, and then we select that on the intercepts. In this regard, we still use an information criterion, such as the BIC, because, as clarified in Sect. 3.3, there is not a full ordering of such constraints that instead follow a nonnested structure. Moreover, as already clarified, in order to favor interpretability of the selected model, we require that the constraint on the regression coefficients is nested in or, at most, equal to the constraint adopted on the intercepts.

For further clarification, Table 1 reports the number of free parameters for the intercept parameters of the transition (sub)model, under the different proposed constraints. Note that, to obtain the number of parameters referred to the regression coefficients, we need to multiply the corresponding number of intercept parameters by the number of covariates included in the model, which is denoted by $s$.

Finally, note that, although we mainly focus on constraints on the parameters for the transition probabilities, the constraint that the covariates do not affect the initial probabilities may be also adopted. In this regard, the stepwise procedure described above can be performed in both cases, with or without covariates affecting the initial probabilities, and then the best model in terms of BIC is selected.

## 5 Simulation study

In the following, we illustrate a Monte Carlo simulation study aimed at evaluating the behavior of the proposed approach. The study concerns both MC and HM models with covariates under two different constraints proposed in Sect. 3.

**Table 1** Number of intercept parameters of the multinomial logit model for the transition probabilities, under the proposed constraints. The name ALL stands for the unconstrained model

| Constraints | Number of parameters |
|---|---|
| ALL | $k(k-1)$ |
| CONST-INT | 1 |
| SYMM-INT | $k(k-1)/2$ |
| RSYMM-INT | $k(k-1)/2$ |
| DIFF-INT | $k-1$ |
| DIST1-INT | $2(k-1)$ |
| DIST2-INT | $(k-1)$ |
| DIST3-INT | $k-1$ |
| TWOPAR-INT | 2 |

### 5.1 Simulation design

We randomly generated $B = 100$ samples from the MC model with individual covariates, under different scenarios in terms of sample size, time occasions, and number of states: $n = 500, 1000$, $T = 5, 10$, and $k = 3, 4$, respectively. For each sample unit, we generated two covariates from an *AR*(1) process with autoregressive parameter equal to 0.5 and stationary variance equal to 1. Both covariates are included in the distribution of the initial and transition probabilities through the multinomial parameterizations described in assumptions (1) and (2).

For the parameters of the initial (sub)model, we let $\alpha_u = 0$ and $\boldsymbol{\beta}_u = (0.5, 1)'$, for $u \geq 2$. These parameters are collected in vector $\boldsymbol{\theta}_1$. About the transition probabilities, we assumed a model without constraints on the intercepts by setting the corresponding parameters equal to $\gamma_{uv} = -\log(0.9/0.1 + \omega)$, for $u, v = 1, \ldots, k$, $u \neq v$, where $\omega$ is a perturbation term generated from a uniform distribution between 0 and 1. This setting leads to a quite high level of persistence in the latent Markov chain. Moreover, for the constraints on the regression coefficients of the transition (sub)model, two different options are considered: CONST-COV and DIST2-COV. Under the first constraint we set $\bar{\boldsymbol{\delta}} = (0.5, 1)'$, whereas under the second we set $\boldsymbol{\delta}_1 = (0.5, 1)'$, $\boldsymbol{\delta}_2 = (0.25, 0.5)'$, when $k = 3$, and $\boldsymbol{\delta}_1 = (0.5, 1)'$, $\boldsymbol{\delta}_2 = (0.25, 0.5)'$, $\boldsymbol{\delta}_3 = (0, 0)'$, when $k = 4$. All these parameters are collected in vector $\boldsymbol{\theta}_2$.

The same settings are considered for the HM model with covariates. In such a case, we assumed a univariate model with a categorical response variable with a number of categories equal to the number of hidden states, that is, $c = k$. When $k = 3$ the corresponding conditional response probabilities are set equal to $\phi_{y,1} = (0.7, 0.2, 0.1)'$, $\phi_{y,2} = (0.3, 0.6, 0.1)'$, and $\phi_{y,3} = (0.2, 0.1, 0.7)'$. When $k = 4$ we let $\phi_{y,1} = (0.7, 0.1, 0.1, 0.1)'$, $\phi_{y,2} = (0.2, 0.6, 0.1, 0.1)'$, $\phi_{y,3} = (0.1, 0.1, 0.7, 0.1)'$, and $\phi_{y,4} = (0.05, 0.05, 0.1, 0.8)'$. In both cases, this choice of parameters leads to quite well separated states.

Finally, we evaluated the performance of our proposal in the presence of unbalanced longitudinal data, namely, when for every individual *i* we have a specific number of time occasions $T_i$. In particular, for both the MC and HM models, after generating a longitudinal sample that, as before, is referred to balanced data with *T* time occasions, we randomly selected half of the subjects and, for these subjects, we removed the last *T*/2 observations. In this case, only scenarios with a maximum of ten time points are considered (i.e., $\max(T_i) = 10$), ensuring that each individual retains a sufficient number of observations.

### 5.2 Results

Under every scenario, the simulation results are compared with those obtained by estimating the model from which data are generated, but without constraints on the regression coefficients of the transition (sub)model (named ALL). The results are evaluated, on the basis of the 100 randomly generated samples, in terms of median bias (Mbias) and median absolute error (MAE) of the estimates. These measures of performance are chosen for their robustness with respect to outliers. In fact, when we fitted the HM model without constraints, we experienced a quite high level of instability of the parameter estimates. This is due to the fact that, in some samples, the information matrix computed during the NR step, used within the EM algorithm to update the parameter estimates, is nearly singular, and the correspond-

**Table 2** Simulation results under the MC model: Average (over the latent states) of the absolute median bias (Mbias) and the median absolute error (MAE) of the estimators of the transition probability parameters, collected in $\hat{\boldsymbol{\theta}}_2$, under the CONST-COV constraint and under the unconstrained model (ALL)

| | | CONST-COV | | | | ALL | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $n = 500$ | | $n = 1000$ | | $n = 500$ | | $n = 1000$ | |
| | | \|Mbias\| | MAE | \|Mbias\| | MAE | \|Mbias\| | MAE | \|Mbias\| | MAE |
| $T = 5$ | $k = 3$ | 0.012 | 0.071 | 0.008 | 0.052 | 0.016 | 0.094 | 0.010 | 0.067 |
| | $k = 4$ | 0.012 | 0.098 | 0.011 | 0.071 | 0.015 | 0.116 | 0.011 | 0.083 |
| $T = 10$ | $k = 3$ | 0.007 | 0.051 | 0.009 | 0.036 | 0.010 | 0.059 | 0.008 | 0.045 |
| | $k = 4$ | 0.012 | 0.067 | 0.008 | 0.046 | 0.010 | 0.077 | 0.007 | 0.054 |
| $\max(T_i) = 10$ | $k = 3$ | 0.008 | 0.063 | 0.006 | 0.046 | 0.013 | 0.073 | 0.007 | 0.051 |
| | $k = 4$ | 0.013 | 0.075 | 0.013 | 0.054 | 0.016 | 0.093 | 0.013 | 0.064 |

The last two rows are referred to the case of unbalanced data, with $\max(T_i) = 10$

**Table 3** Simulation results under the MC model: Average (over the latent states) of the absolute median bias (Mbias) and the median absolute error (MAE) of the estimators of the transition probability parameters, collected in $\hat{\boldsymbol{\theta}}_2$, under the DIST2-COV constraint and under the unconstrained model (ALL)

| | | DIST2-COV | | | | ALL | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $n = 500$ | | $n = 1000$ | | $n = 500$ | | n = 1000 | |
| | | \|Mbias\| | MAE | \|Mbias\| | MAE | \|Mbias\| | MAE | \|Mbias\| | MAE |
| $T = 5$ | $k = 3$ | 0.014 | 0.077 | 0.006 | 0.059 | 0.015 | 0.091 | 0.010 | 0.069 |
| | $k = 4$ | 0.008 | 0.095 | 0.008 | 0.065 | 0.015 | 0.117 | 0.010 | 0.078 |
| $T = 10$ | $k = 3$ | 0.008 | 0.056 | 0.006 | 0.038 | 0.008 | 0.064 | 0.006 | 0.046 |
| | $k = 4$ | 0.009 | 0.062 | 0.007 | 0.044 | 0.012 | 0.072 | 0.005 | 0.052 |
| $\max(T_i) = 10$ | $k = 3$ | 0.006 | 0.059 | 0.008 | 0.050 | 0.012 | 0.071 | 0.008 | 0.056 |
| | $k = 4$ | 0.010 | 0.083 | 0.010 | 0.052 | 0.013 | 0.091 | 0.011 | 0.061 |

The last two rows are referred to the case of unbalanced data, with $\max(T_i) = 10$

ing transition probabilities are very close to 0. For such a reason, in showing the estimation results under the HM model, we also report the proportion of samples for which the parameter estimates are, in absolute value, greater than a specified threshold.

The simulation results obtained under the MC model are reported in Tables 2 and 3, with the first table referred to constraint CONST-COV and the second referred to constraint DIST2-COV. In estimating this model, there were no problems of instability of parameter estimates. Moreover, we observe that, regarding the estimation of the parameters of the transition (sub)model, the Mbias and MAE are relatively small under all scenarios. As expected, these measures tend to decrease as the sample size and the number of time occasions increase. On the other hand, the quality of results, especially in terms of MAE, gets slightly worse as the number of states increases and for the case of unbalanced longitudinal data. In almost all scenarios, the estimation results obtained under the CONST-COV constraint are better, in terms of accuracy of parameter estimates, than those obtained under the corresponding ALL model. A similar conclusion can be drawn with reference to constraint DIST2-COV.

Table 4 reports the average, over the latent states, of the Mbias and MAE of the parameter estimates, obtained under the HM model with the CONST-COV constraint on the regression parameter of the transition (sub)model, compared to those obtained under the full model (ALL). The table also reports the proportion of samples for which the absolute value of at

least one of the estimates of the parameters of interest is greater than the absolute value of the corresponding true parameter value plus a constant, which is set equal to 5. In Table 5, we report the same results referred to the DIST2-COV constraint.

From these tables, we observe that the median bias and the MAE may be quite large under those scenarios characterized by less information due to small sample size ($n = 500$) and number of time occasions ($T = 5$). When estimating the HM model without constraints, we notice that for these scenarios the proportion of samples leading to unstable estimates is quite large, ranging from 0.21, when $k = 3$, to 0.27, when $k = 4$. As expected, the quality of results improves when either *n* or *T* increases. Conversely, performance deteriorates in unbalanced scenarios, where the loss in the number of observations for certain individuals reduces efficiency and may introduce bias. Again, the results obtained by estimating the model with constraints (both CONST-COV and DIST2-COV) are generally better than those obtained under the unconstrained model (ALL). Moreover, the instability of the parameter estimates is much less pronounced under the constrained models, thus emphasizing the usefulness of our proposal.

Finally, in evaluating the performance of the proposed approach, we also consider the computational costs of ML estimation of the proposed models. In this regard, Table 6 reports the median computing time (in seconds), across the simulated samples, required for the estimation algorithm to reach convergence on a standard personal computer. The table is referred to both the MC and HM models, formulated according to the unconstrained specification and under the COST-COV constraint, considering the most interesting scenarios. We observe that the computing time is, on average, of a few seconds for all cases. However, estimating a constrained model is in general faster than estimating its unconstrained version.

**Table 4** Simulation results under the HM model: Average (over the latent states) of the absolute median bias (Mbias) and the median absolute error (MAE) of the estimators of the transition probability parameters, collected in $\hat{\boldsymbol{\theta}}_2$, under the CONST-COV constraint and under the unconstrained model (ALL)

| CONST-COV | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | $n = 500$ | | | $n = 1000$ | | |
| | | \|Mbias\| | MAE | prop | \|Mbias\| | MAE | prop |
| $T = 5$ | $k = 3$ | 0.062 | 0.323 | 0.000 | 0.056 | 0.230 | 0.000 |
| | $k = 4$ | 0.053 | 0.325 | 0.000 | 0.042 | 0.211 | 0.000 |
| $T = 10$ | $k = 3$ | 0.054 | 0.186 | 0.000 | 0.010 | 0.129 | 0.000 |
| | $k = 4$ | 0.021 | 0.170 | 0.000 | 0.014 | 0.126 | 0.000 |
| $\max(T_i) = 10$ | $k = 3$ | 0.064 | 0.222 | 0.000 | 0.016 | 0.153 | 0.000 |
| | $k = 4$ | 0.025 | 0.207 | 0.000 | 0.021 | 0.151 | 0.000 |
| ALL | | | | | | | |
| | | $n = 500$ | | | $n = 1000$ | | |
| | | \|Mbias\| | MAE | prop | \|Mbias\| | MAE | prop |
| $T = 5$ | $k = 3$ | 0.131 | 0.495 | 0.208 | 0.056 | 0.230 | 0.016 |
| | $k = 4$ | 0.103 | 0.440 | 0.274 | 0.085 | 0.287 | 0.037 |
| $T = 10$ | $k = 3$ | 0.078 | 0.240 | 0.014 | 0.019 | 0.154 | 0.000 |
| | $k = 4$ | 0.039 | 0.213 | 0.013 | 0.026 | 0.144 | 0.000 |
| $\max(T_i) = 10$ | $k = 3$ | 0.084 | 0.312 | 0.055 | 0.033 | 0.189 | 0.000 |
| | $k = 4$ | 0.088 | 0.286 | 0.084 | 0.032 | 0.176 | 0.000 |

The column “prop" reports the proportion of unstable samples; the last two rows are referred to the case of unbalanced data, with $\max(T_i) = 10$

**Table 5** Simulation results under the HM model: Average (over the latent states) of the absolute median bias (Mbias) and the median absolute error (MAE) of the estimators of the transition probability parameters, collected in $\hat{\boldsymbol{\theta}}_2$, under the DIST2-COV constraint and under the unconstrained model (ALL)

| DIST2-COV | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | $n = 500$ | | | $n = 1000$ | | |
| | | \|Mbias\| | MAE | prop | \|Mbias\| | MAE | prop |
| $T = 5$ | $k = 3$ | 0.075 | 0.371 | 0.020 | 0.031 | 0.231 | 0.000 |
| | $k = 4$ | 0.082 | 0.361 | 0.000 | 0.029 | 0.224 | 0.000 |
| $T = 10$ | $k = 3$ | 0.027 | 0.190 | 0.000 | 0.022 | 0.132 | 0.000 |
| | $k = 4$ | 0.026 | 0.179 | 0.000 | 0.019 | 0.119 | 0.000 |
| $\max(T_i) = 10$ | $k = 3$ | 0.044 | 0.248 | 0.000 | 0.031 | 0.161 | 0.000 |
| | $k = 4$ | 0.028 | 0.221 | 0.000 | 0.017 | 0.145 | 0.000 |
| ALL | | | | | | | |
| | | $n = 500$ | | | $n = 1000$ | | |
| | | \|Mbias\| | MAE | prop | \|Mbias\| | MAE | prop |
| $T = 5$ | $k = 3$ | 0.070 | 0.453 | 0.204 | 0.061 | 0.308 | 0.043 |
| | $k = 4$ | 0.112 | 0.428 | 0.272 | 0.052 | 0.275 | 0.088 |
| $T = 10$ | $k = 3$ | 0.028 | 0.229 | 0.014 | 0.020 | 0.148 | 0.000 |
| | $k = 4$ | 0.042 | 0.217 | 0.014 | 0.025 | 0.143 | 0.000 |
| $\max(T_i) = 10$ | $k = 3$ | 0.061 | 0.301 | 0.022 | 0.025 | 0.188 | 0.000 |
| | $k = 4$ | 0.058 | 0.270 | 0.052 | 0.034 | 0.176 | 0.000 |

The column "prop" reports the proportion of unstable samples; the last two rows are referred to the case of unbalanced data, with $\max(T_i) = 10$

**Table 6** Simulation results: Median computing time (in seconds) across the simulated samples required for the estimation algorithm to reach convergence on a standard personal computer, under the most informative scenarios

| | MC model | | | |
|---|---|---|---|---|
| | $n = 500$–$k = 3$ | $n = 1000$–$k = 3$ | $n = 500$–$k = 3$ | $n = 500$–$k = 4$ |
| | $T = 5$ | $T = 5$ | $T = 10$ | $T = 5$ |
| CONST-COV | 0.528 | 0.976 | 0.965 | 0.544 |
| ALL | 0.582 | 0.992 | 0.993 | 0.541 |
| | HM model | | | |
| | $n = 500$–$k = 3$ | $n = 1000$–$k = 3$ | $n = 500$–$k = 3$ | $n = 500$–$k = 4$ |
| | $T = 5$ | $T = 5$ | $T = 10$ | $T = 5$ |
| CONST-COV | 1.823 | 3.176 | 2.298 | 2.091 |
| ALL | 3.100 | 5.897 | 3.553 | 3.608 |

In particular, for the HM formulation, estimation of the proposed CONST-COV model is roughly 1.7 times faster than that of the unconstrained (ALL) model.

## 6 Empirical example

The HRS is conducted by the University of Michigan[1] and collects data related to retirement and health among elderly individuals (Fisher & Ryan, 2017). Monitoring perceived health across covariates remains an important topic in the literature; see Böckerman and Ilmak-

[1] More details on the study and questionnaire can be found at the website: http://hrsonline.isr.umich.edu/

unnas (2009) and Zajacova et al. (2017), among others. The sampling design is nationally representative of the American population aged over 50 years. The response variable is the self-reported health status (SRHS) measured on a scale based on five ordered categories: 'poor', 'fair', 'good', 'very good', 'excellent'. The sample includes $n = 7,074$ individuals interviewed at $T = 8$ approximately equally spaced occasions from 1992 to 2006. The available covariates are: gender (coded as 1 for male and 2 for female), race (coded as 1 for white, 2 for black, and 3 for others), educational level (coded as 1 for high school, 2 for general educational diploma, 3 for high school graduate, 4 for some college, 5 for college and above), and age (reported in years). These data are publicly available in the R package `LMest` (Bartolucci et al., 2017).

The marginal distribution of the response variable over the interviews is presented in Table 1 of the SI. The frequencies of the response categories "good" and "very good" are both around 30% and remain almost stable over time; at the first interview the 25.7% of individuals declares an excellent health status, but this value constantly decreases over time to reach around 10% at the last interview.

Table 2 of the SI shows the empirical transition matrix where each row reports the frequencies of the five response categories at occasion $t$ given the response at occasion $t-1$, with $t = 2, \ldots, 8$. The highest values of the off-diagonal elements of this matrix are for transitions from "poor" to "fair" (32.6%), from "fair" to "good" (27.2%), from "very good" to "good" (25.7%), and from "excellent" to "very good" (33.3%).

Table 3 of the SI reports the distribution of the time-fixed and time-varying covariates. We observe that 58.0% of the respondents are females, 82.9% are white, and 33.2% are college graduates. The average age at the first interview is 54.8 years, and at the last interview is 68.6 years. In the following sections, we illustrate the analysis of these data using the MC and HM models previously described.

### 6.1 Analysis with the Markov chain model

In order to choose a suitable MC model for the HRS data, we followed the stepwise procedure described in Sect. 4.3 by selecting a suitable constraint on the regression parameters of the transition (sub)model and then a suitable constraint on the intercepts of the same (sub) model. We considered two versions of the MC model, one with covariates affecting the initial probabilities, and the other without. Additionally, as mentioned in Sects. 3.3 and 4.3, the constraint on the regression coefficients is nested in or, at most, equal to the constraint adopted on the intercepts.

The results of this preliminary fitting are reported in Table 7 for the models with covariates affecting the initial probabilities, and in Table 8 for those with constant initial probabilities. For each model we show the maximum log-likelihood, the number of parameters, and the values of BIC and AIC indices. According to these results, the full model, which has no constraints on any component, as shown in the first row of Table 7 , has a relatively large number of parameters, equal to 216.

The maximum of the log-likelihood of the full model is − 65,150, with corresponding values of AIC and BIC equal to 130,731 and 132,214, respectively. Among the other MC models with different constraints on the regression parameters of the transition (sub)model, the one based on constraint of type DIST1-COV emerges as the best choice in terms of BIC. For this model, the number of free parameters decreases to 120 with a corresponding

**Table 7.** Maximum log-likelihood ($\ell(\hat{\boldsymbol{\theta}})$), number of parameters (*#par*), AIC and BIC under each MC model estimated with different constraints on the regression parameters for the transition probabilities (top panel) and the intercept parameters for these probabilities (bottom panel)

| Intercepts trans. prob. | Covariates init. prob. | Covariates trans. prob. | $\ell(\hat{\boldsymbol{\theta}})$ | *#par* | AIC | BIC |
|---|---|---|---|---|---|---|
| ALL | ALL | ALL | -65,149.50 | 216 | 130,731.00 | 132,213.66 |
| ALL | ALL | CONST | -66,064.80 | 64 | 132,257.60 | 132,696.91 |
| ALL | ALL | SYMM | -65,779.45 | 136 | 131,830.90 | 132,764.43 |
| ALL | ALL | RSYMM | -65,436.66 | 136 | 131,145.32 | 132,078.85 |
| ALL | ALL | DIFF | -65,462.44 | 88 | 131,100.88 | 131,704.93 |
| **ALL** | **ALL** | **DIST1** | **-65,312.90** | **120** | **130,865.81** | **131,689.51** |
| ALL | ALL | DIST2 | -65,875.55 | 88 | 131,927.10 | 132,531.14 |
| ALL | ALL | DIST3 | -65,521.77 | 88 | 131,219.55 | 131,823.59 |
| ALL | ALL | TWO | -65,585.93 | 72 | 131,315.86 | 131,810.08 |
| ALL | ALL | NONE | -66,733.53 | 24 | 133,515.05 | 133,679.79 |
| DIST1 | ALL | DIST1 | -66,919.11 | 108 | 134,054.23 | 134,795.56 |

The name ALL is used for models without constraints while NONE refers to the model without covariates; the list of model names is that of constraints introduced in Sect. 3

**Table 8.** Maximum log-likelihood ($\ell(\hat{\boldsymbol{\theta}})$), number of parameters (*#par*), AIC and BIC under each MC model estimated without covariates on the initial probabilities with different constraints on the regression parameters for the transition probabilities (top panel) and the intercept parameters for these probabilities (bottom panel)

| Intercepts trans. prob. | Covariates init. prob. | Covariates trans. prob. | $\ell(\hat{\boldsymbol{\theta}})$ | *#par* | AIC | BIC |
|---|---|---|---|---|---|---|
| ALL | NONE | ALL | -65,717.36 | 184 | **131,802.72** | 133,065.73 |
| ALL | NONE | CONST | -66,632.66 | 32 | 133,329.32 | 133,548.97 |
| ALL | NONE | SYMM | -66,347.31 | 104 | 132,902.62 | 133,616.50 |
| ALL | NONE | RSYMM | -66,004.52 | 104 | 132,217.04 | 132,930.91 |
| ALL | NONE | DIFF | -66,030.30 | 56 | 132,172.60 | 132,556.99 |
| ALL | NONE | **DIST1** | **-65,880.76** | **88** | 131,937.53 | **132,541.57** |
| ALL | NONE | DIST2 | -66,443.41 | 56 | 132,998.81 | 133,383.21 |
| ALL | NONE | DIST3 | -66,089.63 | 56 | 132,291.26 | 132,675.66 |
| ALL | NONE | TWO | -66,153.79 | 40 | 132,387.57 | 132,662.14 |
| ALL | NONE | NONE | -66,733.53 | 24 | 133,515.05 | 133,679.79 |
| DIST1 | NONE | DIST1 | -67,486.97 | 76 | 135,125.91 | 135,647.60 |

The name ALL is used for models without constraints while NONE refers to the model without covariates; the list of model names is that of constraints introduced in Sect. 3

decrease in the maximum log-likelihood to − 65,313. Additionally, AIC becomes 130,866, and BIC becomes 131,690. Then, in the second step, we considered possible constraints on the intercepts for the transition (sub)model (see the second panel in Table 7). According to the general method previously mentioned, we only consider DIST1-INT constraint on these intercepts. The resulting model has 108 free parameters, but both AIC and BIC increase, and then this model is not adopted.

According to the results in Table 8, AIC and BIC select two different models without covariates in the initial (sub)model. In particular, while the first criterion selects the model without constraints on the regression parameters of the transition (sub)model, the BIC selects the model based on constraint of type DIST1-COV on these parameters. We recall

that, with covariates in the initial (sub)model (see Table 7) we select the same constraint but the values of both indices (AIC and BIC) are higher. According to the BIC, no constraints can be reasonably included on the intercepts of the transition (sub)model.

Overall, we selected a model with a constraint of type DIST1-COV on the regression coefficients of the transition (sub)model and no other constraints on the parameters of the initial (sub)model. The parameter estimates under this model are reported in Tables 4, 5, and 6, of the SI. In particular, Table 4 presents the effects of the covariates on the initial probabilities, Table 5 reports the estimated intercepts of the transition (sub)model, and Table 6 displays the regression parameters of the same (sub)model. From these results we notice that, at the first interview, black people tend to belong to the first three states: the odds ratio of belonging to the 4th state, which corresponds to the best self-perceived health, for black versus white individuals is equal to $\exp(-0.557) = 0.573$, thus showing that black people report poorer self-rated health. Looking at the results reported in Table 6 of the SI, the first column shows that the estimates of the effects of different educational levels are negative and increase in absolute value: with higher education, there is a negative effect on the probability of moving from the 5th state (best health conditions) to the 1st state (worst health conditions). For example, for the highest level of education the estimated coefficient is –3.828, meaning that, *ceteris paribus*, the relative risk of transitioning from the 5th to the 1st state is $\exp(-3.828) = 0.022$. These findings underscore the significant influence of educational attainment on self-perceived health, in line with previous research highlighting the role of schooling in shaping health outcomes.

Finally, we show in Table 9 the average of the estimated initial and transition probabilities of the Markov chain. We note that, at the first interview, the average probability of feeling “very good" is 0.308 (4th state). Over time, the probability to transit from the 4th to the 3rd (“good") state is 0.268.

### 6.2 Analysis with the hidden Markov model

The second analysis uses an HM model with $k = 5$ latent states, equal to the number of categories of the response variable. In this setup, the model can be viewed as an MC model where the states are not directly observable, and the association between responses and latent states is also taken into account. Since perceptions are based on subjective measurements, they are properly considered as not directly observable (latent concepts) under the HM model. Consequently, this model provides a more precise representation of the phenomena under investigation, accommodating potential measurement errors arising from varied interpretations of the response scale in the survey.

**Table 9** Average estimates of the initial probabilities ($\hat{\hat{\lambda}}_u$) and transition probabilities ($\hat{\hat{\pi}}_{uv}$) under the MC model with DIST1-COV constraint and $k = 5$ states

| | | $\hat{\hat{\pi}}_{uv}$ | | | | |
|---|---|---|---|---|---|---|
| $u$ | $\hat{\hat{\lambda}}_u$ | $v = 1$ | $v = 2$ | $v = 3$ | $v = 4$ | $v = 5$ |
| 1 | 0.047 | 0.516 | 0.360 | 0.091 | 0.022 | 0.011 |
| 2 | 0.115 | 0.109 | 0.504 | 0.294 | 0.078 | 0.015 |
| 3 | 0.272 | 0.024 | 0.161 | 0.533 | 0.240 | 0.041 |
| 4 | 0.308 | 0.010 | 0.056 | 0.268 | 0.542 | 0.124 |
| 5 | 0.257 | 0.006 | 0.029 | 0.132 | 0.345 | 0.489 |

Following an approach similar to that used for the MC model, we fitted a series of HM models with various constraints on the parameters for the initial and the transition (sub) models. The results are presented in Table 10 and led us to choose the model with constraint of type CONST-COV on the regression coefficients of the second (sub)model. These results, reported in Table 10, can be compared, in terms of information criteria, with those reported in Table 7, which pertain to the MC model. Considering the BIC, the HM models always provide lower values, indicating that a more comprehensive explanation of the phenomenon can be achieved with more complex models.

In Table 11 we report results of the HM model estimated without covariates on the initial probabilities. Considering the BIC, the HM model with constraint of type CONST-COV is selected; however, the BIC value is higher than that of the best model in Table 10. Table 12 reports the estimated conditional response probabilities and their cumulative values expressed as in equation (5) for the selected HM model with covariates also on the initial (sub)model and constraint of type CONST-COV. The five subpopulations identify individuals differing in perceived health: those with poor and fair perceived health are in the 1st and 2nd latent states, while individuals with a perception of fair and good health are in the 3rd state, and those with good and very good perceived health are in the 4th and 5th states. The cumulative probabilities reported in the right panel of Table 12 help to interpret the five states. These parameter estimates are valuable as they also offer insight into the impact of measurement errors, explaining the differences between the observed response and the estimated latent perception (hidden state). Table 7 of the SI shows the estimated standard errors for the estimates presented in Table 12 (left panel) obtained with $B$=200 bootstrap samples. Additionally, Table 13 shows the estimated average initial and transition probabilities of the HM model. While the initial probabilities are nearly identical between the selected MC and HM models, we notice that the transition matrix reported in Table 13 is much more

**Table 10.** Maximum log-likelihood ($\ell(\hat{\boldsymbol{\theta}})$), number of parameters (*#par*), AIC and BIC under each HM model estimated with different constraints on the regression parameters for the transition probabilities (top panel) and the intercept parameters for these probabilities (bottom panel)

| Intercepts trans. prob. | Covariates init. prob. | Covariates trans. prob. | $\ell(\hat{\boldsymbol{\theta}})$ | *#par* | AIC | BIC |
|---|---|---|---|---|---|---|
| ALL | ALL | ALL | -62,124.69 | 236 | 124,721.38 | 126,341.33 |
| ALL | ALL | **CONST** | **-62,318.94** | **84** | **124,805.87** | **125,382.46** |
| ALL | ALL | SYMM | -62,189.23 | 156 | 124,690.46 | 125,761.27 |
| ALL | ALL | RSYMM | -62,237.85 | 156 | 124,787.70 | 125,858.51 |
| ALL | ALL | DIFF | -62,325.06 | 108 | 124,866.12 | 125,607.45 |
| ALL | ALL | DIST1 | -62,221.99 | 140 | 124,723.98 | 125,684.96 |
| ALL | ALL | DIST2 | -62,247.49 | 108 | 124,710.99 | 125,452.32 |
| ALL | ALL | DIST3 | -62,305.72 | 108 | 124,827.43 | 125,568.76 |
| ALL | ALL | TWO | -62,306.94 | 92 | 124,797.88 | 125,429.38 |
| ALL | ALL | NONE | -63,153.92 | 44 | 126,395.84 | 126,697.86 |
| CONST | ALL | CONST | -63,704.33 | 65 | 127,538.66 | 127,984.84 |
| SYMM | ALL | CONST | -63,115.77 | 74 | 126,379.53 | 126,887.48 |
| RSYMM | ALL | CONST | -62,692.03 | 74 | 125,532.06 | 126,040.01 |
| DIST1 | ALL | CONST | -62,651.74 | 72 | 125,447.48 | 125,941.71 |
| DIST2 | ALL | CONST | -63,130.39 | 68 | 126,396.78 | 126,863.54 |
| TWO | ALL | CONST | -63,340.49 | 66 | 126,812.98 | 127,266.01 |

The name ALL is used for models without constraints while NONE refers to the model without covariates; the list of model names is that of constraints introduced in Sect. 3

**Table 11.** Maximum log-likelihood ($\ell(\hat{\boldsymbol{\theta}})$), number of parameters (*#par*), AIC and BIC under each HM model estimated without covariates on the initial probabilities with different constraints on the regression parameters for the transition probabilities (top panel) and the intercept parameters for these probabilities (bottom panel)

| Intercepts trans. prob. | Covariates init. prob. | Covariates trans. prob. | $\ell(\hat{\boldsymbol{\theta}})$ | *#par* | AIC | BIC |
|---|---|---|---|---|---|---|
| ALL | NONE | ALL | -62,730.95 | 204 | 125,869.90 | 127,270.19 |
| ALL | NONE | **CONST** | **-62,961.24** | **52** | **126,026.47** | **126,383.41** |
| ALL | NONE | SYMM | -62,815.04 | 124 | 125,878.08 | 126,729.23 |
| ALL | NONE | RSYMM | -62,839.97 | 124 | 125,927.95 | 126,779.11 |
| ALL | NONE | DIFF | -62,921.62 | 76 | 125,995.24 | 126,516.91 |
| ALL | NONE | DIST1 | -62,835.69 | 108 | 125,887.38 | 126,628.71 |
| ALL | NONE | DIST2 | -62,877.13 | 76 | 125,906.26 | 126,427.94 |
| ALL | NONE | DIST3 | -62,912.95 | 76 | 125,977.90 | 126,499.58 |
| ALL | NONE | TWO | -62,937.08 | 60 | 125,994.16 | 126,406.01 |
| ALL | NONE | NONE | -63,153.92 | 44 | 126,395.84 | 126,697.86 |
| CONST | NONE | CONST | -64,378.35 | 33 | 128,822.70 | 129,049.22 |
| SYMM | NONE | CONST | -63,811.09 | 42 | 127,706.18 | 127,994.48 |
| RSYMM | NONE | CONST | -63,327.75 | 42 | 126,739.49 | 127,027.79 |
| DIST1 | NONE | **CONST** | **-63,325.71** | **40** | **126,731.42** | 127,005.98 |
| DIST2 | NONE | CONST | -63,830.64 | 36 | 127,733.29 | 127,980.40 |
| TWO | NONE | CONST | -63,961.69 | 34 | 127,991.38 | 128,224.76 |

The name ALL is used for models without constraints while NONE refers to the model without covariates; the list of model names is that of constraints introduced in Sect. 3

**Table 12** Estimated conditional probabilities of SRHS (left panel) and cumulative values (right panel) under the HM model with constraint CONST-COV and $k = 5$ hidden states

| | Conditional response | | | | | Conditional cumulative response | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Probabilities ($\hat{\phi}_{1y,u}$) | | | | | Probabilities ($\hat{\psi}_{1y,u}$) | | | | |
| Category ($y$) | $u$=1 | $u$=2 | $u$=3 | $u$=4 | $u$=5 | $u$=1 | $u$=2 | $u$=3 | $u$=4 | $u$=5 |
| Poor | 0.71 | 0.06 | 0.00 | 0.00 | 0.00 | 0.71 | 0.06 | 0.00 | 0.00 | 0.00 |
| Fair | 0.26 | 0.68 | 0.09 | 0.01 | 0.00 | 0.97 | 0.74 | 0.09 | 0.01 | 0.00 |
| Good | 0.02 | 0.22 | 0.71 | 0.15 | 0.03 | 0.99 | 0.96 | 0.80 | 0.16 | 0.03 |
| Very good | 0.01 | 0.03 | 0.18 | 0.75 | 0.19 | 1.00 | 0.99 | 0.98 | 0.91 | 0.22 |
| Excellent | 0.00 | 0.01 | 0.02 | 0.09 | 0.78 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

persistent than that reported in Table 9. This suggests that, over time, most individuals tend to perceive their health as remaining relatively stable. The most likely transition (0.189) is estimated from “excellent" (5th hidden state) to “very good" (4th hidden state).

Table 8 in the SI file illustrates the impact of the covariates on the initial probabilities. Significance is established according to a parametric bootstrap procedure. Notably, during the initial interview, individuals with graduate and college degrees tend to be associated with states other than the 1st. We also observed that the odds ratio to belong to the 5th state of people with college and above is equal to $\exp(4.554) = 95.012$, thus showing that attaining higher levels of education also guarantees better health conditions, as well-documented in numerous studies (Ross & Wu, 1995). Table 9 in the SI file presents the estimates of the intercepts on the transition (sub)model. Table 10 in the SI file shows the estimates of the covariates effects on the transition probabilities based on the imposed constraints. We notice

**Table 13** Average initial and transition probabilities under the HM model with constraint CONST-COV and $k = 5$ hidden states

| | | $\hat{\hat{\pi}}_{uv}$ | | | | |
|---|---|---|---|---|---|---|
| $u$ | $\hat{\hat{\lambda}}_u$ | $v = 1$ | $v = 2$ | $v = 3$ | $v = 4$ | $v = 5$ |
| 1 | 0.050 | 0.933 | 0.062 | 0.000 | 0.005 | 0.000 |
| 2 | 0.115 | 0.043 | 0.911 | 0.041 | 0.003 | 0.001 |
| 3 | 0.274 | 0.007 | 0.073 | 0.890 | 0.028 | 0.003 |
| 4 | 0.289 | 0.003 | 0.008 | 0.114 | 0.859 | 0.016 |
| 5 | 0.272 | 0.000 | 0.004 | 0.019 | 0.189 | 0.788 |

that the estimated effect of gender is negative, suggesting that females tend to exhibit more persistence in the same state over time compared to males. We also observe that the odds ratio for people with a college degree or higher is $\exp(-1.036) = 0.355$, suggesting that individuals with higher education are less likely to transition out of the same latent state compared to those with only a high school education, holding the other covariates constant.

## 7 Conclusions

In this paper, we propose a set of constraints on the parameters for the transition probabilities of Markov Chain (MC) and Hidden Markov (HM) models for longitudinal data with covariates. These constraints are particularly useful for making the model more parsimonious compared to standard formulations, thereby enhancing interpretability and numerical stability in performing maximum likelihood estimation, especially when the model involves several latent states. We also develop iterative algorithms for estimating the proposed models, which retain the same level of complexity as the corresponding algorithms for standard MC and HM models with covariates, since the constraints under consideration are linear. This ensures that the computational complexity of the proposed approach remains manageable, and may even allow for the introduction of additional constraints beyond those explicitly considered in this paper.

The simulation study allows us to evaluate the usefulness of the proposed constrained model formulations. In almost all simulated scenarios, we observe a higher accuracy of the parameters estimates obtained under the constrained models with respect to the unconstrained ones. Moreover, in certain simulated samples, we note a numerical instability of the estimation process of the HM model without constraints. This problem does not arise under the constrained models.

The approach is illustrated through an application involving self-rated health data obtained from the Health and Retirement Study. This study entails interviewing a representative panel of individuals from 1992 to 2006, during which several individual covariates are also collected. The primary focus is on assessing differences among the perceived health conditions based on these covariates. This application provides a useful example for comparing the MC and HM models, and also for comparing the unconstrained version of both models with their constrained counterparts. Given that we have one ordinal response variable, we adopt a number of states in the HM analysis equal to the number of categories of this variable. In particular, we observe that by employing a much more parsimonious HM model compared to the complete HM model with the usual parameterization we achieve a significant improved fit. Furthermore, the selected HM model outperforms the correspond-

ing MC model. The proposed constraints in the MC and HM models can be particularly useful in reducing the complexity of parameter interpretation while preserving the necessary explanatory power. In this work, we also show that the instability of the parameter estimates is considerably mitigated under the constrained models, further highlighting the practical value of these constraints. The proposed constraints may prove particularly useful in a variety of applied contexts, such as in life course studies (Bolano et al., 2016), where longitudinal categorical data are used to summarize pathways across different dimensions of the life course and to relate them to demographic or concomitant variables.

In the paper, we briefly discuss the formulation of HM models for continuous response variables. Currently, the code has been implemented for models with categorical response variables, and future work will focus on extending the proposed formulations to HM models for continuous responses, both under homoskedasticity and heteroskedasticity, with and without covariates. The development of this code will be addressed in future research.

## Appendix A: Design matrix for the constraints

Depending on the constraint on the intercepts $\gamma_{uv}$ we have the following structure for matrix $\boldsymbol{Z}_1$ in (8):

- CONST-INT: $\boldsymbol{Z}_1 = \mathbf{1}_{k(k-1)}$;
- SYMM-INT: $\boldsymbol{Z}_1$ has $k(k-1)/2$ columns with each column having all elements equal to 0 apart from two suitably selected elements equal to 1;
- RSYMM-INT: $\boldsymbol{Z}_1$ has $k(k-1)/2$ columns with each column having all elements equal to 0 apart from two suitably selected elements equal to 1 and -1;
- DIFF-INT: $\boldsymbol{Z}_1$ has $k-1$ columns with each column having all elements equal to 0 apart from $k-1$ elements equal to -1 and the same number of elements equal to 1;
- DIST1-INT: $\boldsymbol{Z}_1$ has $2(k-1)$ columns with each column having elements equal to 0 apart from certain elements equal to 1; the number of these elements goes from 1 to $k-1$;
- DIST2-INT: $\boldsymbol{Z}_1$ has $k-1$ columns with each column having all elements equal to 0 apart from certain elements equal to 1; the number of these elements goes from 2 to $2(k-1)$;
- DIST3-INT: $\boldsymbol{Z}_1$ has $k-1$ columns with each column having all elements equal to 0 apart from some elements equal to 1 and -1; the number of these elements goes from 2 to $2(k-1)$;
- TWOPAR-INT: $\boldsymbol{Z}_1$ had 2 columns both having all elements equal to 0 apart from $k(k-1)/2$ elements equal to 1. Depending on the constraint on the regression parameters $\boldsymbol{\delta}_{uv}$, $\boldsymbol{Z}_2$ has the same structure as $\boldsymbol{Z}_1$ above.

The following are examples of matrix $\boldsymbol{Z}_1$ (and $\boldsymbol{Z}_2$) depending on the constraints assumed on the parameters for the transition probabilities of an MC or HM model with $k = 4$ states:

CONST

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

SYMM

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

RSYMM

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}$$

DIFF

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$$

DIST1

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

DIST2

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

DIST3

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{pmatrix}$$

TWO

$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}$$

## Appendix B: Forward recursion

We have that

$$f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}) = \sum_{u=1}^{k} q_{iT,u},$$

where $q_{iT,u}$ is obtained by the forward recursion

$$q_{it,v} = f(\boldsymbol{y}_{it}|v) \sum_{u=1}^{k} q_{i,t-1,u} \pi_{it,uv}, \quad v = 1, \ldots, k,$$

initialized with $q_{i1,u} = \lambda_{i,u} f(\boldsymbol{y}_{i1}|u)$.

## Appendix C: Backward recursion and posterior probabilities

First of all we have that

$$p(U_{it} = u|\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}) = \frac{q_{it,u} q^*_{it,u}}{f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT})}, \quad t = 1, \ldots, T, \; u = 1, \ldots, k,$$

where $q^*_{it,u}$ is based on the backward recursion

$$q^*_{it,u} = \sum_{v=1}^{k} \pi_{i,t+1,uv} f(\boldsymbol{y}_{i,t+1}|v) q^*_{i,t+1,v}, \quad t = T-1, \ldots, 1,$$

initialized with $q^*_{iT,u} = 1$ for $u = 1, \ldots, k$. Moreover, we have that

$$p(U_{i,t-1} = u, U_{it} = v|\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}) = \frac{q_{i,t-1,u}\pi_{it,uv} f(\boldsymbol{y}_{it}|v) q^*_{it,v}}{f(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT})}, \quad u, v = 1, \ldots, k,$$

for $t = 2, \ldots, T$.

## Appendix D: Score vectors and information matrices of the complete log-likelihood functions

The score vectors for $\hat{\ell}^*_1(\boldsymbol{\theta}_1)$ and $\hat{\ell}^*_2(\boldsymbol{\theta}_2)$ have the following expressions:

$$\begin{aligned}
\boldsymbol{s}_1(\boldsymbol{\theta}_1) =& \sum_{i=1}^{n} \boldsymbol{X}'_{i1} \boldsymbol{G}' (\hat{\boldsymbol{a}}_{i1} - \boldsymbol{\lambda}_i), \\
\boldsymbol{s}_2(\boldsymbol{\theta}_2) =& \boldsymbol{Z}' \begin{pmatrix} \dfrac{\partial \hat{\ell}^*_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}'_1} & \cdots & \dfrac{\partial \hat{\ell}^*_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}'_k} \end{pmatrix}',
\end{aligned}$$

where

$$\frac{\partial \hat{\ell}^*_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}_u} = \sum_{i=1}^{n} \sum_{t=2}^{T} \boldsymbol{X}'_{it} \boldsymbol{H}'_u (\hat{\boldsymbol{b}}_{it,u} - \hat{b}_{it,u+} \boldsymbol{\pi}_{it,u}), \quad u = 1, \ldots, k,$$

with $\hat{\boldsymbol{a}}_{it}$ denoting the column vector with elements $\hat{a}_{it,u}$, $u = 1, \ldots, k$, and $\hat{\boldsymbol{b}}_{it,u}$ that with elements $\hat{b}_{it,uv}$, $v = 1, \ldots, k$, with $\hat{b}_{it,u+}$ being the sum of these elements.

The corresponding information matrices have expression

$$\begin{aligned}
\boldsymbol{J}^*_1(\boldsymbol{\theta}_1) =& \sum_{i=1}^{n} \boldsymbol{X}'_{i1} \boldsymbol{G}' \boldsymbol{\Omega}_i \boldsymbol{G} \boldsymbol{X}'_{i1}, \\
\boldsymbol{J}^*_2(\boldsymbol{\theta}_2) =& \boldsymbol{Z}' \mathrm{diag} \begin{pmatrix} \dfrac{\partial^2 \hat{\ell}^*_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}_1 \partial \boldsymbol{\eta}'_1} & \cdots & \dfrac{\partial^2 \hat{\ell}^*_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\eta}'_k} \end{pmatrix} \boldsymbol{Z},
\end{aligned}$$

where $\boldsymbol{\Omega}_i = \mathrm{diag}(\boldsymbol{\lambda}_i) - \boldsymbol{\lambda}_i \boldsymbol{\lambda}'_i$ and

$$\frac{\partial^2 \hat{\ell}^*_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\eta}_u \partial \boldsymbol{\eta}'_u} = \sum_{i=1}^{n} \sum_{t=2}^{T} \sum_{u=1}^{k} \hat{b}_{it,u+} \boldsymbol{\pi}_{it,u} \boldsymbol{X}'_{it} \boldsymbol{H}'_u \boldsymbol{\Omega}_{it,u} \boldsymbol{H}_u \boldsymbol{X}_{it}, \quad u = 1, \ldots, k,$$

and $\boldsymbol{\Omega}_{it,u} = \mathrm{diag}(\boldsymbol{\pi}_{it,u}) - \boldsymbol{\pi}_{it,u} \boldsymbol{\pi}'_{it,u}$.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10479-025-06986-x.

**Acknowledgements** We acknowledge the financial support from the grant "Hidden Markov Models for Early Warning Systems" of Ministero dell'Università e della Ricerca (PRIN 2022 2022TZEXKF) funded by European Union - Next Generation EU, Mission 4, Component 2, CUP J53D23004990006.

## Declarations

**Conflict of interest** The authors declare that they have no Conflict of interest.

**Ethical Approval** This article does not contain any studies with human participants or animals performed by any of the authors.



## References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pp. 267–281. Akademinai Kiado.

Anderson, T. (1954). Probability models for analyzing time changes in attitudes. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social science* (pp. 17–66). New York; Free Press.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika, 81*, 767–775.

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.

Bartolucci, F. (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.

Bartolucci, F., Bacci, S., & Pennoni, F. (2014). Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *Journal of the Royal Statistical Society Series C: Applied Statistics, 63*, 267–288.

Bartolucci, F., & Farcomeni, A. (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.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2013). *Latent Markov models for longitudinal data*. 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., & Forcina, A. (2002). Extended RC association models allowing for order restrictions and marginal modeling. *Journal of the American Statistical Association, 97*, 1192–1199.

Bartolucci, F., Lupparelli, M., & Montanari, G. E. (2009). Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *The Annals of Applied Statistics, 3*, 611–636.

Bartolucci, F., Montanari, G. E., & Pandolfi, S. (2015). Three-step estimation of latent Markov models with covariates. *Computational Statistics & Data Analysis, 83*, 287–301.

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.
Bartolucci, F., Pennoni, F., & Lupparelli, M. (2008). Likelihood inference for the latent Markov rasch model. In C. Huber, N. Limnios, M. Mesbah, & M. Nikulin (Eds.), *Mathematical methods for survival analysis, reliability and quality of life* (pp. 239–254). Wiley.
Baum, L., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics, 37*, 1554–1563.
Böckerman, P., & Ilmakunnas, P. (2009). Unemployment and self-assessed health: Evidence from panel data. *Health Economics, 18*, 161–179.
Bolano, D., Berchtold, A., & Ritschard, G. (2016). A discussion on hidden Markov models for life course data. in G. Ritschard and M. Studer (Eds.), *Proceedings of the International Conference on Sequence Analysis and Related Methods*, Lausanne, Switzerland, pp. 241–260.
Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application, 3*, 297–321.
Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. 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.
Di Mari, R., Oberski, D. L., & Vermunt, J. K. (2016). Bias-adjusted three-step latent Markov modeling with covariates. *Structural Equation Modeling: A Multidisciplinary Journal, 23*, 649–660.
Fisher, G. G., & Ryan, L. H. (2017). Overview of the health and retirement study and introduction to the special issue. *Work, Aging and Retirement, 4*, 1–9.
Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231.
Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association, 74*, 537–552.
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.
Juster, F. T., & Suzman, R. (1995). An overview of the health and retirement study. *Journal of Human Resources, 18*, S7–S56.
Koki, C., Meligkotsidou, L., & Vrontos, I. (2020). Forecasting under model uncertainty: Non-homogeneous hidden Markov models with pòlya-gamma data augmentation. *Journal of Forecasting, 39*, 580–598.
Maruotti, A., & Punzo, A. (2021). Initialization of hidden Markov and semi-Markov models: A critical evaluation of several strategies. *International Statistical Review, 89*, 447–480.
Maruotti, A., & Rocci, R. (2012). A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. *Statistics in Medicine, 31*, 871–886.
McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika, 21*, 331–347.
McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. Wiley.
Meligkotsidou, L., & Dellaportas, P. (2011). Forecasting with non-homogeneous hidden Markov models. *Statistics in Medicine, 21*, 439–449.
Meyn, S. P., & Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.
Mor, B., Garhwal, S., & Kumar, A. (2021). A systematic review of hidden Markov models and their applications. *Archives of Computational Methods in Engineering, 28*, 1429–1448.
Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 61*, 479–482.
Pennoni, F., Pandolfi, S., & Bartolucci, F. (2025). LMest: An R package for estimating generalized latent Markov models. *R Journal, 16*, 74–101.
R Core Team 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
Ross, C. E., & Wu, C. L. (1995). The links between education and health. *American Sociological Review, 60*, 719–745.
Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica, 39*, 577–591.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.
Spezia, L. (2006). Bayesian analysis of non-homogeneous hidden Markov models. *Journal of Statistical Computation and Simulation, 76*, 713–725.
Vermunt, J. K. (2025). Stepwise estimation of latent variable models: An overview of approaches. *Statistical Modelling, 25, 530-551.*

Wang, E. T., Chiang, S., Haneef, Z., Rao, V. R., Moss, R., & Vannucci, M. (2023). Bayesian non-homogeneous hidden Markov model with variable selection for investigating drivers of seizure risk cycling. *The Annals of Applied Statistics, 17*, 333.

Zajacova, A., Huzurbazar, S., & Todd, M. (2017). Gender and the structure of self-rated health across the adult life span. *Social Science & Medicine, 187*, 58–66.

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov models for time series: An introduction using R*. CRC Press.