Advances in Data Analysis and Classification (2022) 16:457–485
https://doi.org/10.1007/s11634-021-00446-7

REGULAR ARTICLE

Model-based two-way clustering of second-level units in ordinal multilevel latent Markov models

Giorgio Eduardo Montanari[1] · **Marco Doretti**[1] · **Maria Francesca Marino**[2]

Received: 29 September 2020 / Revised: 26 May 2021 / Accepted: 2 June 2021 / Published online: 15 June 2021

Abstract
In this paper, an ordinal multilevel latent Markov model based on separate random effects is proposed. In detail, two distinct second-level discrete effects are considered in the model, one affecting the initial probability vector and the other affecting the transition probability matrix of the first-level ordinal latent Markov process. To model these separate effects, we consider a bi-dimensional mixture specification that allows to avoid unverifiable assumptions on the random effect distribution and to derive a two-way clustering of second-level units. Starting from a general model where the two random effects are dependent, we also obtain the independence model as a special case. The proposal is applied to data on the physical health status of a sample of elderly residents grouped into nursing homes. A simulation study assessing the performance of the proposal is also included.

Keywords Latent Markov model · Multilevel modeling · Nursing home · Random effect separation

Mathematics Subject Classification 62H30

1 Introduction

Latent Markov models are a well-established tool for the analysis of longitudinal categorical data. In particular, this class of models is well tailored to settings where some repeatedly measured categorical variables are likely to be simultaneously influenced by an underlying latent trait of interest, which is assumed to be also a categorical

✉ Giorgio Eduardo Montanari
giorgio.montanari@unipg.it

1 Department of Political Sciences, University of Perugia, 06123 Perugia, Italy

2 Department of Statistics, Computer Science and Applications "G. Parenti", University of Florence, 50134 Florence, Italy

variable that probabilistically evolves over time according to a first-order Markov chain (Wiggins 1973; Bartolucci et al. 2013).

In the last decades, numerous extensions of the basic latent Markov model have been proposed to incorporate additional model features. These include multilevel modeling, which is in order when sample units are grouped and the number of groups is considerable. Examples of applications of multilevel latent Markov models (MLMMs) concern longitudinal datasets of individuals grouped in households (e.g., Koukounari et al. 2013), students in classes (e.g., Bartolucci et al. 2011) or workers in firms (e.g., Bartolucci and Lupparelli 2016). Hereafter, we will refer to such groups as to second-level units (SLUs). In these settings, the multilevel structure of the data is typically accounted for by introducing a second-level latent variable affecting the observed responses coming from the same group—either directly or through the first-level latent Markov chain—via a set of random effects. This approach is somewhat related to other proposals, which also make use of continuous or discrete random effects in latent Markov models, though not in a formal multilevel setting (Altman 2007; Maruotti and Rydén 2009; Maruotti and Rocci 2012; Marino and Alfò 2016; Marino et al. 2018).

In the multilevel approaches mentioned before, the second-level latent variable is discrete with a fixed number of states, resulting in a clustering of SLUs. In particular, in Bartolucci and Lupparelli (2016) such a clustering structure is assumed to be time-varying and to directly affect the distribution of the observed responses, together with some covariates and the first-level latent process. In the other multilevel approaches, SLU clustering is assumed to be time-invariant and is considered as a determinant—possibly with some covariates—of the first-level latent process, which in turn has a probabilistic influence on the observed categorical variables. In this case, the first-level process can be interpreted as a (possibly multidimensional) underlying trait the outcome variables are a measure of.

When SLU clustering is assumed to be time-invariant, each state of the second-level latent variable is associated to a multivariate support point occurring with a certain probability. In detail, every support point identifies a global effect on all the components of the first-level Markovian process, meaning that there is no separation, for instance, between the effect on the initial probabilities and that on the transition probabilities of the Markov chain. This model feature might represent an important limitation when SLU effects on initial and transition probabilities have a substantially different interpretation and a two-way classification of SLUs according to these different dimensions is therefore needed.

A motivating example is represented by longitudinal datasets collecting health indicators for elderly residents grouped in Nursing Homes (NHs). In this context, the first-level latent trait may be an ordinal one representing one domain of residents' overall health status and the main interest resides in detecting NH effects on the transition probabilities. Indeed, these are typically associated to the quality of the health care service they provide. Conversely, NH effects on the initial probabilities usually reflect admission policies, since NHs may have different tendencies to admit residents in more severe health conditions. With reference to this problem, some recently proposed solutions rely on modeling the transition probabilities as a function of fixed NH effects (Bartolucci et al. 2009) or introducing a bivariate Gaussian second-level latent variable (Montanari et al. 2018). Both these approaches allow to obtain an

NH ranking along the dimension of the SLU effects on transition probabilities only. However, the former is likely to result in poor effect estimation when the number of NHs is high or some NHs have just a few residents, while the latter relies on the Gaussianity assumption for the second-level latent effects, which is not testable and requires a remarkable computational effort to approximate integrals that do not admit a closed-form solution.

In this paper, we modify the setting in Montanari et al. (2018) introducing a discrete (rather than continuous) bivariate random effect. Such an approach has a twofold aim: (i) avoiding unverifiable parametric assumptions on the second-level latent variable; (ii) obtaining a direct two-way clustering of SLUs. In this respect, an ordinal two-way MLMM is introduced. This is based on the specification of a joint probability matrix assigning weights to the discrete support points of a bivariate second-level latent variable. When the two components of this latent variable are independent, the entries of such a joint probability matrix reduce to the product of the corresponding marginal probabilities, leading to an independence model. That is, the latter is nested into the proposed ordinal two-way MLMM, so that standard theory can be applied in a hypothesis testing perspective.

The paper is organized as follows. In Sect. 2, we set the notation and formulate the model, reporting also the details for its estimation procedure (Sect. 2.2). Moreover, we introduce the independence model (Sect. 2.3). In Sect. 3, we reconsider the motivating example described above and fit the proposed two-way MLMM to a longitudinal health care dataset referring to residents hosted in the NHs of Umbria, a region of central Italy. In Sect. 4, we report evidence from a small simulation study conducted to evaluate the model parameter estimators in a setting similar to the one considered in the empirical application, while in Sect. 5 some concluding remarks are given.

2 The two-way MLMM

2.1 The model

Let $\boldsymbol{Y}_{hi}^{(t)} = (Y_{hi1}^{(t)}, \ldots, Y_{hiJ}^{(t)})$ denote the vector containing the J categorical response variables for the i-th unit in the h-th SLU at time occasion t. Every item can have a different number of response categories labelled from 1 to c_j, with $j = 1, \ldots, J$. Each of the H SLUs has its own number of units n_h, so that the overall sample size is $n = \sum_{h=1}^{H} n_h$. The number of measurement occasions $T_{hi} \leq T$ is unit-specific, with T denoting the maximum number of occasions. Response vectors can be collected across time, i.e. $\boldsymbol{Y}_{hi} = (\boldsymbol{Y}_{hi}^{(1)}, \ldots, \boldsymbol{Y}_{hi}^{(T_{hi})})$, and across first-level units, i.e. $\boldsymbol{Y}_h = (\boldsymbol{Y}_{h1}, \ldots, \boldsymbol{Y}_{hn_h})$. The same notation applies to vectors of individual covariates, that we denote by $\boldsymbol{X}_{hi}^{(t)}$.

For each unit, the Markovian latent process $\boldsymbol{V}_{hi} = (V_{hi}^{(1)}, \ldots, V_{hi}^{(T_{hi})})$ is a collection of T_{hi} categorical unobserved variables with k_v latent states labelled from 1 to k_v. SLU random effects affecting such a process are assumed to be time-invariant and determined by a bivariate latent variable denoted by $\boldsymbol{Z}_h = (U_h, W_h)$. Here, U_h affects the initial probabilities, while W_h affects the transition probabilities. Formally, U_h and

W_h represent two separate latent variables defined on k_u and k_w support points; these are denoted by ψ_u $(u = 1, \ldots, k_u)$ and ξ_w $(w = 1, \ldots, k_w)$, respectively.

To formalize the two-way SLU clustering procedure, we denote by SLU-uw $(u = 1, \ldots, k_u, w = 1, \ldots, k_w)$ the group of SLUs sharing the u-th level of U_h and the w-th level of W_h. Every SLU belongs to the SLU-uw group with probability

$$\tau_{uw} = P(U_h = \psi_u, W_h = \xi_w), \qquad u = 1, \ldots, k_u, w = 1, \ldots, k_w.$$

To emphasize the two-way nature of our clustering approach, the above joint probabilities can be arranged in the probability matrix

$$\boldsymbol{\mathcal{T}} = \begin{pmatrix} \tau_{11} & \ldots & \tau_{1k_w} \\ \vdots & \ddots & \vdots \\ \tau_{k_u 1} & \ldots & \tau_{k_u k_w} \end{pmatrix},$$

whose rows and columns sum to

$$\tau_{u.} = \sum_{w=1}^{k_w} \tau_{uw} \qquad u = 1, \ldots, k_u \qquad \text{and} \qquad \tau_{.w} = \sum_{u=1}^{k_u} \tau_{uw} \qquad w = 1, \ldots, k_w$$

respectively. These marginals are collected in the vectors $\boldsymbol{\tau}_{u.} = (\tau_{1.}, \ldots, \tau_{k_u.})$ and $\boldsymbol{\tau}_{.w} = (\tau_{.1}, \ldots, \tau_{.k_w})$.

As typical in settings where the primary interest lies in modeling the latent trait, we assume covariates and SLU latent variables to influence the individual Markovian processes $\boldsymbol{V}_{hi}$ but not the measurement model, that is, the model for the response variables given the latent trait. Specifically, such a dependence structure leads to define the initial probabilities

$$\pi_{hi}(v|u) = P(V_{hi}^{(1)} = v|\boldsymbol{X}_{hi}^{(1)} = \boldsymbol{x}_{hi}^{(1)}, U_h = \psi_u),$$

the first-order transition probabilities

$$\pi_{hi}^{(t)}(v|\bar{v}, w) = P(V_{hi}^{(t)} = v|V_{hi}^{(t-1)} = \bar{v}, \boldsymbol{X}_{hi}^{(t)} = \boldsymbol{x}_{hi}^{(t)}, W_h = \xi_w)$$

$(t = 2, \ldots, T_{hi})$, and the conditional response probabilities

$$\phi_{jyv} = P(Y_{hij}^{(t)} = y|V_{hi}^{(t)} = v)$$

$(j = 1, \ldots, J;\ y = 1, \ldots, c_j;\ v = 1, \ldots, k_v)$, that are time-invariant. The resulting path diagram is depicted in Fig. 1 for a unit with $T_{hi} = 4$ measurement occasions. This diagram highlights the random effect separation characterizing our two-way MLMM. Indeed, U_h affects $V_{hi}^{(1)}$ only, whereas W_h affects all the other latent variables via its effect on $\pi_{hi}^{(t)}(v|\bar{v}, w)$ $(t = 2, \ldots, T_{hi})$.

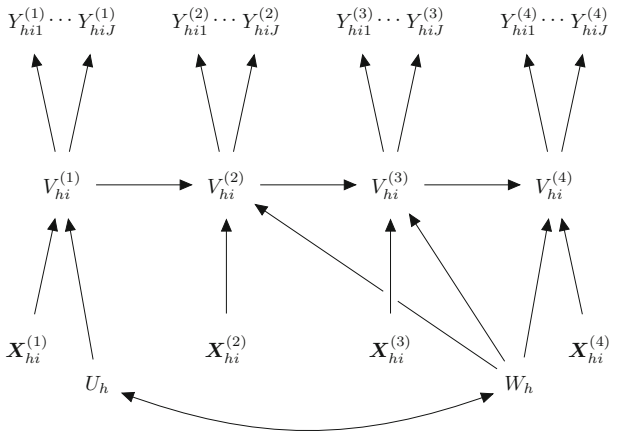

Fig. 1 Path diagram of the two-way MLMM for a unit with $T_{hi} = 4$ measurement occasions

As mentioned in Sect. 1, we assume that the latent trait $V_{hi}^{(t)}$ is ordinal, with states corresponding to increasing intensities of a certain attribute. In the presence of an ordinal trait, a number of parametrizations for the initial and transition probabilities can be adopted. These include, among others, the adjacent category, global and continuation logit parametrizations; see Bartolucci et al. (2013) for an overview. Here, for the initial probabilities a global logit parametrization based on the proportional odds model (McCullagh 1980) is used; that is,

$$\log \frac{\pi_{hi}(v+1|u) + \cdots + \pi_{hi}(k_v|u)}{\pi_{hi}(1|u) + \cdots + \pi_{hi}(v|u)} = \beta_{0v} + \boldsymbol{x}_{hi}^{(1)} \boldsymbol{\beta}_1 + \psi_u \tag{1}$$

$(u = 1, \ldots, k_u; v = 1, \ldots, k_v - 1)$. The parameter β_{0v} in the equation above is an intercept varying with the logit equations, ψ_u represents the effect due to SLU h belonging to latent group u, while $\boldsymbol{\beta}_1$ is a vector of fixed regression coefficients related to individual covariates. The global logit parametrization requires the components of $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0k_v-1})'$ be non-increasing to ensure the cumulative probabilities are non-decreasing. Further, to identify the model, we set $\psi_1 = 0$ together with the order constraints $\psi_1 \leq \cdots \leq \psi_{k_u}$.

With regard to the $(k_v \times k_v)$ transition probability matrices, we specify a global logit model similar to the previous one; that is,

$$\log \frac{\pi_{hi}^{(t)}(v+1|\bar{v}, w) + \cdots + \pi_{hi}^{(t)}(k_v|\bar{v}, w)}{\pi_{hi}^{(t)}(1|\bar{v}, w) + \cdots + \pi_{hi}^{(t)}(v|\bar{v}, w)} = \gamma_{0\bar{v}} + \gamma_{1v} + \boldsymbol{x}_{hi}^{(t)} \boldsymbol{\gamma}_2 + \xi_w, \tag{2}$$

with $w = 1, \ldots, k_w$, $\bar{v} = 1, \ldots, k_v$, $v = 1, \ldots, k_v - 1$ and $t = 2, \ldots, T_{hi}$. Again, the components of $\boldsymbol{\gamma}_1 = (\gamma_{11}, \ldots, \gamma_{1k_v-1})'$ must be non-increasing, whereas $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0k_v})'$ is a vector of additive shifts introduced to diversify the transition probability distributions among the rows of the transition matrices. No mathematical constraints are posed on the elements of $\boldsymbol{\gamma}_0$, apart from the necessary identifiabil-

ity constraint $\gamma_{01} = 0$. The vector $\boldsymbol{\gamma}_2$ is a time-invariant vector of fixed regression coefficients—constant across logit equations—for the effect of individual-level covariates, while ξ_w represents the effect due to SLU h belonging to latent group w. These effects are also time-invariant and constant across logit equations, subject to the identifiability constraints $\xi_1 = 0$ and $\xi_1 \leq \cdots \leq \xi_{k_w}$. These order constraints, together with those on the support points of U_h, ensure that the SLU-uw clusters are univocally identified, tackling the well-known problem of label switching (Stephens 2000) at the second level.

2.2 Two-step maximum likelihood estimation

The model log-likelihood can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{h=1}^{H} \log P(\boldsymbol{Y}_h = \boldsymbol{y}_h \mid \boldsymbol{X}_h = \boldsymbol{x}_h), \tag{3}$$

where $\boldsymbol{\theta}$ is a vector collecting all the parameters of the model and

$$P(\boldsymbol{Y}_h = \boldsymbol{y}_h \mid \boldsymbol{X}_h = \boldsymbol{x}_h) = \sum_{\boldsymbol{v}_h} \sum_{\boldsymbol{z}_h} P(\boldsymbol{Y}_h = \boldsymbol{y}_h, \boldsymbol{V}_h = \boldsymbol{v}_h, \boldsymbol{Z}_h = \boldsymbol{z}_h \mid \boldsymbol{X}_h = \boldsymbol{x}_h)$$

is obtained by summing over all the possible configurations of $\boldsymbol{v}_h$ and $\boldsymbol{z}_h$. The logarithm of each term in the summation above, that we denote by $\ell_h^\star(\boldsymbol{\theta})$, can be explicitly related to the model parameters. In detail, given the model assumptions we have

$$\begin{aligned}
\ell_h^\star(\boldsymbol{\theta}) = & \sum_{i=1}^{n_h} \sum_{t=1}^{T_{hi}} \sum_{j=1}^{J} \sum_{v=1}^{k_v} \sum_{y=1}^{c} a_{hi}^{(t)}(v) f_{hij}^{(t)}(y) \log \phi_{jyv} \\
& + \sum_{i=1}^{n_h} \sum_{v=1}^{k_v} \sum_{u=1}^{k_u} a_{hi}^{(1)}(v) c_h(u) \log \pi_{hi}(v \mid u) \\
& + \sum_{i=1}^{n_h} \sum_{t=2}^{T_{hi}} \sum_{v=1}^{k_v} \sum_{\bar{v}=1}^{k_v} \sum_{w=1}^{k_w} b_{hi}^{(t)}(v, \bar{v}) d_h(w) \log \pi_{hi}^{(t)}(v \mid \bar{v}, w) \\
& + \sum_{u=1}^{k_u} \sum_{w=1}^{k_w} e_h(u, w) \log \tau_{uw}.
\end{aligned} \tag{4}$$

In the expression above, a number of binary indicator variables are present. Specifically, $a_{hi}^{(t)}(v)$ takes value 1 if unit hi is in latent state v at time t, while $b_{hi}^{(t)}(v, \bar{v})$ indicates whether the same unit is in latent state v at time t and in latent state $\bar{v}$ at time $t-1$. Moreover, $c_h(u)$ and $d_h(w)$ indicate whether the h-th SLU belongs to the u-th and w-th cluster formed with respect to the SLU effect on the initial and transition probabilities, respectively, with $e_h(u, w)$ denoting the joint membership of the same

SLU to these two clusters (i.e., to the SLU-uw group). Finally, $f_{hij}^{(t)}(y)$ takes value 1 if at time t unit hi responds with category y at the j-th item.

Given the overall complexity of the proposed model, full maximum likelihood (ML) methods are likely to be unfeasible or highly unstable (Di Mari et al. 2016). Therefore, a two-step approach (Bartolucci et al. 2014) is undertaken. In the first step, conditional response probabilities ϕ_{jyv} (i.e., the measurement model) are estimated by fitting a latent class (LC) model (Lazarsfeld and Henry 1968; Goodman 1974) where all the individual records are pooled together. In the second step, ML estimates of these conditional response probabilities, $\hat{\phi}_{jyv}$, are plugged through (4) in the log-likelihood (3), which is then maximized with respect to the remaining model parameters, $\boldsymbol{\theta}_\ell$ (i.e., the latent model). The two-step method was proved to be consistent (Bartolucci et al. 2014, 2015) and is less likely to suffer from the instability issues mentioned above (Montanari and Pandolfi 2018).

At both steps, the log-likelihood maximization process is performed via an Expectation-Maximization (EM) algorithm (Dempster et al. 1977). While EM estimation of LC models is well-established and standard software is available for its implementation (see Sect. 3.2), that of the latent model requires a more in-depth discussion due to the specifics of the two-way MLMM we have introduced. Formally, taking the conditional response probability estimates as fixed, at the second step the complete-data log-likelihood can be regarded as a function of $\boldsymbol{\theta}_\ell$ only and written as

$$\hat{\ell}^\star(\boldsymbol{\theta}_\ell) = \sum_{h=1}^{H} \hat{\ell}_h^\star(\boldsymbol{\theta}_\ell).$$

In the above expression, $\hat{\ell}_h^\star(\boldsymbol{\theta}_\ell)$ is as in the right-hand side of (4) with $\hat{\phi}_{jyv}$ in place of ϕ_{jyv}, while the binary indicator variables $a_{hi}^{(t)}(v)$, $b_{hi}^{(t)}(v, \bar{v})$, $c_h(u)$, $d_h(w)$, $e_h(u, w)$ and $f_{hij}^{(t)}(y)$ are unchanged. Since they refer to the components of the latent processes, all these variables but $f_{hij}^{(t)}(y)$ are not observed. The first stage of the EM algorithm (E-step) consists in computing their conditional expectations given the observed data. These are denoted by adding the hat symbol to the corresponding indicator variables (e.g., $\hat{a}_{hi}^{(t)}(v)$ in place of $a_{hi}^{(t)}(v)$). In practice, the expressions for $\hat{a}_{hi}^{(t)}(v)$, $\hat{b}_{hi}^{(t)}(v, \bar{v})$, $\hat{c}_h(u)$, $\hat{d}_h(w)$, and $\hat{e}_h(u, w)$ reduce to a list of posterior probabilities, which is reported in Appendix A. These are used in the following stage of the algorithm (M-step). At this stage, the value of $\boldsymbol{\theta}_\ell$ is updated according to a constrained maximization of the expected second-step complete-data log-likelihood

$$\begin{aligned} E(\hat{\ell}^\star(\boldsymbol{\theta}_\ell) \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{x}) = & \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{t=1}^{T_{hi}} \sum_{j=1}^{J} \sum_{v=1}^{k_v} \sum_{y=1}^{c} \hat{a}_{hi}^{(t)}(v) f_{hi}^{(t)}(y) \log \hat{\phi}_{jyv} \\ & + \sum_{h=1}^{H} \sum_{u=1}^{k_u} \hat{c}_h(u) \sum_{i=1}^{n_h} \sum_{v=1}^{k_v} \hat{a}_{hi}^{(1)}(v) \log \pi_{hi}(v \mid u) \\ & + \sum_{h=1}^{H} \sum_{w=1}^{k_w} \hat{d}_h(w) \sum_{i=1}^{n_h} \sum_{t=2}^{T_{hi}} \sum_{v=1}^{k_v} \sum_{\bar{v}=1}^{k_v} \hat{b}_{hi}^{(t)}(v, \bar{v}) \log \pi_{hi}^{(t)}(v \mid \bar{v}, w) \\ & + \sum_{h=1}^{H} \sum_{u=1}^{k_u} \sum_{w=1}^{k_w} \hat{e}_h(u, w) \log \tau_{uw} \end{aligned}$$

(where $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_H)$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_H)$), which is performed in a way such that all the order constraints due to the global logit parametrization (see Sect. 2.1) as well as the probability constraint $\sum_{u=1}^{k_u} \sum_{w=1}^{k_w} \tau_{uw} = 1$ are met. The EM algorithm alternates the E-step and the M-step until convergence. Notice that the expression above directly involves the initial and transition probabilities. Thus, the present estimation framework based on the EM algorithm is valid regardless of the parametrization used for these probabilities. However, parametrizations alternative to the global logit might not require the aforementioned order constraints, that would be removed from the M-step.

As typical in complex latent variable models, the EM solution might correspond to a local maximum of the likelihood function rather than to the global one. A circumvention of this problem usually consists in performing multiple initializations of the algorithm, with the number of runs being a multiple of the number of latent groups at any unit level. In the two-step approach we undertake, this strategy has to be carried out separately for the measurement and the latent model; see Appendix B for further details in the context of our application in Sect. 3.

Once the ML estimate for the latent model parameter vector, $\hat{\boldsymbol{\theta}}_\ell$, is obtained, its estimated variance-covariance matrix can be computed via the sandwich formula

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\theta}}_\ell) = \{-\boldsymbol{H}(\hat{\boldsymbol{\theta}}_\ell)\}^{-1} \boldsymbol{S}(\hat{\boldsymbol{\theta}}_\ell) \{-\boldsymbol{H}(\hat{\boldsymbol{\theta}}_\ell)\}^{-1} \tag{5}$$

(White 1980; Royall 1986; see also Spagnoli et al. 2018). In the above equation, denoting by $\boldsymbol{s}_h(\boldsymbol{\theta}_\ell)$ the independent contribution of the h-th SLU to the score function

$$\boldsymbol{s}(\boldsymbol{\theta}_\ell) = \frac{\partial \ell(\boldsymbol{\theta}_\ell)}{\partial \boldsymbol{\theta}_\ell} = \sum_{h=1}^{H} \boldsymbol{s}_h(\boldsymbol{\theta}_\ell),$$

$\boldsymbol{H}(\boldsymbol{\theta}_\ell) = \partial \boldsymbol{s}(\boldsymbol{\theta}_\ell) / \partial \boldsymbol{\theta}_\ell'$ represents the Hessian information matrix while $\boldsymbol{S}(\boldsymbol{\theta}_\ell)$ is given by

$$\boldsymbol{S}(\boldsymbol{\theta}_\ell) = \sum_{h=1}^{H} \boldsymbol{s}_h(\boldsymbol{\theta}_\ell) \{\boldsymbol{s}_h(\boldsymbol{\theta}_\ell)\}'.$$

In practice, $\boldsymbol{H}(\boldsymbol{\theta}_\ell)$ is obtained by numerical derivation of the score $\boldsymbol{s}(\boldsymbol{\theta}_\ell)$, whose expression is reported, together with that of its components $\boldsymbol{s}_h(\boldsymbol{\theta}_\ell)$, in Appendix C. Although it is known to be robust to some degree of model misspecification, it is important to underline that the sandwich estimator in (5) ignores the sampling variability of the $\hat{\phi}_{jyv}$ estimates, which are taken as fixed in place of the true ϕ_{jyv} values in its expression. In principle, this might lead to systematically underestimate the variances in the true variance-covariance matrix $\text{Cov}(\hat{\boldsymbol{\theta}}_\ell)$; see Bakk and Kuha (2018) for a related discussion and an adjusted variance estimator in the context of LC models. However, this problem becomes negligible when an estimator with good finite-sample properties is used at the first step. This seems to be the case in our setting, where the first-step estimator of the ϕ_{jyv} probabilities is based on a pooled LC model including

all the available observations, which results in an increased sample size. An empirical account of the first-step estimator performance is reported in the simulation study, where the explored context is rather similar to the one considered in the empirical application (see Sect. 4 for details).

Finally, it is worth to mention that within this estimation framework also the model selection procedure consists of two phases. The first phase involves the choice of k_v, that is, the number of latent states characterizing the first-level unit latent process $\boldsymbol{V}_{hi}$. The second phase is about determining the value of the pair (k_u, k_w), that is, the number of groups of second-level units formed across the two dimensions of the proposed two-way MLMM. With regard to the empirical application considered here, these two phases are addressed in Sects. 3.2 and 3.3.

2.3 The independence model

As mentioned in Sect. 1, a nice feature of the proposed MLMM is that it nests the independence model. That is, it nests the model assuming independence among the latent variables influencing the initial and the transition probabilities of the first-level latent Markov chain, which allows the two SLU clustering procedures not to be influenced from one another. In the path diagram of Fig. 1, this would result in deleting the arrow connecting U_h and W_h.

As stated above, when U_h and W_h are independent, every joint probability τ_{uw} reduces to the product of the corresponding marginals

$$\tau_{uw} = \tau_{u.} \times \tau_{.w}. \tag{6}$$

The estimation of model parameters may proceed as detailed in Sect. 2.2, by simply replacing (6) in Eq. (3) (via (4)) and solving with respect to $\tau_{u.}$ and $\tau_{.w}$, under the constraints $\sum_{u=1}^{k_u} \tau_{u.} = \sum_{w=1}^{k_w} \tau_{.w} = 1$.

This result can be fruitfully employed also in a hypothesis testing perspective. To this end, we start by defining the following logit transformation for the joint masses τ_{uw}:

$$\log \frac{\tau_{uw}}{\tau_{k_u.}\tau_{.k_w}} = \lambda_u^I + \lambda_w^T + \zeta_{uw}$$

$(u = 1, \ldots, k_u - 1;\ w = 1, \ldots, k_w - 1)$, where

$$\lambda_u^I = \log \frac{\tau_{u.}}{\tau_{k_u.}}, \qquad \lambda_w^T = \log \frac{\tau_{.w}}{\tau_{.k_w}},$$

and

$$\zeta_{uw} = \log \frac{\tau_{uw}}{\tau_{u.}\tau_{.w}}.$$

The ζ_{uw} parameters directly provide a measure of the dependence between U_h and W_h and are null when Eq. (6) holds (see, e.g., Spagnoli et al. 2018). In this sense,

the independence model occurs when $\boldsymbol{\zeta} = (\zeta_{11}, \ldots, \zeta_{k_u-1,k_w-1})' = \mathbf{0}$. Defining the system of hypotheses

$$\begin{cases} H_0 : \boldsymbol{\zeta} = \mathbf{0} \\ H_1 : \boldsymbol{\zeta} \neq \mathbf{0}, \end{cases}$$

standard Wald-type, score, or likelihood ratio test (LRT) statistics may be employed. As regards the latter, we may compute

$$\text{LRT} = -2\big\{\ell(\hat{\boldsymbol{\theta}}_0) - \ell(\hat{\boldsymbol{\theta}}_1)\big\},$$

where $\ell(\hat{\boldsymbol{\theta}}_0)$ and $\ell(\hat{\boldsymbol{\theta}}_1)$ denote the likelihood values obtained at convergence of the EM algorithm under H_0 and H_1, respectively. From standard theory, it is known that all these statistics are asymptotically equivalent and follow a χ^2 distribution with $(k_u - 1) \times (k_w - 1)$ degrees of freedom under the null hypothesis.

3 Application to NH data

3.1 The LTCF dataset

The proposed two-way MLMM is used to analyze a longitudinal dataset collecting information on the health status of elderly residents hosted in the NHs of Umbria, a region of central Italy. Data are gathered within the *Suite interRAI* protocol (Carpenter and Hirdes 2013), an international system the regional government of Umbria has been adhering to for many years. In detail, NH residents are administered the Long Term Care Facilities (LTCF) questionnaire (Hirdes et al. 2008; Kim et al. 2015), which is specifically designed to investigate their health conditions within these facilities.

For this application, we consider a set of $n = 1548$ residents, grouped in $H = 43$ NHs, whose first observation falls in the second semester of 2017. These residents are then followed up for the years 2018 and 2019. According to the LTCF protocol, each resident should be administered a questionnaire every six months *and* whenever a significant change in their health status is acknowledged by the NH staff. For this reason, we have collected all the available observations for each resident, removing those with a subsequent one taken within seven days. In these cases, the second observations are almost identical to those deleted, probably because they are just amendments rather than actual new measurements. The resulting dataset comprises 5582 observations. Notice that the rationale leading to the construction of such a dataset is somewhat different from that in Montanari et al. (2018), where a panel dataset was obtained by taking the observations that were closest to predetermined six-month spaced dates.

Although the LTCF questionnaire investigates many health domains (physical limitations, psychological conditions, auditory and view sphere *etc.*), we here focus on a specific section named Activities of Daily Living (ADL). This section includes $J = 10$ ordinal response variables measuring resident difficulties in taking everyday actions like walking, getting dressed or maintaining personal hygiene. All these items have $c_j = c = 6$ categories, with labels increasing with the experienced difficulty level.

Table 1 Frequency distributions of the 5582 observations for the ADL items; summary statistics for resident age (years) and gender (1 = male) at baseline and for the distance from the previous observation (months, $t > 1$); summary statistics for the number of residents and observations across the NHs

Activities of Daily Living		Response category					
		1	2	3	4	5	6
1	Use of the shower stall/bath tub	0.015	0.083	0.109	0.207	0.157	0.428
2	Personal hygiene	0.037	0.109	0.107	0.207	0.146	0.395
3	Dressing (upper part)	0.057	0.126	0.123	0.205	0.131	0.359
4	Dressing (lower part)	0.049	0.105	0.093	0.200	0.149	0.403
5	Walking	0.156	0.114	0.069	0.134	0.083	0.443
6	Locomotion	0.165	0.130	0.072	0.136	0.071	0.426
7	Transfer to WC	0.143	0.093	0.092	0.165	0.112	0.395
8	WC use	0.113	0.100	0.084	0.164	0.101	0.438
9	Bed mobility	0.245	0.131	0.110	0.152	0.098	0.265
10	Eating	0.251	0.365	0.085	0.084	0.023	0.192
Covariates & NHs		Mean	Min	Q1	Median	Q3	Max
$x_{hi1}^{(1)}$	Age	82.63	25	77	85	90	106
$x_{hi2}^{(1)}$	Gender	0.284					
$x_{hi3}^{(t)}$	Distance from previous obs.	5.30	0.27	5.23	5.83	6	12.33
n_h	Number of residents	36	16	20	25	47.5	86
N_h	Number of obs.	139.8	45	79	106	177	344

A description of the items is contained in the upper part of Table 1, together with the associate frequency distributions. In this setting, the Markovian individual process $\boldsymbol{V}_{hi}$ is meant to represent the overall level of physical impairment, justifying the assumption that the latent trait is ordinal (see Sect. 2.1) as well as the global logit parametrization in Eqs. (1) and (2).

It is important to point out that, within the *Suite interRAI* protocol, questionnaires are filled in by the NH staff rather than by the residents personally. This fact should guarantee that compilation criteria are objective and time-invariant, thereby preventing from the well-known issue of response shift (Sprangers and Schwartz 1999; Visser et al. 2005). Response shift occurs with self-reported questionnaires and consists in respondents changing their internal standards and reconceptualizing quality of life domains between measurement occasions. It is likely to threaten the identification of true changes in the latent trait, resulting in the misspecification of the measurement model (Oort 2005). Since in our approach the conditional response probabilities are assumed to be time-invariant and their estimation is performed by pooling all the individual records together (see Sects. 2.1 and 2.2), ruling out the presence of response shift appears to be of the essence.

Another relevant aspect concerns resident dropout, that can be due to death or to other reasons not explicitly linked to health status transition. While the latter can essentially be treated like an ignorable missing data mechanism (Rubin 1976; Little

1995), the former clearly corresponds to a worsening in personal health conditions. In other words, death can be considered as an outcome of residents' health status trajectories rather than a source of missing data. For these reasons, we identify it with the k_v-th latent state of the individual Markov process. Such a latent state has two specific features. In detail, it is an absorbing state and it has null probability at $t = 1$. This means that residents can not be dead at the first measurement occasion and can not revert to other latent states once they have reached it. Clearly, this requires the constraints $\pi_{hi}(k_v|u) = 0$, as well as $\pi_{hi}^{(t)}(v|k_v, w) = 0$ and $\pi_{hi}^{(t)}(k_v|k_v, w) = 1$, for every value of u and w and for $v = 1, \ldots, k_v - 1$. These constraints are met simply by setting $\beta_{0k_v-1} = -\infty$ and $\gamma_{0k_v} = +\infty$. Thus, these two parameters are not to be estimated anymore. To identify dropout in the data, the 430 residents dropping out due to death are assigned an extra observation dated on the day of death with all the ADL items set to the additional response category $c + 1 = 7$. In terms of conditional response probabilities, this univocal association corresponds in practice to $\phi_{j,c+1,v} = 0$ and $\phi_{j,c+1,k_v} = 1$ ($v = 1, \ldots, k_v - 1$; $j = 1, \ldots, J$). Therefore, the final dataset contains $N = 6012$ observations, with about 3.88 observations per resident on average.

As for the covariates in the latent model, resident age (in years) and gender (a binary variable taking value 1 for males) form the individual covariate vector $\boldsymbol{x}_{hi}^{(1)}$ appearing in Eq. (1). These two covariates have been proved in different studies mentioned in Sect. 1 to affect both the initial probabilities and the probability of transition between latent states. Furthermore, in Eq. (2), the time distance from the previous collection occasion ($x_{hi3}^{(t)}$, measured in months) is included as an additional covariate, together with age and gender, in the vector $\boldsymbol{x}_{hi}^{(t)}$ to account for the different amount of time occurring from the previous observation on the same unit. Given the data collection mechanism discussed above, such a variable is expected to play a role also in measuring the overall quality of NH care services. Indeed, since questionnaires are administered prior to the six-month schedule if a sensible change in the health conditions is observed, such a variable is expected to be related to the probability that a transition occurs, though with some important provisos discussed in Sect. 3.3. Within this scheme, the latent variable U_h is meant to capture NH effects on the initial probabilities that are unexplained by age and gender, whereas W_h models NH effects on the transition probabilities unexplained by age, gender and time distance. In other words, time distance and W_h are two concomitant components of the overall NH performance. This is to some degree different from the panel-based approach in Montanari et al. (2018), where time distance was only included in the covariate set to adjust for unequally spaced measurements and not used in the NH performance evaluation process.

Since the date of death is recorded by the NH staff, all the covariates are available also for the dropout-related extra observations. The bottom part of Table 1 contains some summary statistics of these covariates, as well as of the number of residents and observations available for each NH. It is worth to highlight that some young residents are present in the dataset, although the *interRAI* protocol is in principle designed for elderly people. However, these cases are quite rare (there are only 9 residents aged less than 50), and are typically associated to rather serious impairments or mental handicap.

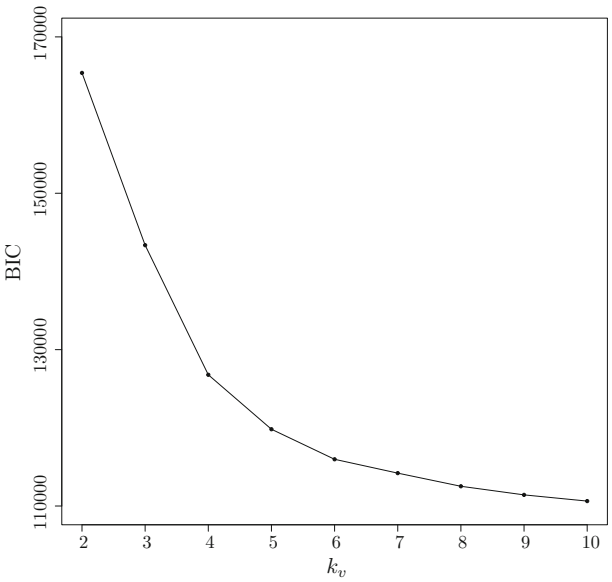

Fig. 2 BIC of the LC measurement models ($k_v = 2, \ldots, 10$)

3.2 Measurement model

As stated in Sect. 2.2, the two-way MLMM is estimated via a two step procedure. In this section, we show results for the first estimation step, i.e., for the measurement model, obtained via a pooled LC analysis without covariates performed on the ADL items presented in Sect. 3.1. ML estimation of the pooled LC models was performed via the poLCA R package (Linzer and Lewis 2011).

To address model selection, we have explored a set of models by letting k_v vary from 2 to 10. The model with $k_v = 1$ was not included since in the LTCF data we always have to account for the death-related extra latent state. That is, we always have at least two latent states; see Sect. 3.1. In Fig. 2, the Bayesian Information Criteria (BIC, Schwarz 1978) of the estimated models are plotted. In theory, models with a lower BIC should be preferred. Nevertheless, in the context of latent variable models this index is known to have a tendency to inflate the number of latent states (Pohle et al. 2017; Bacci et al. 2014). Specifically, when the BIC reaches a minimum within a reasonably large set of candidate models, then the model with the minimum value is typically a reliable choice. However, if it keeps decreasing with the consecutive reductions being always smaller, then it is a good practice to consider also alternative factors in the final model choice. This appears to be the case for the ADL-LC models considered here, where we observe a sensible reduction of the BIC until $k_v = 6$, with some kind of stabilization for $k_v \geq 7$ (Fig. 2).

In principle, the poLCA package is designed for nominal (not ordinal) responses and latent variables. However, starting from the estimated conditional response prob-

abilities, ordinality can be assessed a posteriori by means of the scores

$$\hat{s}_{jv} = \frac{1}{c}\sum_{y=1}^{c+1}(y-1)\hat{\phi}_{jyv} \qquad v = 1, \ldots, k_v, \quad j = 1, \ldots, J.$$

For each pair (j, v), $\hat{s}_{jv}$ is a normalized score laying in the 0-1 range and representing the average impairment level of residents in latent state v with respect to the j-th ADL. In line with the model formulation highlighted in Sect. 2.1, lower labels are assigned to states with lower values of $\hat{s}_{jv}$. Clearly, in the ordinal setting considered here it would be desirable that the same latent state ordering is maintained for every item. As a matter of fact, this ordering univocality among items is another relevant factor the choice of k_v might be based upon.

To investigate this issue in the LTCF data, the $\hat{s}_{jv}$ scores for all the fitted models are reported in Fig. 3. Specifically, in each plot of the 3×3 panel latent states are ordered according to the $\hat{s}_{jv}$ scores of the first item, with a line for each latent state ($v = 1, \ldots, k_v$) joining the $\hat{s}_{jv}$ values across the remaining items. Looking at the plots, it is possible to notice that for $k_v \geq 7$ (as well as for $k_v = 3$) the lines tend to overlap each other, meaning that the latent state ordering is not univocal. Another noteworthy feature is that when $k_v \geq 4$, the LC model correctly identifies the last extra latent state associated to death, where $\hat{s}_{jk_v} = 1$ by virtue of $\phi_{j,c+1,k_v} = 1$ ($j = 1, \ldots, J$).

In light of such considerations, a natural choice for the final measurement model is $k_v = 6$. The resulting conditional response probability structure (not shown) is aligned with those obtained in similar settings (Montanari et al. 2018). With regard to the $\hat{s}_{jv}$ scores, we evince that the last two items (bed mobility and eating) present the smallest values for every latent state. This finding is sensible since they represent abilities that are usually lost latest by residents. As a sensitivity check, the whole LC analysis was repeated including $\phi_{j,c+1,k_v} = 1$ ($j = 1, \ldots, J$) as a formal probability constraint. This procedure involves fitting models for the remaining $k_v - 1$ latent states on the reduced dataset obtained removing the 430 extra observations. The conclusions we may draw are substantially equivalent.

3.3 Latent model

Taking the estimated conditional response probabilities as fixed parameters, a number of two-way MLMMs are fitted. In detail, we let k_u and k_w vary between 1 and 3, so that 9 possible models are inspected overall. Before showing the estimation results, a more detailed discussion about the nature of the effect of $x_{hi3}^{(t)}$ (distance from the previous observation) in Eq. (2) is necessary. Specifically, it is important to observe that almost 80% of the values of $x_{hi3}^{(t)}$ lie between 5 and 7 months, that is, in the neighbourhood of the canonical measurement distance of 6 months. Because of this inflation, the effect of $x_{hi3}^{(t)}$ is likely to be blurred in this neighbourhood, and more distinct for $x_{hi3}^{(t)} < 5$. This intuition is confirmed by a sensitivity analysis we performed on the estimated measurement model. Specifically, we assigned each record a latent state based on a maximum-a-posteriori rule. In this way, we were able to build a binary variable

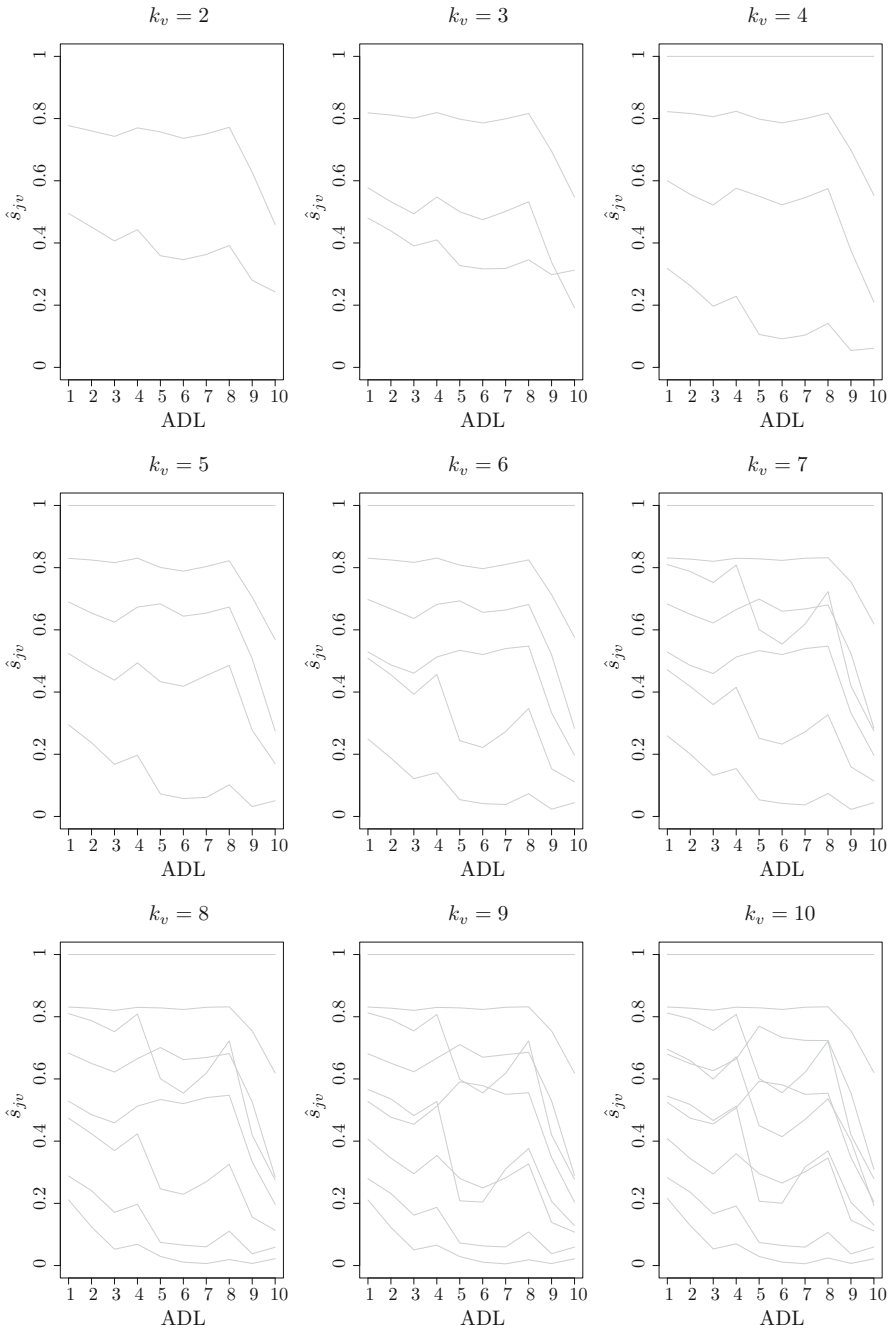

Fig. 3 Normalized item scores $\hat{s}_{jv}$ ($k_v = 2, \ldots, 10$)

Table 2 Log-likelihood, parameters and BIC for the models with $(k_u, k_w) \in \{1, 2, 3\}^2$

(k_u, k_w)	$\ell(\hat{\boldsymbol{\theta}})$	# par	BIC
(1,1)	− 52,647.99	18	105,428.19
(1,2)	− 52,626.84	20	105,400.58
(1,3)	− 52,625.58	22	105,412.74
(2,1)	− 52,617.12	20	105,381.13
(2,2)	− 52,592.56	23	105,354.05
(2,3)	− 52,589.02	26	105,369.01
(3,1)	− 52,607.82	22	105,377.23
(3,2)	− 52,584.05	26	105,359.07
(3,3)	− 52,582.59	30	105,385.53

taking value 1 if a latent state worsening with respect to the previous observation was observed and 0 otherwise, in the spirit of the global logit parametrization in (2). An extensive investigation about the relationship between $x_{hi3}^{(t)}$ and this variable, including the estimation of several logistic regression models and other empirical analyses, led to opt for the linear spline encoded by the covariate

$$\tilde{x}_{hi3}^{(t)} = \begin{cases} 3 & \text{if } x_{hi3}^{(t)} < 3 \\ x_{hi3}^{(t)} & \text{if } 3 \leq x_{hi3}^{(t)} \leq 5 \\ 5 & \text{if } x_{hi3}^{(t)} > 5. \end{cases}$$

In practice, the effect of $x_{hi3}^{(t)}$ on the worsening probability (that is, the probability of moving towards more impaired latent states) is assumed to be constant before 3 months, then it is assumed to be linear (on the logit scale) between 3 and 5 months, and then constant again after 5 months. This functional form allows to account for the data inflation around $x_{hi3}^{(t)} = 6$, as well as for a sort of flattening of the effect for lower values of $x_{3hi}^{(t)}$ emerging from empirical analyses. In what follows, we undertake this spline approach, including $\tilde{x}_{hi3}^{(t)}$ rather than $x_{hi3}^{(t)}$ as the third component of the $\boldsymbol{x}_{hi}^{(t)}$ vector for every fitted model. Also, age and $\tilde{x}_{hi3}^{(t)}$ were centered at the values of 80 years and 6 months, respectively, in order to ease the overall interpretation of $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$.

Table 2 reports the log-likelihood, the number of parameters in the latent model, and the BIC index for all the fitted models. With regard to the log-likelihood, for each value of k_u we notice that its increase is sensibly greater when moving from $k_w = 1$ to $k_w = 2$ than when moving from $k_w = 2$ to $k_w = 3$. The same argument applies swapping k_u and k_w. This intuition is confirmed by the BIC index, that reaches its minimum for the (2, 2) model. In other words, the log-likelihood gain due to the introduction of a third SLU cluster for the effect on either the initial and the transition probabilities is not worth the increase in the overall model complexity. As a consequence, we consider the (2, 2) model as the final one. For such a model, diagnostic checks based

Table 3 Estimates, standard errors and posterior SLU probabilities for the (2, 2) model

Parameters			Posterior SLU probabilities $\hat{e}_h(u, w)$							
par.	est.	s.e.	h	$\hat{e}_h(1,2)$	$\hat{e}_h(2,1)$	$\hat{e}_h(2,2)$	h	$\hat{e}_h(1,2)$	$\hat{e}_h(2,1)$	$\hat{e}_h(2,2)$
$\boldsymbol{\beta}_0$	1.553	0.165	1	1.000	0.000	0.000	23	0.995	0.000	0.005
	0.576	0.152	2	0.973	0.000	0.027	24	0.998	0.000	0.002
	−0.151	0.173	3	0.999	0.000	0.001	25	0.993	0.000	0.007
	−0.978	0.177	4	0.996	0.000	0.004	26	0.000	1.000	0.000
$\boldsymbol{\beta}_1$	0.012	0.007	5	1.000	0.000	0.000	27	0.016	0.000	0.984
	−0.493	0.109	6	1.000	0.000	0.000	28	0.873	0.000	0.127
ψ_2	1.036	0.205	7	1.000	0.000	0.000	29	1.000	0.000	0.000
$\boldsymbol{\gamma}_0$	5.278	0.337	8	0.332	0.000	0.668	30	1.000	0.000	0.000
	8.282	0.453	9	0.833	0.000	0.167	31	1.000	0.000	0.000
	10.765	0.574	10	0.059	0.000	0.941	32	0.806	0.000	0.194
	13.427	0.670	11	0.536	0.001	0.463	33	0.860	0.026	0.114
$\boldsymbol{\gamma}_1$	−0.661	0.264	12	0.000	1.000	0.000	34	0.000	0.000	1.000
	−4.759	0.411	13	0.795	0.017	0.188	35	1.000	0.000	0.000
	−7.354	0.507	14	0.997	0.000	0.003	36	0.995	0.000	0.005
	−9.664	0.611	15	0.000	0.000	1.000	37	0.976	0.000	0.024
	−13.257	0.695	16	0.000	0.353	0.647	38	0.979	0.000	0.021
$\boldsymbol{\gamma}_2$	0.016	0.002	17	0.124	0.000	0.876	39	0.011	0.001	0.988
	0.107	0.049	18	0.000	1.000	0.000	40	0.281	0.004	0.715
	−1.464	0.257	19	0.032	0.001	0.967	41	0.535	0.002	0.464
ξ_2	0.917	0.268	20	0.988	0.000	0.012	42	0.952	0.000	0.048
			21	0.989	0.000	0.011	43	0.997	0.000	0.003
			22	1.000	0.000	0.000				

on ordinary normal pseudo-residuals (Zucchini and MacDonald 2009, Chapter 6) show a satisfactory fitting for every item; see Fig. 4.

The left-hand side of Table 3 reports the estimates, together with their standard errors, for all the parameters of the latent (2, 2) model but the probability matrix $\boldsymbol{\mathcal{T}}$. From this table, we may conclude that, for both the initial and transition probabilities, the estimated effect of age and gender is in line with that observed in similar analyses of the LTCF data referring to other years (Bartolucci et al. 2009; Montanari et al. 2018). In detail, given the global logit parametrization in Eqs. (1) and (2), the positive effect of age shows that older residents are more likely to find themselves in a worse physical health status at admission ($\hat{\beta}_{11} = 0.012$), as well as to move towards more serious latent states at the following occasions ($\hat{\gamma}_{21} = 0.016$). However, the $\hat{\beta}_{11}$ estimate is barely significant (p-value 0.098). As for the effect of gender, we can conclude that males typically have lower physical impairment at the first occasion ($\hat{\beta}_{12} = -0.493$, p-value 5.68×10^{-6}), but are also more likely to worsen their condition ($\hat{\gamma}_{22} = 0.107$, p-value 0.030) with respect to females. Regarding the effect of $\tilde{x}_{hi3}^{(t)}$, its estimated coefficient $\hat{\gamma}_{23} = -1.464$ is highly significant (p-value 1.29×10^{-8}) and shows that shorter

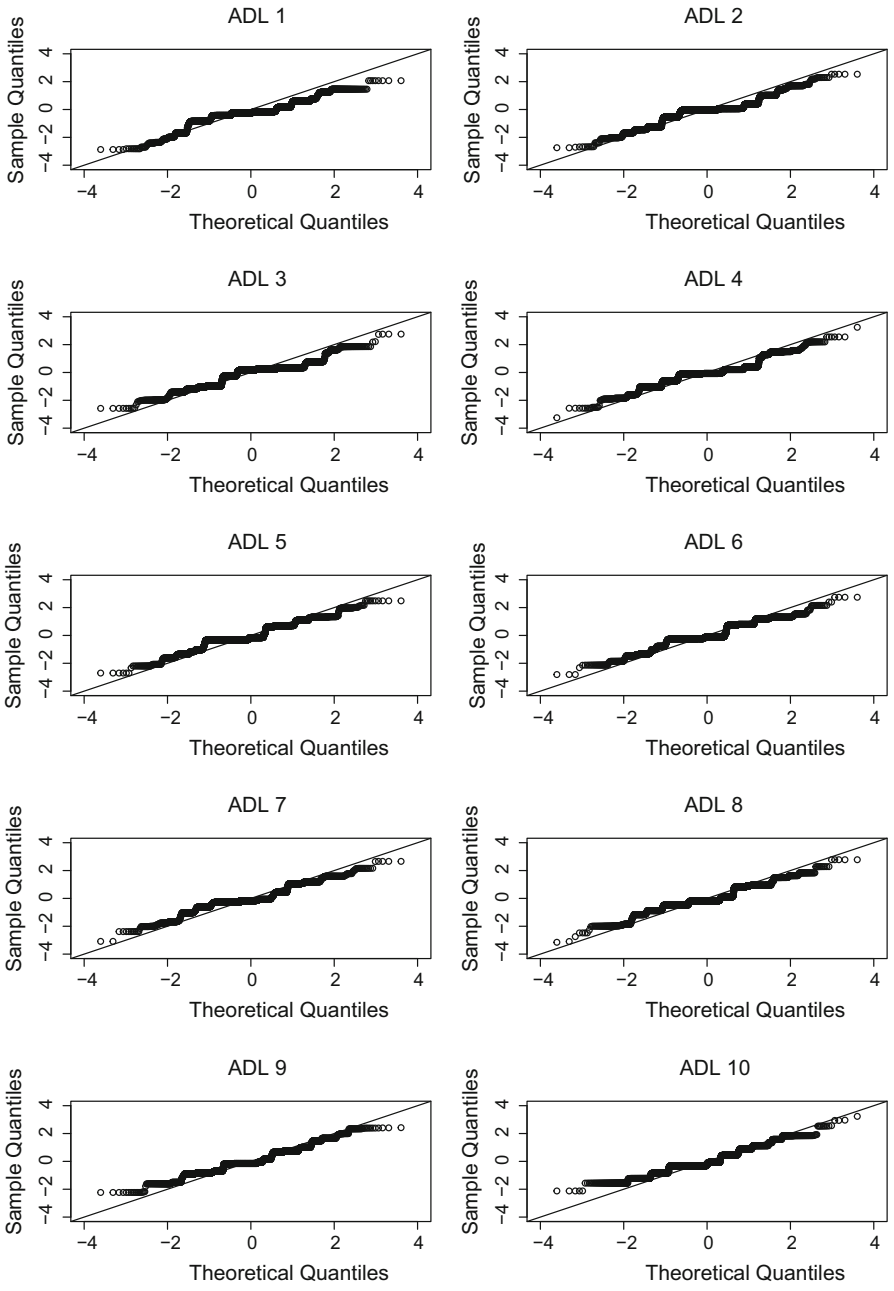

Fig. 4 QQplots of ordinary normal pseudo-residuals for the (2, 2) model

distances between observations—within the interval [3, 5] months—are associated to higher probabilities of moving towards higher (i.e., more impaired) latent states. This effect is in line with subject-matter related expectations, since sudden changes in residents' health status are more likely to correspond to a worsening in their physical conditions.

Looking at the second-level latent variables, it is possible to notice that $\hat{\psi}_2 = 1.0356$ and $\hat{\xi}_2 = 0.9174$ are both significantly different from zero (p-values 4.16×10^{-7} and 6.25×10^{-4} respectively), thereby corroborating the adoption of the $(2, 2)$ model. In other words, there is evidence of the presence of 4 SLU clusters corresponding to the 4 combinations of the joint effect $\boldsymbol{Z}_h = (U_h, W_h)$ on the initial and transition probabilities of the first-level-unit Markov process. Recalling the notation introduced in Sect. 2.1, we label these clusters by SLU-uw ($u, w = 1, 2$).

The estimated SLU joint probability matrix is

$$\hat{\boldsymbol{\mathcal{T}}} = \begin{pmatrix} 0.000 & 0.672 \\ 0.079 & 0.249 \end{pmatrix}.$$

Combining the estimates above with $\hat{\psi}_2$ and $\hat{\xi}_2$, it is possible to compute the estimated correlation between U_h and W_h, which is equal to –0.420. Such an estimate suggests a deviation from the independence model, which is confirmed by a formal LRT. Specifically, the log-likelihood of the $(2, 2)$ independence model is –52596.1, so that the LRT statistic is equal to 7.072 (p-value 0.008 under the null $\chi^2(1)$ distribution).

Since $\hat{\tau}_{11} \approx 0$, we can conclude that there are no NHs belonging to the SLU-11 cluster. Conversely, the model suggests that around 8% of the NHs belong to cluster SLU-21. The remaining 92% of NHs are placed in the other two clusters. Of these, around 73% are estimated to belong to cluster SLU-12.

In order to understand which group each NH is more likely to belong to, one can rely on the posterior SLU probabilities $\hat{e}_h(u, w)$ estimated at convergence of the EM algorithm. All these probabilities but $\hat{e}_h(1, 1)$, which are always lower than 10^{-3}, are reported in the right-hand side of Table 3. It is possible to pinpoint three NHs belonging to the SLU-21 group with a very low degree of uncertainty ($h \in \{12, 18, 26\}$). The other NHs are estimated to split between clusters SLU-12 and SLU-22 with a varying degree of uncertainty.

To characterize the identified clusters, the estimated initial probability distribution of the latent states for an 80-year-old female resident hosted in an NH belonging to the SLU-1w clusters ($w = 1, 2$) is

$$\hat{\boldsymbol{\pi}}_{hi}(u = 1) = \begin{pmatrix} 0.175 & 0.185 & 0.178 & 0.189 & 0.273 & 0.000 \end{pmatrix},$$

whereas the vector for a resident with the same individual covariate pattern hosted in an NH of the SLU-2w clusters ($w = 1, 2$) is

$$\hat{\boldsymbol{\pi}}_{hi}(u = 2) = \begin{pmatrix} 0.070 & 0.097 & 0.126 & 0.193 & 0.514 & 0.000 \end{pmatrix}.$$

In line with the parameter constraints $\psi_2 \geq \psi_1 = 0$, this shows that the NHs in the latter clusters tend to admit residents in worse conditions with respect to physical impairment.

For the same reference resident, the 6-month ahead ($\tilde{x}_{hi3}^{(t)} = 0$) transition matrix associated to the SLU-$u1$ clusters ($u = 1, 2$) is

$$\hat{\boldsymbol{\Pi}}_{hi}^{(t)}(w = 1) = \begin{pmatrix} 0.659 & 0.332 & 0.008 & 0.001 & 0.000 & 0.000 \\ 0.010 & 0.364 & 0.515 & 0.099 & 0.012 & 0.000 \\ 0.000 & 0.028 & 0.255 & 0.516 & 0.194 & 0.007 \\ 0.000 & 0.002 & 0.030 & 0.218 & 0.674 & 0.076 \\ 0.000 & 0.000 & 0.002 & 0.020 & 0.435 & 0.543 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix},$$

while that for the SLU-$u2$ groups ($u = 1, 2$) is

$$\hat{\boldsymbol{\Pi}}_{hi}^{(t)}(w = 2) = \begin{pmatrix} 0.436 & 0.543 & 0.019 & 0.002 & 0.000 & 0.000 \\ 0.004 & 0.188 & 0.569 & 0.209 & 0.029 & 0.001 \\ 0.000 & 0.011 & 0.125 & 0.478 & 0.369 & 0.017 \\ 0.000 & 0.001 & 0.012 & 0.104 & 0.711 & 0.172 \\ 0.000 & 0.000 & 0.001 & 0.008 & 0.243 & 0.748 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}.$$

It is worth to recall that the two matrices above differ only for the residual NH effect W_h. Again, conditional on the previous latent state and the covariates, in the first matrix a lower chance of worsening the physical conditions or of dying is observed, corresponding to a better performance of the NHs belonging to the SLU-$u1$ clusters. However, to properly evaluate the overall performance of an NH in avoiding the worsening of physical impairment, one should also take into account the effect of the $\tilde{x}_{hi3}^{(t)}$ covariate: as previously mentioned, worsening is more likely as its value decreases.

Although primarily intended as a SLU clustering tool, in the context of this application the proposed two-way MLMM may be used also for NH performance evaluation and ranking purposes. For example, an overall index can be built given by the sum of the averaged contribution, on the global logit scale of Eq. (2), of the observed values of $\tilde{x}_{hi3}^{(t)}$ in a NH and the corresponding posterior expected value of W_h, controlling for age and gender. However, in principle such a procedure should be conducted separately for the groups formed with respect to the NH effect on the initial probabilities. In this way, the lack of independence between the latent variables U_h and W_h is properly accounted for. In a policy evaluation perspective, this approach would discourage the adoption of unfair practices like adverse selection of residents at baseline (Montanari et al. 2018; Montanari and Doretti 2019). We argue that these group-dependent measures would be more naturally obtained from the present discrete-effect model rather than from continuous-effect models. However, this kind of development is beyond the scope of the present paper.

4 Simulation study

In this section, we present some evidence from a small simulation study conducted within the two-way MLMM framework. For the reasons outlined in Sect. 2.2, the main purpose of this simulation is checking the performance of the proposed two-step procedure in a setting close to the LTCF application rather than comparing different estimation methods. To this end, $B = 500$ datasets are generated from a two-way MLMM with $k_v = 6$, $k_u = 2$, $k_w = 2$ and the same probability constraints as in Sect. 3.2 for the sixth state of the individual level latent process.

Each dataset includes $H = 60$ SLUs with $n_h = 50$ first-level units. At baseline, a normal variate with mean equal to 83 and variance equal to 49 and a Bernoulli variate with probability 0.3 are generated for every first-level unit to mimic the distributions of age (in years) and gender (1=male) in the LTCF data. However, to simplify the whole setting, each unit is assigned $T = 3$ equally spaced measurement occasions. Thus, the $x_{hi3}^{(t)}$ covariate is removed and a constant number of observations across datasets ($N = 9000$) is generated. Since we assume the canonical distance of 6 months between measurements, at $t = 3$ age is deterministically increased of 1 unit with respect to $t = 1$, whereas at $t = 2$ such an increase is random. This mechanism reflects the fact that residents are born in different times of the year.

The data generating process follows the scheme outlined in Sect. 2.1 according to a sequential procedure. First, SLUs are assigned to SLU-uw clusters based on the SLU probabilities τ_{uw}. Conditional on this assignment, the values of the covariates and the remaining latent model parameters, initial and transition probabilities are then determined and used to simulate the first-level unit Markov chains $\boldsymbol{v}_{hi}$ ($h = 1, \ldots, H$; $i = 1, \ldots, n_h$). The numerical values for the latent model parameters are reported in the second column of Table 4. Notice that two scenarios are considered which differ for the values of the $\boldsymbol{\mathcal{T}}$ matrix only, with the first scenario corresponding to an independence model. Finally, given the realized $\boldsymbol{v}_{hi}$ processes, $J = 10$ ordinal items with $c + 1 = 7$ categories are generated according to a conditional response probability array similar to the one of the LTCF data (not shown). In detail, items are set to the seventh category with probability 1 if a unit is in the sixth latent state and with probability 0 otherwise, in line with the conditional response probability constraints introduced in Sect. 3.1.

The two-step estimator of the two-way MLMM fitted on the simulated datasets is as described in Sect. 2.2. For each dataset, at the first step 10 random starts are run, whereas at the second step the true latent model parameter vector is taken as initial guess. This combined strategy allows to avoid local maxima issues as well as to speed up the overall computational time. As expected, the performance of the first-step LC model estimator is highly remarkable. Indeed, for the conditional response probabilities the maximum absolute bias is 0.0016 for the first scenario and 0.0022 for the second scenario, while the maximum simulation standard deviation is 0.0195 for the first scenario and 0.0201 for the second scenario. Importantly, discrepancies about latent state ordering across different items are absent.

While the LC measurement model is always fitted with 6 latent states like in the true data generating process, at the second step all the latent models with $(k_u, k_w) \in$

Table 4 Simulation results for the latent model (independence model in the upper part of the table)

Latent model									
par.	True	Min.	Median	Mean	Max.	SD	Bias	RRMSE	AESE
$\boldsymbol{\beta}_0$	0.35	−0.861	0.323	0.332	1.314	0.395	−0.018	1.129	0.393
	−0.75	−1.895	−0.773	−0.769	0.236	0.393	−0.019	0.525	0.393
	−1.50	−2.633	−1.525	−1.522	−0.498	0.393	−0.022	0.263	0.393
	−2.45	−3.628	−2.469	−2.470	−1.432	0.394	−0.020	0.161	0.395
$\boldsymbol{\beta}_1$	0.02	0.008	0.020	0.020	0.034	0.005	0.000	0.232	0.005
	−0.38	−0.577	−0.377	−0.376	−0.182	0.071	0.004	0.188	0.071
ψ_2	1.05	0.804	1.050	1.051	1.269	0.074	0.001	0.070	0.069
$\boldsymbol{\gamma}_0$	4.25	3.573	4.251	4.252	4.818	0.190	0.002	0.045	0.185
	7.05	6.289	7.054	7.055	7.787	0.210	0.005	0.030	0.208
	9.15	8.380	9.157	9.160	9.912	0.219	0.010	0.024	0.216
	11	10.122	11.007	11.013	11.781	0.221	0.013	0.020	0.220
$\boldsymbol{\gamma}_1$	−3.65	−4.786	−3.662	−3.670	−2.506	0.339	−0.020	0.093	0.334
	−7.20	−8.505	−7.230	−7.217	−5.935	0.374	−0.017	0.052	0.372
	−9.55	−10.896	−9.579	−9.568	−8.318	0.382	−0.018	0.040	0.383
	−11.45	−12.940	−11.472	−11.473	−10.192	0.389	−0.023	0.034	0.389
	−13.60	−15.037	−13.632	−13.627	−12.336	0.394	−0.027	0.029	0.396
$\boldsymbol{\gamma}_2$	0.02	0.007	0.020	0.020	0.033	0.004	0.000	0.192	0.004
	0.15	−0.017	0.155	0.152	0.299	0.060	0.002	0.403	0.058
ξ_2	1.50	1.319	1.501	1.501	1.696	0.058	0.001	0.039	0.057
τ_{11}	0.20	0.057	0.200	0.201	0.363	0.052	0.001	0.257	0.116
τ_{12}	0.30	0.114	0.300	0.299	0.508	0.062	0.000	0.206	0.072
τ_{21}	0.20	0.042	0.203	0.203	0.372	0.051	0.003	0.253	0.059
τ_{22}	0.30	0.101	0.293	0.296	0.509	0.059	−0.003	0.196	0.072
$\boldsymbol{\beta}_0$	0.35	−1.212	0.338	0.348	1.571	0.408	−0.002	1.166	0.397
	−0.75	−2.265	−0.751	−0.750	0.507	0.403	0.000	0.538	0.396
	−1.50	−3.005	−1.505	−1.499	−0.211	0.406	0.001	0.271	0.396
	−2.45	−3.942	−2.458	−2.451	−1.182	0.407	−0.001	0.166	0.397
$\boldsymbol{\beta}_1$	0.02	0.005	0.020	0.020	0.038	0.005	0.000	0.242	0.005
	−0.38	−0.653	−0.381	−0.383	−0.172	0.073	−0.003	0.193	0.071
ψ_2	1.05	0.844	1.052	1.054	1.306	0.071	0.004	0.068	0.068
$\boldsymbol{\gamma}_0$	4.25	3.756	4.263	4.266	4.781	0.190	0.016	0.045	0.190
	7.05	6.287	7.064	7.071	7.781	0.209	0.021	0.030	0.213
	9.15	8.497	9.172	9.175	9.849	0.214	0.025	0.024	0.220
	11	10.318	11.027	11.030	11.725	0.218	0.030	0.020	0.225
$\boldsymbol{\gamma}_1$	−3.65	−4.724	−3.646	−3.647	−2.508	0.350	0.003	0.096	0.338
	−7.20	−8.332	−7.215	−7.207	−5.967	0.380	−0.007	0.053	0.378
	−9.55	−10.770	−9.569	−9.565	−8.275	0.389	−0.015	0.041	0.390
	−11.45	−12.721	−11.456	−11.468	−10.279	0.395	−0.018	0.035	0.396
	−13.60	−14.824	−13.631	−13.621	−12.368	0.401	−0.021	0.030	0.403

Table 4 continued

Latent model									
par.	True	Min.	Median	Mean	Max.	SD	Bias	RRMSE	AESE
γ_2	0.02	0.007	0.020	0.020	0.031	0.004	0.000	0.198	0.004
	0.15	−0.034	0.152	0.149	0.318	0.057	−0.001	0.383	0.058
ξ_2	1.50	1.304	1.507	1.503	1.675	0.059	0.003	0.040	0.059
τ_{11}	0.20	0.063	0.199	0.199	0.395	0.052	0.003	0.268	0.116
τ_{12}	0.29	0.120	0.290	0.291	0.480	0.057	−0.001	0.197	0.071
τ_{21}	0.12	0.011	0.119	0.122	0.317	0.044	0.003	0.369	0.045
τ_{22}	0.39	0.232	0.395	0.389	0.601	0.061	−0.005	0.156	0.082

$\{1, 2, 3\}^2$ are estimated. In this way, we are able to control the performance of the BIC index as a model selection tool for the second step. In detail, for the first scenario only for 3 of the 500 datasets (3, 2) model is selected in place of the true one. For the second scenario, the (2, 3) model is selected 6 times, the (3, 2) model is selected 2 times, while in the other cases the correct (2, 2) model is chosen. These results denote a good performance of the BIC index in this context.

With regard to the performance of the proposed estimation procedure at the latent layer, results are summarized in Table 4. For each parameter, the true value is accompanied by the main summary statistics of the simulation distribution, including the standard deviation (SD) and the relative root mean squared error (RRMSE). From this table, it is possible to observe that parameter estimators exhibit very small bias and a modest degree of variability, with some deviations occurring for the first element of $\boldsymbol{\beta}_0$. Also, results are very stable across the two scenarios, with little differences only concerning the minimum and maximum values.

The last column of Table 4 contains the average estimated standard error (AESE) for each parameter. As mentioned in Sect. 2.2, in principle these values might be prone to underestimate the true standard deviations of the parameter estimators, due to the estimator in (5) ignoring the sampling variability concerning the conditional response probabilities. However, the results in Table 4 show that AESEs are quite close to Monte Carlo standard deviations (that is, the SD column in the table), apart from some degree of overestimation observed for the SLU-group probabilities. This is not surprising given the very good performance of the measurement model estimator and the adoption of a robust estimation method like the sandwich estimator. Similar patterns are observed for the models fitted with $(k_u, k_w) \neq (2, 2)$. However, when k_u or k_w are greater than 2, the variability of the estimators of the ψ_3 and ξ_3 parameters is sometimes underestimated. It seems reasonable to ascribe this fact to a model misspecification (we recall that data are generated under the (2,2) model) rather than to first-step variability.

5 Conclusions

In this paper, a two-way multilevel latent Markov model (MLMM) is introduced, which is in order when sample units are grouped into second level units (SLUs). In these

settings, the multilevel structure of the data is typically accounted for by introducing second-level latent variables affecting the observed responses via a set of random effects. These random effects are assumed to be discrete, resulting in a clustering of SLUs. The main methodological innovation of the proposed model consists in the fact that the vector of discrete second-level random effects does not have a univocal influence on the whole first-level Markovian process. Indeed, distinct effects for the initial and transition probabilities are specified, so that SLUs can be clustered along the two dimensions separately, using the Cartesian product of the support points for the random effects on the initial probabilities and those for the transition probabilities. This feature is particularly appealing for those applications where SLU random effects on these two components have a substantially different interpretation. An independence test between the effects on the initial probabilities and the effects on the transition probabilities is also proposed. This test allows to detect when the SLU clustering process along a dimension is independent of that along the other.

An application of the proposal is illustrated with reference to health care data collected on residents hosted in Nursing Homes (NHs), which are the SLUs at issue. NH effects on the initial probabilities of the latent states of first-level units can be ascribed to different admission policies (i.e., NHs tend to admit residents in different health conditions at baseline), whereas those on the transition probabilities are related to the quality of the health care service provided (i.e., NHs with their actions generate different impacts on their residents' probabilities of moving to a different health status). For such data, modeling the NH effects beside those of the available covariates on the transition probabilities may allow to build indicators of NH performances in taking care of their residents.

Although it appears to be absent in the considered application (see Sect. 3.1), response shift might characterize other settings involving latent Markov models. Therefore, solutions to this issue in the latent Markov framework are desirable. In particular, one could adopt a time-varying parametrization for the conditional response probabilities, and perform its estimation accordingly. This approach seems to be the categorical counterpart to the one developed for structural equation models (Oort 2005), which refers to linear models and continuous latent traits and response variables.

The proposed model is tailored to an ordinal first-level latent trait. As a consequence, the marginal SLU clusters defined along the two random effect dimensions also have a conceptual ordering and, thus, a clear interpretation. While the extension to non-ordinal settings is possible, interpretation issues might arise. In this sense, a partially ordered set approach could be helpful. Such an approach has already been introduced within the hidden Markov model framework (Ip et al. 2013) and would be particularly sensible for those datasets where indicators measuring several latent domains are collected. Its extension to a multilevel setting would be a promising topic for future research.

Acknowledgements This research is funded by a grant of Fondazione Cassa di Risparmio di Perugia, Italy. We also thank the Region of Umbria for making available the NHs dataset used in the application.

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

Appendix A: E-step details

As mentioned in Sect. 2.2, the indicator variables $a_{hi}^{(t)}(v)$, $b_{hi}^{(t)}(v,\bar{v})$, $c_h(u)$, $d_h(w)$, $e_h(u,w)$ and $f_{hij}^{(t)}(y)$ are defined as

$$\begin{aligned}
a_{hi}^{(t)}(v) &= I(V_{hi}^{(t)} = v) && t = 1,\ldots,T_{hi};\\
b_{hi}^{(t)}(v,\bar{v}) &= I(V_{hi}^{(t-1)} = \bar{v}, V_{hi}^{(t)} = v) && t = 2,\ldots,T_{hi};\\
c_h(u) &= I(U_h = \psi_u);\\
d_h(w) &= I(W_h = \xi_w);\\
e_h(u,w) &= I(U_h = \psi_u, W_h = \xi_w);\\
f_{hij}^{(t)}(y) &= I(Y_{hij}^{(t)} = y) && t = 1,\ldots,T_{hi}.
\end{aligned}$$

The conditional expectations of the first five variables, given the data and the parameter vector $\boldsymbol{\theta}$, are the posterior probabilities

$$\begin{aligned}
\hat{a}_{hi}^{(t)}(v) &= P(V_{hi}^{(t)} = v \mid \boldsymbol{Y}_h = \boldsymbol{y}_h, \boldsymbol{X}_h = \boldsymbol{x}_h)\\
&= \sum_{u=1}^{k_u}\sum_{w=1}^{k_w}\Big\{P(V_{hi}^{(t)} = v \mid \boldsymbol{Y}_h = \boldsymbol{y}_h, \boldsymbol{X}_h = \boldsymbol{x}_h, U_h = \psi_u, W_h = \xi_w)\times\\
&\quad P(U_h = \psi_u, W_h = \xi_w \mid \boldsymbol{Y}_h = \boldsymbol{y}_h, \boldsymbol{X}_h = \boldsymbol{x}_h)\Big\} && t = 1,\ldots,T_{hi};\\
\hat{b}_{hi}^{(t)}(v,\bar{v}) &= P(V_{hi}^{(t-1)} = \bar{v}, V_{hi}^{(t)} = v \mid \boldsymbol{Y}_h = \boldsymbol{y}_h, \boldsymbol{X}_h = \boldsymbol{x}_h)\\
&= \sum_{u=1}^{k_u}\sum_{w=1}^{k_w}\Big\{P(V_{hi}^{(t-1)} = \bar{v}, V_{hi}^{(t)} = v \mid \boldsymbol{Y}_h = \boldsymbol{y}_h, \boldsymbol{X}_h = \boldsymbol{x}_h, U_h = \psi_u, W_h = \xi_w)\times\\
&\quad P(U_h = \psi_u, W_h = \xi_w \mid \boldsymbol{Y}_h = \boldsymbol{y}_h, \boldsymbol{X}_h = \boldsymbol{x}_h)\Big\} && t = 2,\ldots,T_{hi};\\
\hat{c}_h(u) &= \sum_{w=1}^{k_w} P(U_h = \psi_u, W_h = \xi_w \mid \boldsymbol{Y}_h = \boldsymbol{y}_h, \boldsymbol{X}_h = \boldsymbol{x}_h);\\
\hat{d}_h(w) &= \sum_{u=1}^{k_u} P(U_h = \psi_u, W_h = \xi_w \mid \boldsymbol{Y}_h = \boldsymbol{y}_h, \boldsymbol{X}_h = \boldsymbol{x}_h);\\
\hat{e}_h(u,w) &= P(U_h = \psi_u, W_h = \xi_w \mid \boldsymbol{Y}_h = \boldsymbol{y}_h, \boldsymbol{X}_h = \boldsymbol{x}_h).
\end{aligned}$$

The probabilities involved in the expressions above can be obtained from the parametrized ones (i.e., the initial, transition and SLU probabilities), by using Bayes' theorem and/or the forward and backward Baum-Welch recursions (Baum et al. 1970; Welch 2003), which are commonly adopted in the hidden Markov model literature.

Appendix B: multiple initialization of EM algorithms

With regard to the measurement model, $20 \times k_v$ replications are run for every model ($k_v \in 2, \ldots, 10$) to overcome local maxima issues. As for the latent model, a combined strategy is undertaken which consists of the following steps: (i) fit the three models with $k_w = 1$ adopting multiple random starting points, (ii) use the corresponding solutions as a basis for the starting vectors of the other models, with the remaining parameters varying according to a deterministic grid. Such a grid allows to explore all the plausible values in the parametric subspaces.

With regard to step (i), we perform $50 \times k_u$ random starts with initial values given by

$$\begin{aligned}
\boldsymbol{\beta}_0^{\text{START}} &= (-3, -1, 1, 3)' + \epsilon_1 \\
\boldsymbol{\gamma}_0^{\text{START}} &= (3, 4, 5, 6)' + \epsilon_2 \\
\boldsymbol{\gamma}_1^{\text{START}} &= (0, -1.5, -3, -4.5, 6)' + \epsilon_3
\end{aligned}$$

(recall that $k_v = 6$ with $\beta_{05} = -\infty$ and $\gamma_{06} = +\infty$). In the above, ϵ_1, ϵ_2 and ϵ_3 are drawn from a standard normal distribution. Moreover, the elements of regression coefficient vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}_2$ are sampled from a $N(0, 0.01)$ distribution, whereas $(\psi_2, \ldots, \psi_{k_u})$ are drawn from non-overlapping uniform distributions whose centers are spaced by 0.7. Finally, the probabilities $\boldsymbol{\mathcal{T}} = (\tau_{11}, \ldots, \tau_{k_u 1})'$ are sampled from a Dirichlet distribution with parameter vector given by $3 \cdot \mathbf{1}_{k_u}$.

With regard to step (ii), for the ($k_u = 1, k_w = 2$) model we take 48 starting points obtained by combining $\xi_2 \in \{0.25, 0.5, \ldots, 2.75, 3\}$ and $\tau_{11} \in \{0.2, 0.4, 0.6, 0.8\}$. For the ($k_u = 1, k_w = 3$) model, we consider the combinations arising from $\xi_2 \in \{0.5, 1, \ldots, 2.5, 3\}$, $\xi_3 \in \{\xi_2 + 0.5, \xi_2 + 1, \ldots, 2.5, 3\}$ and $\boldsymbol{\mathcal{T}} = (\tau_{11}, \tau_{12}, \tau_{13})$ given by the rows of

$$\begin{pmatrix}
0.2 & 0.2 & 0.6 \\
0.2 & 0.4 & 0.4 \\
0.2 & 0.6 & 0.2 \\
0.4 & 0.2 & 0.4 \\
0.4 & 0.4 & 0.2 \\
0.6 & 0.2 & 0.2
\end{pmatrix},$$

so that 90 starting points are obtained. For the models with $k_u > 1$, the scheme above is replicated for the $\boldsymbol{\xi}$ vector and for the conditional probability distributions

$$\tau_{u\cdot}^{-1}(\tau_{u1}, \ldots, \tau_{uk_w}) \qquad u = 1, \ldots, k_u.$$

Essentially, since these conditional distributions are kept constant for every u, the algorithm starts from the independence model. As mentioned above, the starting values for $\boldsymbol{\tau}_{u\cdot} = (\tau_{1\cdot}, \ldots, \tau_{k_u\cdot})$ are taken from the final estimates of the corresponding models with $k_w = 1$.

Appendix C: log-likelihood score function

The log-likelihood score function is

$$\begin{aligned}
\boldsymbol{s}(\boldsymbol{\theta}_\ell) &= \sum_{h=1}^{H} \boldsymbol{s}_h(\boldsymbol{\theta}_\ell) = \sum_{h=1}^{H} \frac{\partial}{\partial \boldsymbol{\theta}_\ell} \log P(\boldsymbol{Y}_h = \boldsymbol{y}_h \mid \boldsymbol{X}_h = \boldsymbol{x}_h) \\
&= \sum_{h=1}^{H} \frac{\frac{\partial}{\partial \boldsymbol{\theta}_\ell} \sum_{u=1}^{k_u} \sum_{w=1}^{k_w} P(\boldsymbol{Y}_h = \boldsymbol{y}_h \mid \boldsymbol{X}_h = \boldsymbol{x}_h, U_h = \psi_u, W_h = \xi_w)\tau_{uw}}{P(\boldsymbol{Y}_h = \boldsymbol{y}_h \mid \boldsymbol{X}_h = \boldsymbol{x}_h)} \\
&= \sum_{h=1}^{H} \left\{ P(\boldsymbol{Y}_h = \boldsymbol{y}_h \mid \boldsymbol{X}_h = \boldsymbol{x}_h)^{-1} \sum_{u=1}^{k_u} \sum_{w=1}^{k_w} \frac{\partial}{\partial \boldsymbol{\theta}_\ell} \exp(\ell_{huw})\tau_{uw} \right\},
\end{aligned}$$

with $\ell_{huw} = \log P(\boldsymbol{Y}_h = \boldsymbol{y}_h \mid \boldsymbol{X}_h = \boldsymbol{x}_h, U_h = \psi_u, W_h = \xi_w)$. Therefore, we have

$$\boldsymbol{s}(\boldsymbol{\theta}_\ell) = \sum_{h=1}^{H} \left\{ P(\boldsymbol{Y}_h = \boldsymbol{y}_h \mid \boldsymbol{X}_h = \boldsymbol{x}_h)^{-1} \sum_{u=1}^{k_u} \sum_{w=1}^{k_w} \exp(\ell_{huw}) \boldsymbol{d}_{huw} \right\},$$

where

$$\boldsymbol{d}_{huw} = \tau_{uw} \frac{\partial}{\partial \boldsymbol{\theta}_\ell} \ell_{huw} + \frac{\partial}{\partial \boldsymbol{\theta}_\ell} \tau_{uw}$$

and $\frac{\partial}{\partial \boldsymbol{\theta}_\ell} \ell_{huw}$ is computed by deriving the SLU-uw group specific expected complete log-likelihood. These two derivatives are identical (Oakes 1999). Finally, notice that to avoid numerical problems the score is computed as

$$\boldsymbol{s}(\boldsymbol{\theta}_\ell) = \sum_{h=1}^{H} \left[\exp(\ell_{h,\max} - \ell_h) \sum_{u=1}^{k_u} \sum_{w=1}^{k_w} \left\{ \exp(\ell_{huw} - \ell_{h,\max}) \boldsymbol{d}_{huw} \right\} \right],$$

where $\ell_{h,\max} = \max_{u,w} \ell_{huw}$ and $\ell_h = \log P(\boldsymbol{Y}_h = \boldsymbol{y}_h \mid \boldsymbol{X}_h = \boldsymbol{x}_h)$ is calculated relying on the same numerical trick, i.e.

$$\ell_h = \ell_{h,\max} + \log \sum_{u=1}^{k_u} \sum_{w=1}^{k_w} \left\{ \exp(\ell_{huw} - \ell_{h,\max}) \tau_{uw} \right\}.$$

References

Altman RM (2007) Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. J Am Stat Assoc 102:201–210

Bacci S, Pandolfi S, Pennoni F (2014) A comparison of some criteria for states selection in the latent Markov model for longitudinal data. Adv Data Anal Classif 8:125–145

Bakk Z, Kuha J (2018) Two-step estimation of models between latent classes and external variables. Psychometrika 83(4):871–892

Bartolucci F, Farcomeni A, Pennoni F (2013) Latent Markov models for longitudinal data. Statistics in the social and behavioural sciences. Chapman & Hall/CRC

Bartolucci F, Lupparelli M (2016) Pairwise likelihood inference for nested hidden Markov chain models for multilevel longitudinal data. J Am Stat Assoc 111:216–228

Bartolucci F, Lupparelli M, Montanari GE (2009) Latent Markov model for longitudinal binary data: an application to the performance evaluation of nursing homes. Ann Appl Stat 3:611–636

Bartolucci F, Montanari GE, Pandolfi S (2014) A comaparison of some estimation methods for latent Markov models with covariates. In: Proceedings of COMPSTAT 2014—21st international conference on computational statistics, pp 531–538

Bartolucci F, Montanari GE, Pandolfi S (2015) Three-step estimation of latent Markov models with covariates. Comput Stat Data Anal 83:287–301

Bartolucci F, Pennoni F, Vittadini G (2011) Assessment of school performance through a multilevel latent Markov Rasch model. J Educ Behav Stat 36:491–522

Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat 41:164–171

Carpenter I, Hirdes JP (2013) Using interRAI assessment systems to measure and maintain quality of long-term care. In: A good life in old age? Monitoring and improving quality in long-term care, chap 3. OECD Health Policy Studies, pp 93–139

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc Ser B 39:1–38

Di Mari R, Oberski DL, Vermunt JK (2016) Bias-adjusted three-step latent Markov modeling with covariates. Struct Equ Modeli Multidiscip J 23(5):649–660

Goodman LA (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61(2):215–231

Hirdes JP, Ljunggren G, Morris JN, Frijters DH, Finne Soveri H, Gray L, Björkgren M, Gilgen R (2008) Reliability of the interRAI suite of assessment instruments: a 12-country study of an integrated health information system. BMC Health Serv Res 8:277

Ip E, Zhang Q, Rejeski J, Harris T, Kritchevsky S (2013) Partially ordered mixed hidden Markov model for the disablement process of older adults. J Am Stat Assoc 108(502):370–384

Kim H, Jung Y-I, Sung M, Lee J-Y, Yoon J-Y, Yoon J-L (2015) Reliability of the interRAI long term care facilities (LTCF) and interRAI home care (HC). Geriatr Gerontol Int 15:220–228

Koukounari A, Moustaki I, Grassly NC, Blake IM, Basáñez MG, Gambhir M, Mabey DC, Bailey RL, Burton MJ, Solomon AW (2013) Using a nonparametric multilevel latent Markov model to evaluate diagnostics for trachoma. Am J Epidemiol 177:913–922

Lazarsfeld PF, Henry NW (1968) Latent strucuture analysis. Houghton Mifflin

Linzer DA, Lewis JB (2011) poLCA: an R package for polytomous variable latent class analysis. J Stat Softw 42(10):1–29

Little RJ (1995) Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc 90(431):1112–1121

Marino MF, Alfò M (2016) Gaussian quadrature approximations in mixed hidden Markov models for longitudinal data: a simulation study. Comput Stat Data Anal 94:193–209

Marino MF, Tzavidis N, Alfò M (2018) Mixed hidden Markov quantile regression models for longitudinal data with possibly incomplete sequences. Stat Methods Med Res 27(7):2231–2246

Maruotti A, Rocci R (2012) A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. Stat Med 31:871–886

Maruotti A, Rydén T (2009) A semiparametric approach to hidden Markov models under longitudinal observations. Stat Comput 19(4):381–393

McCullagh P (1980) Regression models for ordinal data (with discussion). J Roy Stat Soci Ser B 42(2):109–142

Montanari GE, Doretti M (2019) Ranking nursing homes' performances through a latent Markov model with fixed and random effects. Social Indic Res 146(1–2):307–326
Montanari GE, Doretti M, Bartolucci F (2018) A multilevel latent Markov model for the evaluation of nursing homes' performance. Biometr J 60(5):962–978
Montanari GE, Pandolfi S (2018) Evaluation of long-term health care services through a latent Markov model with covariates. Stat Methods Appl 27:151–173
Oakes D (1999) Direct calculation of the information matrix via the EM algorithm. J Roy Stat Soc Ser B 61:479–482
Oort FJ (2005) Using structural equation modeling to detect response shifts and true change. Qual Life Res 14(3):587–598
Pohle J, Langrock R, van Beest FM, Schmidt NM (2017) Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. J Agric Biol Environ Stat 22:270–293
Royall RM (1986) Model robust confidence intervals using maximum likelihood estimators. Int Stat Rev 54(2):221–226
Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Spagnoli A, Marino MF, Alfò M (2018) A bidimensional finite mixture model for longitudinal data subject to dropout. Stat Med 37(20):2998–3011
Sprangers MAG, Schwartz CE (1999) Integrating response shift into health-related quality of life research: a theoretical model. Social Sci Med 48(11):1507–1515
Stephens M (2000) Dealing with label switching in mixture models. J Roy Stat Soc Ser B 62:795–809
Visser MRM, Oort FJ, Sprangers MAG (2005) Methods to detect response shift in quality of life data: a convergent validity study. Qual Life Res 14(3):629–639
Welch LR (2003) Hidden Markov models and the Baum-Welch algorithm. IEEE Inform Theory Soc Newslett 53:1–13
White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48(4):817–838
Wiggins LM (1973) Panel analysis: latent probability models for attitude and behavior processes. Jossey-Bass
Zucchini W, MacDonald IL (2009) Hidden Markov models for time series, 1st edn. Chapman & Hall/CRC