MULTIVARIATE BEHAVIORAL RESEARCH
2024, VOL. 59, NO. 1, 17–45

Go Multivariate: Recommendations on Bayesian Multilevel Hidden Markov Models with Categorical Data

Sebastian Mildiner Moraga and Emmeke Aarts

Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University

ABSTRACT

The multilevel hidden Markov model (MHMM) is a promising method to investigate intense longitudinal data obtained within the social and behavioral sciences. The MHMM quantifies information on the latent dynamics of behavior over time. In addition, heterogeneity between individuals is accommodated with the inclusion of individual-specific random effects, facilitating the study of individual differences in dynamics. However, the performance of the MHMM has not been sufficiently explored. We performed an extensive simulation to assess the effect of the number of dependent variables (1–8), number of individuals (5–90), and number of observations per individual (100–1600) on the estimation performance of a Bayesian MHMM with categorical data including various levels of state distinctiveness and separation. We found that using multivariate data generally alleviates the sample size needed and improves the stability of the results. Moreover, including variables only consisting of random noise was generally not detrimental to model performance. Regarding the estimation of group-level parameters, the number of individuals and observations largely compensate for each other. However, only the former drives the estimation of between-individual variability. We conclude with guidelines on the sample size necessary based on the level of state distinctiveness and separation and study objectives of the researcher.

KEYWORDS

Hidden Markov models; multilevel modeling; intensive longitudinal data; categorical data; Bayesian statistics; Monte Carlo studies; individual random effects

Introduction

Due to technological advances such as smartphones and smartwatches, sensors, and automatic coding of video recordings, it has become relatively easy and affordable to collect data on groups of individuals with a high temporal resolution (see, for example, Ariens et al., 2020; Cabrera-Quiros et al., 2021; Hamaker & Wichers, 2017; Lemaignan et al., 2018; Orfanos et al., 2017; Shiffman et al., 2008; Walls & Schafer, 2012). This novel type of data, obtained using methods such as experience sampling, ecological momentary assessment, daily dairy studies, or ambulatory assessments, receives the name of intensive longitudinal data (ILD). ILD is increasingly common: from 54 publications indexed on PubMed including one of these terms in the title or abstract in 2000, rising to 984 and 1203 publications in 2020 and 2021, respectively. The steep increase in use of ILD reflects that researchers recognize the unique value these data have for studying behavioral phenomena that unfold over time (see e.g., Hamaker & Wichers, 2017; Shiffman et al., 2008; Walls & Schafer, 2012). ILD avoids the risk of recall bias that is characteristic of traditional survey methodologies such as retrospective self-reported diaries, and ecological validity is increased compared to traditional experimental designs (Ebner-Priemer & Trull, 2009; Mehl & Conner, 2012). In addition, ILD facilitates the assessment of contextual relationships between factors influencing the process under study as it develops over time (Shiffman et al., 2008). Most importantly, due to the high sampling frequency of ILD, individual differences on the dynamics of social and behavioral phenomena can be assessed (Hamaker & Wichers, 2017; Walls et al., 2006). The increasing availability of this novel type of data spurred the development of new and expanding existing modeling techniques to fully exploit the wealth of information contained within ILD.

One such technique is the hidden Markov model (HMM; Rabiner, 1989; Zucchini et al., 2017). The HMM is a probabilistic method that can be used to study processes characterized by switches between discrete latent (hidden) states. Since the objects of study of social and behavioral research often consist of processes that cannot be directly observed, the inclusion of a (discrete) latent process provides a promising framework. In addition, multiple observed variables can link to the hidden states, often present when inferring latent constructs in social and behavioral processes. The HMM has a long record of success in the analysis of sequence (e.g., intense longitudinal) data in varying fields of research, such as speech recognition (Rabiner, 1989), human activity recognition (Ronao & Cho, 2017), animal behavior (Bode & Seitz, 2018), and DNA labeling (Rueda et al., 2013). When applied to social and behavioral processes, the HMM can be used to quantify information on the latent temporal dynamics of behavior into two sets of parameters: (1) the probability of transitioning between each of the hidden states and (2) the probability of emitting an observation given the current hidden state.

Recently, HMMs have been extended to the multilevel framework and take the name of multilevel HMMs (MHMMs, also known as mixed HMMs; Altman, 2007). The multilevel framework is particularly useful in social and behavioral processes, as individual level heterogeneity is to be expected in many applications. Within this framework, the overall temporal dynamics are reflected by a set of group-level parameters, and variability between individuals is accommodated by the inclusion of individual level random effects (Gelman et al., 2013). Additionally, quantifying the variability between individuals poses the invitation to explain the observed heterogeneity by including time-invariant (i.e., measured at the individual level) or time-variant (i.e., measured at the occasion level) covariates. All this combined, MHMMs constitute a powerful vehicle to study the temporal dynamics of social and behavioral processes and can be used in a wide range of applications. For example, to gain insights on the dynamics between "positive," "neutral," or "negative" interactions between adolescents and their parents in interviews (de Haan-Rietdijk et al., 2017), stages of drinking behavior (Shirley et al., 2012), patterns of bird behavior ("resting," "minimal activity," "moderate activity," and "flying"; McClintock et al., 2020), or "good" and "poor" driving behavior (Jackson et al., 2015). On a comprehensive simulation study, McClintock (2021) showed that accommodating continuous random effects within the model is valuable: random effects accounting for unexplained individual variation can improve estimation of state transition probabilities and measurable covariate effects, but discrete random effects can be a relatively poor (and potentially misleading) approximation for continuous variation.

However, to be able to reliably use this novel method, a sufficient amount of data needs to be collected. In the MHMM, the amount of data is a composite of the number of observations per individual, the number of individuals, and the number of outcome variables measured. As the MHMMs are relatively novel models, what constitutes a sufficient amount in the context of ILD is still largely unexplored. An overview of previous Monte Carlo studies examining model performance of the MHMM is provided in Table 1. The majority of Monte Carlo studies either concern longitudinal data (i.e., a large number of individuals combined with few longitudinal measurements; see number 1–9 in Table 1), or were designed to validate a specific MHMM for a given empirical application without manipulating any of the conditions (i.e., number 1–9 and 15 in Table 1).

Our study aims to set a primer for the construction of systematic guidelines concerning the number of dependent variables, individuals and observations per individual required to ensure reliable results fitting a MHMM. We focus on intense longitudinal data tailored to social and behavioral processes characterized by categorical observations, and factor in various levels of noisiness in the data. To this end, we explore the effect of three design factors (number of dependent variables measured, number of individuals, and number of observations per individual) on the accuracy of the MHMM with data of different levels of state distinctiveness and separation. Our Monte Carlo simulation is designed to address the following research questions: RQ1: *How does state distinctiveness and separation impact the estimation performance of the Bayesian MHMM?* RQ2: *How does the use of multivariate data impact the estimation performance of the Bayesian MHMM?*, and RQ3: *How does the estimation performance of the Bayesian MHMM vary with sample size (i.e., number of individuals and observations)?*

Over the next sections we examine the previous work done on each of these factors and draw hypotheses in relation to the estimation performance of the MHMM.

State distinctiveness and separation

The reliability of the estimated parameters of an HMM depends on the level of state distinctiveness and separation. The level of state distinctiveness in an HMM reflects the amount of randomness or

Table 1. Previous literature of Monte Carlo simulations concerning the multilevel hidden Markov model.

Reference	Estimation framework	Number of data sets	Number of individuals	Number of observations per individual	Number of states	Emission distribution	Number of dependent variables	Between ind. variance
Longitudinal data								
1 Altman (2007)	ML	200	30	20	2	Count	1	N.A.
2 Jackson et al. (2015)	ML	1000	60	20	2	Binary and count	2	1.00
3 Maruotti and Rydén (2009)	ML	250	100, 500, 1000	10	2	Count	1	0.10, 0.50
4 Xia et al. (2016)	ML	100	500, 1000	4, 5, 10	2	Binary	1	N.A.
5 Lin et al. (2020)	Bayesian	100	100	10	2	Continuous	1	1.00
6 Xia and Tang (2019)	Bayesian	N.A.	100, 300, 400, 500, 800, 1000	4	2	Continuous	6	N.A.
7 Kang et al. (2019)	Bayesian	100	700	9	2	Continuous	1	N.A.
8 Raffa and Dubin (2015)	Bayesian	200	354	6	3	Binary and continuous	2	N.A.
9 Zhang et al. (2014)	Bayesian	100	400	6	2	Binary	1	1.00
Intense longitudinal data								
10 Brekkan et al. (2019)	ML	100	500	60	2	Continuous	1	N.A.
11 Inaba (2017)	ML	300	10	5, 10, 20, 30, 40	2	Count	1	1.10
12 McClintock (2021)	ML	400	5, 15, 30, 50, 100	30–250	2	Continuous	1	0.20, 0.42
13 Chiang et al. (2018)	Bayesian	N.A.	100	10–100	3	Count	1	0.01
14 Park (2012)	Bayesian	N.A.	10, 20, 30, 40	20, 40, 60, 80, 100	2	Continuous	1	N.A.
15 Rueda et al. (2013)	Bayesian	100	25	150	3	Continuous	1	0.10–1.00

Note: Studies were classified on "longitudinal data" and "intense longitudinal data" depending on a number of observations per individual <20 or ≥ 20, respectively; in column "estimation framework," "ML" refers to maximal likelihood estimation, "Bayesian" refers to Bayesian estimation; "N.A." refers to "not available" which indicates that the corresponding value was not reported.

variability in the observed data within a state. This "noise" is reflective of how well the observed data fit the underlying state conditional probability distribution. High state distinctiveness means that the emission distribution of a state has a clear, recognizable, and predictable signal, which generally reflects a low level of noise in the emission distribution. In the context of categorical data, this is evident when one or a few categories display a high state-dependent emission probability, while all other categories exhibit near-zero emission probabilities. On the other hand, low state distinctiveness implies that the observed data is less predictable for a state, indicating a high amount of noise in the emission distribution. In categorical data, this is evident when there is a reduced difference between state-dependent emission probabilities.

The degree of state separation in an HMM refers to the extent to which the probability densities of the observed data generated by those states overlap. When two or more states have similar emission probability distributions, their overlap is high and state separation low, which can negatively impact the estimation performance of the HMM as the states are harder to infer. Although this concept is generally accepted in the literature, few studies have examined this effect in a simulation study (e.g., Beyer et al., 2013; Jonsen, 2016; McClintock, 2021; Ruiz-Suarez et al., 2022). The one study in Table 1 that examined such an effect, McClintock (2021), found that the degree of state separation was the most important factor affecting state estimation, with the performance of each model declining as state separation decreased. Additionally, low state separation appeared to have a negative, albeit more moderate, effect on the estimation of transition and emission parameters, particularly for smaller sample sizes.

The effect of the degree of state distinctiveness and separation on the estimation performance of HMM parameters depend on several factors, such as the number of hidden states, the size of the data set, and the type of algorithm used for parameter estimation (McClintock, 2021; Ruiz-Suarez et al., 2022). We expect that for data with a lower degree of state distinctiveness and separation, the model will require a larger sample size to produce the same degree of accuracy in the estimation of parameters compared with data with a higher degree of state distinctiveness and separation.

Using multivariate data

One component that defines the estimation performance in latent variable methods, and possibly the MHMM, is the number of dependent variables observed (also known as indicators or outcome variables) used to train the model. To the best of our knowledge, none of the previous simulation studies explored the effect of the number of dependent variables on the estimation of the MHMM (i.e., all studies in Table 1 concerning ILD considered univariate data. Study 2, 6, and 8 in Table 1 concerning longitudinal data did contain multivariate data but did not vary the number of dependent variables). Notwithstanding, empirical applications of

MHMMs in the literature have considered a range between 1 (de Haan-Rietdijk et al., 2017; Jackson et al., 2015; Schafer et al., 2020; Shirley et al., 2012) and 7 (DeRuiter et al., 2017) dependent variables.

Evidence for the benefits of increasing the number of dependent variables linked to the state on model performance is present in closely related literature: a number of studies belonging to the field of latent variable modeling (i.e., without a time component) investigated the effect of the number of latent class indicators (i.e., dependent variables) on several aspects of the estimation performance. On the one hand, the inclusion of a larger number of indicators makes for a more complex model with more response pattern possibilities, which has been shown to lead to boundary parameter estimation issues in latent class analysis under certain circumstances (Galindo Garre & Vermunt, 2006), and can cause numerical problems in estimation algorithms (Vermunt & Magidson, 2004).

However, increasing the number of indicators also has been shown to reduce bias and convergence issues in structural equation models (Marsh et al., 1998), improve classification accuracy and parameter coverage in multilevel latent class analyses (Finch & French, 2014), and improve convergence and reduce bias in latent class analysis (Wurpts & Geiser, 2014). In these studies, the number of indicators and the sample size compensated for each other, with the most beneficial effect of the number of indicators for smaller sample sizes. In addition, in the field of latent transition analysis, Collins and Wugalter (1992) show that increasing the number of indicators reduces standard errors and improve parameter recovery.

We hypothesize that increasing the number of dependent variables will have a positive effect over the estimation performance of the model, since more information is available to ensure an accurate inference of the hidden states.

Number of individuals and number of observations

The number of individuals and the number of observations per individual are well known factors that can affect the estimation performance of MHMMs and longitudinal latent variable models in general. Although a compensation effect between the two has been documented in longitudinal latent variable models such as continuous time models (Hecht & Zitzmann, 2021) and dynamic structural equation modeling (Schultzberg & Muthén, 2018), the effect of sample size on both levels and how they compensate for each other in MHMM parameter estimation is far from well understood. A number of studies have explored the relationship between the number of individuals and number of observations per individual over the estimation performance of MHMMs with simulation studies (i.e., rows 3, 4, 6, 11, 12, and 14 in Table 1). However, most studies only concerned a limited number of manipulated factors and selected levels, none of the studies considered categorical observations, and the studies were restricted to relatively short time series of maximally 250 observations per individual.

For example, Park (2012) introduced and examined the MHMM as a tool to track and describe the existence of hidden changes in panel data analysis, manipulating the number of observations up to 100. Simulated data sets consisted of two hidden regimes with only one transition occurring between regimes over the entire sequence. Both Inaba (2017) and Chiang et al. (2018) studied model performance of the MHMM for clinical research, including up to 40 and 100 occasions, respectively, lower than the number of observations we envision (i.e., $\geq$ 100). In addition, the number of individuals was kept constant at 10 and 100, respectively. Inaba (2017) concluded that for the studied scenarios, model performance was acceptable when the number of observations was at least 20.

In the field of ecology, McClintock (2021) published the single large comprehensive Monte Carlo simulation regarding the effects of sample size on the estimation performance of a (frequentist) MHMM. Although the study's scenarios focus solely on a univariate continuous distribution (specifically, a Gamma distributed variable), and the random effects only concern the probabilities of transitioning between states, some general insights into the impact of sample size can be distilled. Based on the simulation results presented in the study, it appears that larger sample sizes, both in terms of the number of individuals and the number of observations per individual, tend to result in more accurate state assignment and parameter estimation (specially for data with a highly overlapped emission distribution). A compensation effect between the two factors is also supported by the results. In addition, reliable estimation of between-animal variation (i.e., individual-specific random effects) requires at least 50 animals in the studied conditions, with scenarios characterized by high between-animal variation showing the best performance in this regard.

We hypothesize that increasing the sample size in terms of number of individuals and number of

observations per individual will have a positive impact on the estimation of the model. We expect a compensation effect between the two, with the number of individuals, the most important factor for the estimation of between individual variances. The remainder of the paper is organized as follows: to begin, we introduce the HMM and MHMM. We then conduct a large scale simulation study evaluating the performance of the MHMM under various sampling and data outcome conditions. An empirical example illustrating the application of the MHMM on non-verbal patient-therapist communication is also presented. We conclude with findings and recommendations.

The hidden Markov and multilevel hidden Markov model

In this section we present a brief introduction to the HMM and describe approaches to apply the HMM to a group of individuals. Subsequently, we show how the MHMM extends the modeling capabilities of the HMM and discuss inference of the MHMM using Bayesian estimation.

The hidden Markov model

The HMM is a statistical method that is used to infer a sequence of latent or hidden states $S_t \in (1, 2, ..., M)$ for time points $t = 1, ..., T$ which are defined by the probability to observe an outcome Y_t, and account for the dynamics of the observations in terms of the dynamics of the hidden states. The former is based on the assumption that a given observation Y_t in the sequence is generated by an underlying, latent state S_t. The latter is based on the assumption that the hidden states are not identically and independently distributed, but they instead follow a Markov process. That is, the probability of switching from state i at time point t to state j at $t+1$ only depends on the departing state i at time point t. Figure 1 details the general structure of a basic HMM.

The likelihood function L_T of the HMM is defined by three sets of parameters: the initial state probabilities $\boldsymbol{\pi}$, the transition probability matrix (TPM) $\boldsymbol{\Gamma}$ and the state-dependent emission distributions $p(y)$. If we write the set of emission distributions in matrix notation $\mathbf{P}(Y)$, such that $\mathbf{P}(Y)$ is a M by M diagonal matrix with ith diagonal element the state-dependent emission probability density $p_i(y)$, the likelihood function has a convenient matrix notation:

$$L_T = \boldsymbol{\pi}\mathbf{P}(Y_1)\boldsymbol{\Gamma}\mathbf{P}(Y_2)\boldsymbol{\Gamma}\mathbf{P}(Y_3)...\mathbf{P}(Y_T)\mathbf{1}'. \quad (1)$$

The initial probability $\pi_i = P(S_1 = i)$ denote the probability of each state $S \in (1, 2, ..., M)$ for $t = 1$. Often, the initial probabilities of the states are not estimated freely but are assumed to be the stationary distribution implied by the transition probability matrix $\boldsymbol{\Gamma}$, which we will adhere to for this study as well. The transition probability matrix $\boldsymbol{\Gamma}$ with transition probabilities

$$\gamma_{ij} = P(S_{t+1} = j | S_t = i) \quad (2)$$

denote the probability of switching from state $i \in (1, 2, ..., M)$ at time t to state $j \in (1, 2, ..., M)$ at time $t+1$. Here, we follow the implementation in Altman (2007) to facilitate the inclusion of random effects in subsequent multilevel models below, which assumes that the rows of the transition probability matrix $\boldsymbol{\Gamma}$ are independent of each other, and are modeled using a multinomial logit model

$$\gamma_{ij} = \frac{\exp(\alpha_{ij})}{1 + \sum_{l=2}^{m} \exp(\alpha_{il})}. \quad (3)$$

The numerator is set equal to 1 for $j = 1$, making the first state of every row of the transition probability matrix $\boldsymbol{\Gamma}$ the baseline category. Hence, the probability to transition from hidden state $i \in (1, 2, ..., M)$ to state $j \in (1, 2, ..., M)$ at time t is modeled with M batches of fixed intercepts $\boldsymbol{\alpha}_{i.}$, each containing $M-1$ elements, $\alpha_{i2}, ..., \alpha_{iM}$.

In this paper we focus on categorical data, as such the state-dependent emission distribution

$$p_i(y) = P(Y_t = y | S_t = i, \boldsymbol{\theta}_i) = P(Y_t = q | S_t = i, \boldsymbol{\theta}_i) = \theta_{iq} \quad (4)$$

denotes the probability of observing category $q \in 1, ..., Q$ for Y given the hidden state $i \in 1, ..., M$ at time t with parameter set $\boldsymbol{\theta}_i$. Given the categorical nature of the data, $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iQ})$ is a vector of multinomial emission probabilities. Similar to the transition probabilities γ_{ij}, the emission probabilities θ_{iq} are modeled using a multinomial logit model

$$\theta_{iq} = \frac{\exp(\beta_{iq})}{1 + \sum_{l=2}^{Q} \exp(\beta_{il})}. \quad (5)$$

The numerator is set equal to 1 for $q = 1$, making the first category the baseline category. Hence, the probability to observe $y_q \in (1, 2, ..., Q)$ given the state $i \in (1, 2, ..., M)$ at time t is modeled with M batches of fixed intercepts $\boldsymbol{\beta}_{i.}$, each containing $Q-1$ elements, $\beta_{i2}, ..., \beta_{iM}$. However, note that the shape of the emission distributions is flexible and depends on the data observed (e.g., Normal for continuous data, or Poisson for count data).

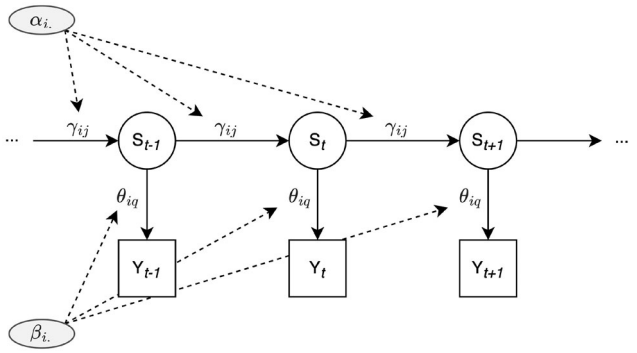

Figure 1. Directed acyclic graph of a basic hidden Markov model with a univariate emission distribution. Each hidden state *S*, depicted with circles, depends only on the state at the previous time point. The observed data *Y*, depicted with squares, depends only on the value of the current latent state *S*. Superimposed are the model parameters, with transition probabilities γ_{ij} and emission probabilities θ_{iq}, and the multinomial logit intercepts $\boldsymbol{\alpha}_{i.}$ and $\boldsymbol{\beta}_{i.}$ pertaining to the transition probabilities and the emission probabilities, respectively.

To extend the HMM framework to accommodate multivariate data, it is necessary to specify a joint distribution $P(Y_{t1} = y_{t1}, Y_{t2} = y_{t2}, ..., Y_{tk} = y_{tK})$ for the $k \in (1, 2, ..., K)$ state-dependent emission distributions. Here, the shape of the emission distributions can vary over the dependent variables, which are typically assumed to be conditionally independent given the sequence of hidden states (Zucchini et al., 2017). Thus, the joint distribution can be expressed as the product of the K marginal state-dependent emission probability densities $p_{ki}(y) = P(Y_{tk} = y_{tk}|S_t = i)$:

$$P(Y_{t1} = y_{t1}, Y_{t2} = y_{t2}, ..., Y_{tk} = y_{tk}) = \prod_{k=1}^{K} P(Y_{tk} = y_{tk}|S_t = i) \quad (6)$$

As a result, the likelihood of the multivariate HMM takes the form:

$$L_T = \prod_{k=1}^{K} \boldsymbol{\pi}\mathbf{P}_k(Y_{1k})\boldsymbol{\Gamma}\mathbf{P}_k(Y_{2k})...\mathbf{P}_k(Y_{Tk})\mathbf{1}'. \quad (7)$$

In this paper, all observed outcome variables are composed of categorical data. As such, the vector of multinomial emission probabilities $\boldsymbol{\theta}_i$ and the used multinomial logit model given in Equation (11) become outcome variable k dependent in the multivariate case

$$\theta_{kiq} = \frac{\exp(\beta_{kiq})}{1 + \sum_{l=2}^{Q} \exp(\beta_{kil})}. \quad (8)$$

Hence, the probability to observe $y_{kq} \in (1, 2, .., Q)$ given the state $i \in (1, 2, .., M)$ for outcome variable $k \in (1, 2, ..., K)$ at time t is modeled with $K \times M$ batches of fixed intercepts $\boldsymbol{\beta}_{ki.}$, each containing $Q - 1$ elements, $\beta_{ki2}, ..., \beta_{kiM}$. Figure 2 details the structure of a multivariate HMM.

Note that in the HMM, the number of states is to be determined *a priori* by the researcher. When a theoretical justification for a specific number of hidden states is lacking, deciding on the number of states becomes a model selection problem for which standard model selection criteria are often used (e.g., the AIC or BIC). However, currently there is no consensus on the optimal way of selecting the number of hidden states used; see Pohle et al. (2017) for a discussion on this matter.

In addition, note that the HMM is also known as the latent Markov model (LMM; Vermunt et al., 1999; Wiggins, 1973) and latent transition analysis (LTA; Hagenaars & McCutcheon, 2002; Nylund-Gibson et al., 2022). Although the names refer to the same model type, there are some key differences important to our study. Both LMMs and LTA models have predominantly been applied to longitudinal data (e.g., a large number of individuals combined with few longitudinal measurements), also referred to as panel data. As such, computing the likelihood and parameter estimation also proceeds differently (see e.g., Visser, 2011). In addition, LMMs have predominantly been applied to take measurement error into account, and in LTA the probability to transition between hidden states is generally estimated separately for each point in time, feasible as only a small number of transitions over time are inferred. In this paper, we focus on the application of the HMM in data with long sequences of observations (e.g. $\geq$100), the number of observations per individual outnumbering the number of individuals measured, the hidden states resembling a theoretical construct beyond measurement error reduction, and the probability to transition between states assumed to be time-homogeneous.

HMMs applied to a group of individuals

Traditionally, HMMs were applied to single time-series such as speech (e.g., Rabiner, 1989) or DNA and amino-acidic sequences to accomplish a labeling task (e.g., Rueda et al., 2013). In contrast, social and behavioral data often consist of sequential observations collected for several individuals. When applying the HMM to multiple sequences of data, researchers can fit a "complete pooling" standard HMM assuming a common set of parameters shared among the N individuals. When explanatory individual covariates

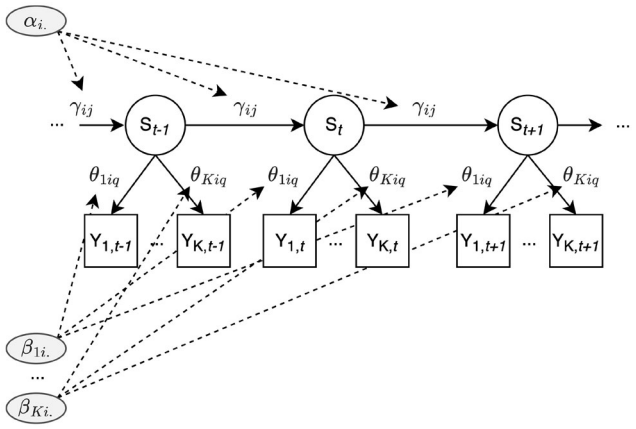

Figure 2. Directed acyclic graph of a basic hidden Markov model with a multivariate emission distribution. Each hidden state S, depicted with circles, depends only on the state at the previous time point. The observed data $Y_1, ..., Y_K$ for outcome variables $k \in (1, 2, ..., K)$, depicted with squares, depends only on the value of the current latent state S. Superimposed are the model parameters, with transition probabilities γ_{ij} and emission probabilities θ_{kiq}, and the multinomial logit intercepts $\boldsymbol{\alpha}_{i.}$ and $\boldsymbol{\beta}_{1i.}, \ldots, \boldsymbol{\beta}_{Ki.}$ pertaining to the transition probabilities and the emission probabilities of the K dependent variables, respectively.

are available, model parameters can be made conditional on these attributes. With this approach, (residual unexplained) individual variation is not accounted for or quantified. On the other side of the spectrum, one can accommodate generic individual heterogeneity in HMMs by fitting a separate model for each individual. However, fitting a separate HMM for each individual results in inefficient use of the collected data: in fitting the HMM on one individual, information available for the other individuals is discarded (Gelman & Carlin, 2014; Gelman et al., 2013). Moreover, nothing ensures that the content of the states is similar across models, making individual comparisons cumbersome. A related approach is the individual fixed effects model (e.g., Jonsen, 2016; McClintock & Michelot, 2018), in which for one of the components of the model the heterogeneity between individuals is handled using individual fixed effects ("no pooling") whereas for the other, the parameters are restricted to be the same across individuals ("complete pooling"). These approaches lead to highly parameterized models that avoid distributional assumptions but are hard to interpret. An intermediate approach to accommodate individual variation that improves on the parameterization burden is to use a mixture Markov model. In the mixture HMM a number of latent classes that differ with regard to the model parameters are specified, while individuals within each class are restricted to share a common set of parameters (e.g., Bartolucci & Farcomeni, 2015; Bartolucci et al., 2012; DeRuiter et al., 2017; Langrock et al., 2012; Maruotti, 2011; Maruotti et al., 2022; Maruotti & Rocci, 2012; McKellar et al., 2015; Towner et al., 2016). Hence, the mixture HMM assumes that the data contains a limited number of homogeneous subgroups, and is equal to including a set of discrete, non-parametric random effects. However, McClintock (2021) showed that approximating individual level effects with a finite mixture HMM is a relatively poor and potentially misleading approximation in case that variation between individuals is continuous.

Neither of these three approaches quantify the heterogeneity between individuals, and, in case of the latter two, ignore the unique opportunity of ILD to obtain an individual description of each person's process. To fully exploit the information contained within ILD, a method that includes these features is key (for a discussion on this topic we refer to Hamaker & Wichers, 2017). The HMM within the multilevel framework, discussed in the next section, does include these opportunities.

The multilevel hidden Markov model

Altman (2007) proposed a multilevel HMM (also mixed-effects HMM) to simultaneously model the sequences of multiple individuals. Her approach consists of a frequentist, hierarchical implementation that allows for estimating a group-level set of parameters, while accommodating the variability at the individual level parameters through (independent and identically distributed normal) random effects. In addition, the variability observed between individuals can be further explained by including time-varying and time-invariant covariates. This model is indistinctly referred to as the mixed-effects HMM or the multilevel HMM (MHMM), the latter name we use throughout this study. Figure 3 details the structure of a multivariate MHMM.

Within the multilevel framework, the intercepts of the multinomial logit model pertaining to transition probabilities γ_{ij} (see Equation 11) are allowed to vary over individuals, resulting in individual n specific transition probabilities γ_{nij}

$$\gamma_{nij} = \frac{\exp(\alpha_{nij})}{1 + \sum_{s=2}^{M} \exp(\alpha_{nis})}, \tag{9}$$

where

$$\alpha_{nij} = \bar{\alpha}_{ij} + \epsilon_{[\alpha]nij} \tag{10}$$

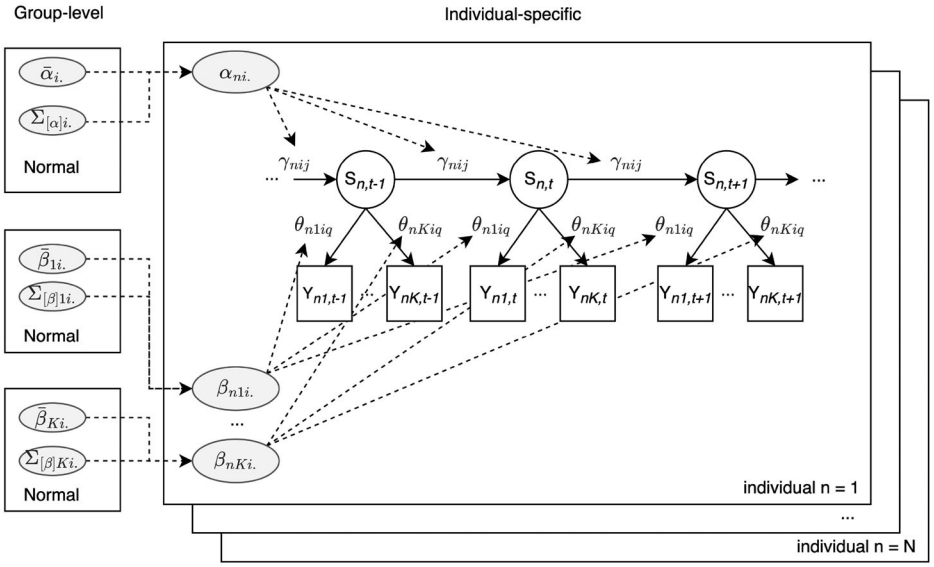

Figure 3. Directed acyclic graph of a multilevel hidden Markov model with a multivariate emission distribution. Each hidden state S_n for individual $n \in (1, 2, ..., N)$, depicted with circles, depends only on the state at the previous time point. The observed data $Y_{n1}, ..., Y_{nK}$ for outcome variables $k \in (1, 2, ..., K)$, depicted with squares, depends only on the value of the current latent state S_n. Superimposed are the model parameters, with individual n specific transition probabilities γ_{nij} and emission probabilities θ_{nkiq}, and the individual specific multinomial logit intercepts $\boldsymbol{\alpha}_{ni.}$ and $\boldsymbol{\beta}_{n1i.}, \ldots, \boldsymbol{\beta}_{nKi.}$ pertaining to the transition probabilities and the emission probabilities of the K dependent variables, respectively. Also depicted are the group-level mean $\bar{\alpha}_{i.}$ and covariance $\boldsymbol{\Sigma}_{[\alpha]i.}$ of the Normal distribution on $\boldsymbol{\alpha}_{ni.}$, and the group-level means $\bar{\boldsymbol{\beta}}_{1i.}, \ldots, \bar{\boldsymbol{\beta}}_{Ki.}$ and covariances $\boldsymbol{\Sigma}_{[\beta]1i.}, \ldots, \boldsymbol{\Sigma}_{[\beta]Ki.}$ of the Normal distribution on $\beta_{n1i.}, \ldots, \beta_{nKi.}$.

for each individual $n \in (1, ..., N)$ and each $i, j \in (1, ..., M)$, with $\bar{\alpha}_{ij}$ being the (group-level) average logit for transitioning from state i to state j, and $\epsilon_{[\alpha]nij}$ denoting the individual n's deviation from that average. The individual level random effects $\epsilon_{[\alpha]nij}$ follow a state i dependent multivariate normal distribution with zero mean vector of length $(M-1)$ and covariance matrix $\boldsymbol{\Sigma}_{[\alpha]}$ with $(M-1)$ rows and columns. Throughout the remainder of this text we refer to the group-level average logit $\bar{\alpha}_{ij}$ as the *TPM group-level fixed effects*, and the (diagonal) variance components $\sigma^2_{[\alpha]ij}$ from the covariance matrix $\Sigma_{[\alpha]}$ as the *TPM individual random effects*.

For the individual categorical emission probabilities θ_{nkiq} we follow an identical specification

$$\theta_{nkiq} = \frac{\exp(\beta_{nkiq})}{1 + \sum_{l=2}^{Q} \exp(\beta_{nkil})}. \tag{11}$$

where

$$\beta_{nkiq} = \bar{\beta}_{kiq} + \epsilon_{[\beta]nkiq} \tag{12}$$

with $\bar{\beta}_{kiq}$ being the (group-level) average logit for observing outcome category q within state i for outcome variable k, and $\epsilon_{[\beta]nkiq}$ denoting the individual n's deviation from that average. The individual level random effects $\epsilon_{[\beta]nkiq}$ follow a state i and outcome variable k dependent multivariate normal distribution with zero mean vector of length $(Q-1)$ and covariance matrix $\boldsymbol{\Sigma}_{[\beta]}$ with $(Q-1)$ rows and columns. We refer to the group-level average logit $\bar{\beta}_{iq}$ as the *EPM group-level fixed effects*, and the variance components $\sigma^2_{[\beta]kiq}$ from the covariance matrix $\boldsymbol{\Sigma}_{[\beta]}$ as the *EPM individual random effects*.

Bayesian estimation of the MHMM

We focus on Bayesian estimation of the model parameters, as it poses several advantages over classical (frequentist) estimation methods (e.g., maximization of the likelihood by either direct numerical maximization or the expectation maximization (EM) algorithm Dempster et al. (1977), known as the Baum–Welch algorithm in the context of HMMs (Baum et al., 1970; Rabiner, 1989). See Cappé et al. (2005) for a review of estimation methodologies, and Rydén (2008) for a comparison on frequentist and Bayesian approaches for the HMM.

While Altman (2007) showed that estimation of the MHMM is feasible using frequentist methods, the dimensions of the integral in the MHMM make fitting these methods largely intractable for more complex models (e.g., including more than two states or including more than three or four random effects). As such, frequentist applications accommodating individual level variation typically limit the random effects to a set of discrete, non-parametric random effects

within the framework of mixture HMMs as described above. In contrast, the estimation of continuously distributed random effects is commonplace in the Bayesian framework (Gelman et al., 2013), achieved by implementing hierarchical priors in a multilevel (hierarchical) structure. Bayesian estimation of the MHMM with multiple continuous random effects was first introduced by Zhang et al. (2010), who showed that it was attainable and efficient. Hence, the Bayesian framework enables the inclusion of continuous random effects for both the parameters of the transition distribution and the emission distribution simultaneously while retaining a computationally feasible model.

In addition to the computational advantage, Bayesian estimation of MHMMs poses five additional advantages we believe relevant to social and behavioral research. First, preliminary information and researcher beliefs can be structurally included in the model through its hierarchical priors (Gelman et al., 2013; Lynch, 2007; McElreath, 2020). As a result, previous evidence on the process under study can be combined with novel data to train the model and extract new insights. Nevertheless, non-informative and weakly informative priors can also be chosen. Second, the hierarchical priors produce a regularization of the individual-level parameters, pooling them toward the group-level means making the model more robust to outliers (Gelman et al., 2013; Lemoine, 2019). Third, Bayesian estimation does not rely on asymptotic assumptions making it more suitable for inference on small samples (van de Schoot & Miočević, 2020). Fourth, Bayesian estimation of the model produces valuable by-products such as parameters' standard errors and local decoding of the sequence of hidden states (Scott, 2002). A fifth valuable feature of the Bayesian framework is the possibility to assess the goodness of fit of the model to the data by posterior predictive checks (PPCs; see e.g., Gelman & Carlin, 2014; Lynch & Western, 2004). That is, PPCs allow the researcher to assess whether the model recovers the data correctly on an array of characteristics and can aid in revealing model missspecification.

Given that some traditional SEM-derived fit indices such as the Likelihood ratio test may not be appropriate for multilevel longitudinal data (Wu et al., 2009), evaluating model fit with PPCs is particularly relevant. To perform a PPC, we first fit a MHMM to the data and generate simulated data sets based on the obtained parameter estimates. For each simulated data set, summary statistics of interest, such as the mean or variance, are then calculated and compared to those of the observed data. An empirical posterior predictive value (PPV) can be constructed ranging from 0 to 1. The PPV takes intermediate values if the observed and simulated data sets produce similar summary statistics, indicating a good model fit, and extreme values (i.e., <0.025 or >0.975) if the observed and simulated data sets differ significantly, suggesting a poor fit. Posterior predictive checks can be used to assess model fit at both the individual and group-levels in multilevel models, allowing us to determine whether the model captures different aspects of the data for individuals and groups (e.g. in MHMMs: de Haan-Rietdijk et al., 2017; Shirley et al., 2012).

One specific issue that may arise with Bayesian estimation of the MHMM is a phenomenon known as "label switching". Because the posterior distribution of the states is not identifiable, when a considerable overlapping exists between hidden states, the labels of the hidden states can switch around while sampling from the Markov chain Monte Carlo (MCMC). Notice that this occurs even though the complete data likelihood remains the same (Allman et al., 2009; Jasra et al., 2005). The occurrence of label switching makes interpreting the posterior parameter estimates meaningless. To reduce the chances of label switching, users can follow one or more of the following strategies: (a) choose a sensible set of starting values for the MCMC, (b) implement a constraint on the order of the emission distribution (e.g., ordering the means), or (c) use (weakly) informative priors to introduce a distinction on the likelihood of the states. Here we follow the approach (a).

A second issue that may arise with Bayesian estimation –and Markov chain Monte Carlo methods in a broader sense– is the lack of convergence of model parameters over the iterations of the MCMC. That is, when training a Bayesian model using different sets of starting values for the Markov chains and they fail to converge toward the same parameter estimates. Improving the likelihood of parameter convergence can be often achieved by choosing a sensible set of starting values for the MCMC. For example, basing the starting values on theoretical likely state compositions and transitions giving the studied process and small variations hereof, or using starting values based on the maximal likelihood estimates of a single-level expectation-maximization HMM trained on the full data set (i.e., a "complete pooling" standard HMM).

When using sensible starting values does not suffice, lack of convergence in model estimation can indicate either trying to estimate a model that is too complex for the available data or model

misspecification (Gelman & Hill, 2006, Ch. 19). To address the former, researchers may need to reduce the complexity of the model by, for example, decreasing the number of hidden states or constraining some parameters in the model. If all remedies fail, researchers may need to use a single-level (completely pooled) HMM with covariates to explain variability between individuals, but this comes at the cost of losing the ability to measure individual-specific parameters and may not ensure convergence of the parameters. To address the latter, exploring potential sources of misspecification through simulation and model fit evaluation, such as PPCs, prior predictive checks, and cross-validated probability integral transforms, can be helpful (for a review on these methods, we refer the reader to Conn et al., 2018). Misspecification can for example arise from selecting the wrong number of states, assuming conditional independence between variables that do not hold in the data, or falsely assuming a homogeneous transition distribution.

Simulation

The aim of the Monte Carlo simulation study was to empirically assess the performance of the MHMM on data of varying levels of state distinctiveness and separation, number of dependent variables, and sample sizes. For the main simulation (which we call "baseline" for the remainder of the text) we manipulated two outcome and three design factors in a fully factorial design. The outcome factors are defined by the degree of state distinctiveness and separation. The design factors concern the number of dependent variables measured and the sample size as defined by the number of individuals and number of observations per individual. A description of each of their levels follows below.

State distinctiveness and separation

The degree of state distinctiveness and separation both include three levels: high, moderate and low. In case of multivariate data, the level of state distinctiveness and separation is equal over the dependent variables. Within our simulation study the emission distributions are composed such that only one or two categories have a high emission probability for a given state *S*. We call these the "defining categories" of a state. We quantify the level of state separation by how many categories represent a defining category for multiple states (see Figure 4a). State separation is defined as follows in the first dependent variable. High state separation corresponds to each defining category uniquely characterizing one state only. In moderate state separation, the third category represents a defining category for both state two and three. In low state separation, the second category represents a defining category for states one and two, and the third category, is a defining category for states two and three. The identity of the states sharing the population-level emission distribution and the specific defining categories are reshuffled in the remaining the dependent variables, but follow the same pattern (see Table S1 of the Supplementary Materials).

In the context of a categorical emission distribution, the degree of state distinctiveness refers to the magnitude of the probabilities for the non-defining categories within a state. Given that all emission probabilities within a state sum to one, larger noise probabilities inevitably result in lower probabilities for the defining categories and hence lower state distinctiveness (see Figure 4b). As such, the noisiness can be seen as a signal to noise ratio of the defining category. We manipulate the level of state distinctiveness with the inclusion of the factor λ in the emission distribution probability matrix, allowing us to include noise by adding or subtracting fixed quantities from each emission probability. For the levels high, moderate and low state distinctiveness in the probability scale, $\lambda = \{0.03, 0.09, 0.15\}$, respectively. See Table S1 for an overview of the parameter values for the emission distributions of each of the dependent variables.

Number of dependent variables

For the "baseline" scenarios we manipulated the number of dependent variables (N_{dep}) to assess whether increasing N_{dep} to infer the hidden state sequence on alters model performance. We included three levels for number of dependent variables $N_{dep} \in (1, 2, 4)$ across the factorial design, and fourth level, $N_{dep} = 8$, only for the boundary scenarios with low and high state distinctiveness and separation in the emission distributions (see below). In the literature, the MHMM was fit using between 1 (de Haan-Rietdijk et al., 2017; Jackson et al., 2015; Schafer et al., 2020; Shirley et al., 2012) and 7 (DeRuiter et al., 2017) dependent variables. The observed data for each of the dependent variables are composed of $Q = 5$ categories each. We defined the emission distribution for each of the dependent variables to have a unique pattern within each of the states such that all dependent variables contribute unique information to the model.

In addition, we considered the effect of adding a second set of dependent variables $N_{dep} \in (1, 2, 4)$ only consisting in random noise on top of the same level of number of dependent variables in the baseline, only for the boundary scenarios with low and high state distinctiveness and separation in the emission distributions. We call this condition "noise," and refer to them as $N_{dep} \in (1+1, 2+2, 4+4)$. See Table S1 in the Supplementary Materials for an overview of parameter values for each of the emission distributions relating to the dependent variables.

Number of individuals and number of observations

Based on the ILD literature, we considered three levels for the number of individuals $N_{ind} \in (5, 30, 90)$ and four levels for the number of observations per individual $N_{obs} \in (100, 400, 800, 1600)$. In the literature on MHMM the range for the number of individuals was 6 to 500. We used a number of individuals of 5 to represent a smaller number of individuals (e.g., Schafer et al., 2020), although this number is generally too small to provide reliable inferences on the magnitude of individual variation in multilevel models (Hox et al., 2018). A number of individuals of 30 roughly represents the median number used in applications of the MHMM (e.g., Altman, 2007; Holsclaw et al., 2017; Jackson et al., 2015; Rueda et al., 2013). Limited by the computational intensiveness of the model, we set the upper limit of number of individuals to 90 to represent the inclusion of many individuals (e.g., Chiang et al., 2018; de Haan-Rietdijk et al., 2017; Shirley et al., 2012).

With respect to the number of observations per individual, applications of the MHMM exhibit a wide variability: in Altman (2007) N_{obs} ranged from 1 to 24, in Shirley et al. (2012) N_{obs} was set to 168, $N_{obs} =$ 539 in de Haan-Rietdijk et al. (2017), $N_{obs} =$ 365–3003 in Chiang et al. (2018), and N_{obs} was over 7000 in Schafer et al. (2020). Such a wide range makes it hard to come with a one-fits-all approach. In our Monte Carlo study we focus on applications suitable to investigate the temporal dynamics at an individual level, which requires a relatively large number of observations per individual. As such we have set the lower limit of number of observations per individual to 100. Limited by computational constraints, we set the upper level for the number of observations per individual to 1600.

Data generation and model fitting

For each of the 324 scenarios on the baseline and 168 additional scenarios ($N_{dep} = 8$ and "noise" dependent variables) we generated 100 simulated data sets. We limited the number of simulated data sets to 100 due to computational constraints. Data was generated fixing the number of hidden states at $m = 3$, with self transition probabilities $\bar{\gamma}_{11} = 0.90, \bar{\gamma}_{22} = 0.70, \bar{\gamma}_{33} =$ 0.50. Transition distribution individual random effects were drawn from a Normal distribution with mean

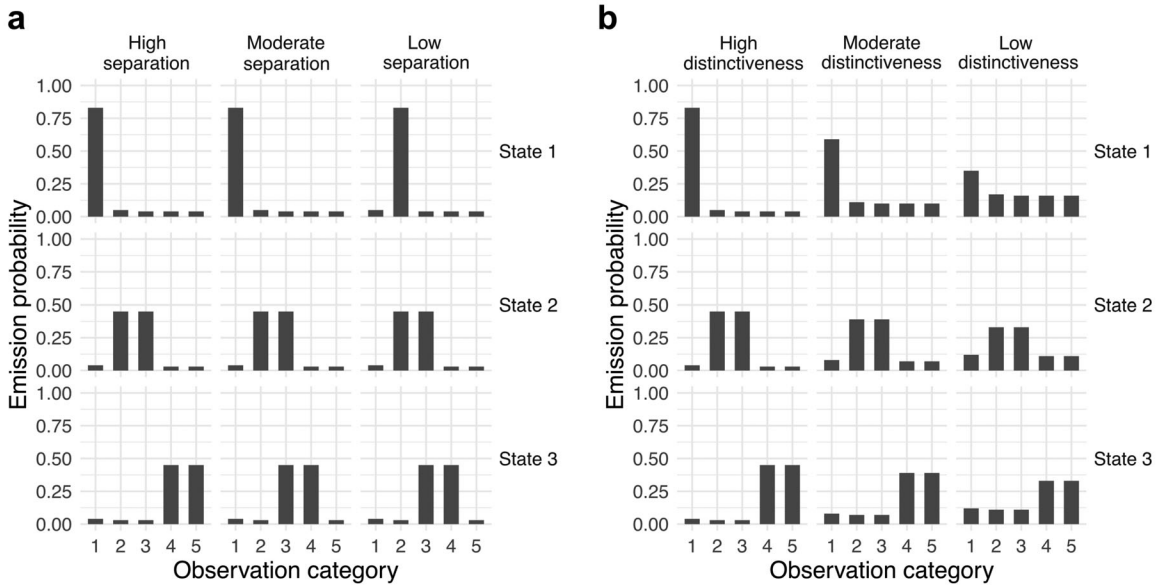

Figure 4. Emission probabilities for five observation categories for data with different levels of state separation and distinctiveness in its emission distribution. Panel (a) shows three levels of state separation ("high," "moderate," "low") on the emission probabilities of five observation categories, for a fixed state distinctiveness ("high"); panel (b) shows three levels of state distinctiveness ("high," "moderate," "low") on the emission distribution, for a fixed state separation ("high").

zero and variance ranging $\sigma^2_{[\alpha]ij} \in \{0.13 - 1.12\}$ on the logit scale, which translate to standard deviations on the self-transitions $\sigma_{\gamma_{11}} = \sigma_{\gamma_{22}} = \sigma_{\gamma_{33}} = 0.10$ in the probability scale. The same sampling scheme was followed for the individual random effects of the emission probability matrix on each dependent variable. A variance ranging $\sigma^2_{[\beta]ij} \in \{0.11 - 1.01\}$ was used, so that the standard deviation of the defining categories also equaled 0.10 in the probability scale. In fitting the model, non-informative normal hyper-priors were specified for all group-level parameters (refer to Section S3: Model specification on the Supplementary materials for the full hierarchical Bayesian specification of the model).

Simulating data and fitting of models is done in the statistical software R (R Core Team, 2021) using the package "mHMMbayes" specifically tailored to MHMMs using Bayesian estimation (Aarts, 2019). Corresponding R code to reproduce and analyze the simulation study can be found in Mildiner Moraga and Aarts (2022). In the mHMMbayes package, models are fitted using a hybrid Metropolis within Gibbs MCMC algorithm, and expands on the HMM implementation using Bayesian estimation as outlined in Scott (2002). That is, parameters are estimated using a forward pass of the forward backward algorithm (Baum & Petrie, 1966) to obtain the (forward) probabilities of each of the states for every time point in the sequence given the observed data, followed by sampling the hidden state sequence in a backward pass from their full conditional posteriors given the (current) parameters of the model. Conditional on the sampled hidden state sequence, the parameter estimates are updated by sampling them from their full conditional posteriors using either a Gibbs or a random walk Metropolis–Hastings step.[1] We used a total of 4000 MCMC iterations to train a three-state MHMM and dropped the first 2000 to attenuate the effect of the starting values (burn-in). To check convergence of the model parameters, for three randomly selected repetitions of each scenario an additional chain was fitted using different starting values. The convergence of the parameters of interest was assessed analytically with the multivariate potential scale reduction factor (MPSRF; Brooks & Gelman, 1998) using two different convergence criteria. Under the strict criterion-I, we considered that a model failed to converge if at list one of the group-level parameters of interest presented a MSRF > 1.2. Under the less restrictive criterion-II, we considered that a model failed to converge when the group-level parameters of interest presented a mean MSRF > 1.2, averaged separately for transitions and emission parameters and fixed and random effects. Finally, we note that model misspecification was not evaluated in this study, as the number of states used to simulate and fit the models were fixed at three.

Model performance was evaluated based on mean and relative mean bias, precision (empirical standard error), and coverage of the 95% confidence interval (CI) for the group-level parameters and the individual random effects on both components in the model. Model performance of the group-level parameters $\bar{\alpha}_{ij}$ and $\bar{\beta}_{ij}$ is presented on the probability domain ($\bar{\gamma}_{ij}$ and $\bar{\theta}_{iq}$) to aid interpretation. Note however that parameter estimation is on the logit domain. Model performance of the variance of the individual level random effects $\sigma^2_{[\alpha]ij}$ and $\sigma^2_{[\beta]ij}$ are presented on the logit domain. The mean proportion of hidden states correctly assigned by the Viterbi algorithm (Viterbi, 1967), also known as global state decoding, along with the mean proportion of instances in which the decoding probability of the true state was ≥ 0.2 were used as measures of the model performance for inference at the individual level. Kappa statistic was used to control for the expected classification accuracy due to random chance (Cohen, 1960). Part of the simulation results is presented using the nested loop plot (Rücker & Schwarzer, 2014). These novel plots offer a way of displaying a large number of simulation results. Scenarios are presented in a lexicographical order, arranged consecutively along the horizontal axis, while the evaluation metric is plotted on the vertical axis.

Results

The majority of the scenarios lead to accurate estimations of the group-level parameters of the transition probabilities and the emission distributions, see Figure 5a,b. Across all scenarios the relative bias was below $\pm 10\%$ for 70.6% and 75.8% of the defining categories of the group-level transition probabilities and the emission distributions, respectively. As expected, an increased state distinctiveness and separation and increments in N_{dep}, N_{ind}, and N_{obs} resulted in better model performance in terms of point estimation of parameters. When bias was present, bias displayed a directional pattern, with high probabilities typically being underestimated, such as probabilities for self-transitioning and observing a defining category within a state, and low probabilities generally being overestimated, such as off-diagonal values of the TPM and probabilities for non-defining

[1]For a detailed description of estimation algorithm, see https://cran.r-project.org/web/packages/mHMMbayes/vignettes/estimation-mhmm.pdf

categories of the emission distribution. Additionally, estimating parameters close to the boundary of the probability domain (i.e., values near 0 and 1) was more challenging and thus more prone to bias. For example, the first state's transitions with probabilities $\boldsymbol{\gamma}_{1.} = (0.9, 0.05, 0.05)$ exhibited more extreme and biased behavior compared to the remaining two states (e.g., third state, $\boldsymbol{\gamma}_{3.} = (0.2, 0.3, 0.5)$).

The variance of the individual level random effects tended to be overestimated and was more challenging compared with accurate estimation of group-level parameters, as shown in Figure 5c,d. The accuracy of the estimation improved for larger values of the parameter, such as the individual random effects for transitions from the first state ($\boldsymbol{\sigma}^2_{[\alpha]1.} = 1.12$), compared to random effects for transitions from state two ($\boldsymbol{\sigma}^2_{[\alpha]2.} = 0.16$) and state three ($\boldsymbol{\sigma}^2_{[\alpha]3.} = 0.13$). Increasing N_{ind} improved the accuracy of estimates of individual level variance, and to a lesser extent with increased N_{dep} and N_{obs}, and increased state distinctiveness and separation.

The following sections first describe model convergence, followed by a detailed discussion on model performance as a function of the level of state distinctiveness and separation, the number of dependent variables N_{dep}, and the number of individuals N_{ind} and observations N_{obs}. Model performance will be discussed with respect to bias, the empirical standard error (precision), coverage, and state decoding accuracy. Due to the large number of conditions we only present a subset of them, although results presented are indicative of the general set of results.

Model convergence

Overall, parameter convergence was achieved for all group-level parameters in 52.8% of the main 324 scenarios assessed (convergence criterion-I). Meanwhile, 77.9% of the scenarios presented convergence measured as models with a mean MPSRF $\leq$ 1.2 averaged separately for transitions and emission parameters and fixed and random effects (convergence criterion-II). However, overall low convergence rates were to be expected, as we explored the minimum amount of data required to obtain reliable parameter estimates for the model's complexity level and data characteristics, pushing the boundaries. When a sufficient amount of information is present within the data ($N_{obs} \geq 800$ and $N_{dep} \geq 4$), more than 95% of the models met the convergence requirements on all their group-level parameters.

Hence, lack of convergence in at least one model parameter varied across the simulation design. Rates of model convergence generally improved with increasing N_{dep} and N_{obs} and a increased state distinctiveness and separation (Figure S1, Supplementary materials). N_{dep} appears to be the most relevant factor in ensuring the convergence of the model. Aggregated over the remaining factors, model convergence for all parameters in a model (criterion-I) within levels of N_{dep} ranged from 17.0% ($N_{dep} = 1$) to 89.6% ($N_{dep} = 4$), within levels of N_{obs}, from 40.2% ($N_{obs} = 100$) to 66.7% ($N_{obs} = 1600$), and within levels of N_{ind}, from 52.4% ($N_{ind} = 90$) to 60.2% ($N_{ind} = 5$). Adding non-informative (noise) additional dependent variables did not affect convergence of the models in a meaningful way. On the contrary, it appeared to improve the likelihood of convergence of the parameters, as 75.3% of scenarios achieved convergence in all their parameters (criterion-I), while 91.7% of them presented a mean MSPRF $\leq$ 1.2 averaged over parameters (criterion-II).

State distinctiveness and separation

Bias

In scenarios in which states are well-defined and well-separated, bias in the group-level model parameters is generally small over all other varied factors. That is, univariate data with a minimum sample size of $N_{ind} \geq 30$ and a number of observations $N_{obs} \geq 400$ resulted in little (absolute) bias for all group-level model parameters (see Figure S3 and Table S2 of the Supplementary materials). In scenarios with a low state distinctiveness and separation, accurate estimation of model parameters proved more difficult: more data by expanding either N_{dep}, N_{ind}, or N_{obs} is required to obtain acceptable bias as detailed below (see Figure 6, and Tables S4 and S5). Compared to the group-level model parameters, the effect of low state distinctiveness and separation on accurate estimation is less pronounced in the variance of the individual random effects (see Figure S4, and Tables S4 and S5). However, the beneficial effect of increasing the amount of data by expanding either N_{dep}, N_{ind}, or N_{obs} is larger for scenarios where states are well-defined and well-separated compared to low state distinctiveness and separation scenarios (see Figure S4). Hence, obtaining accurate parameter estimates for individual random effects is more difficult in the latter.

Empirical standard error

The level of distinctiveness and separation of the states has a large effect on the precision of the group-level transition and emission parameters (see Tables

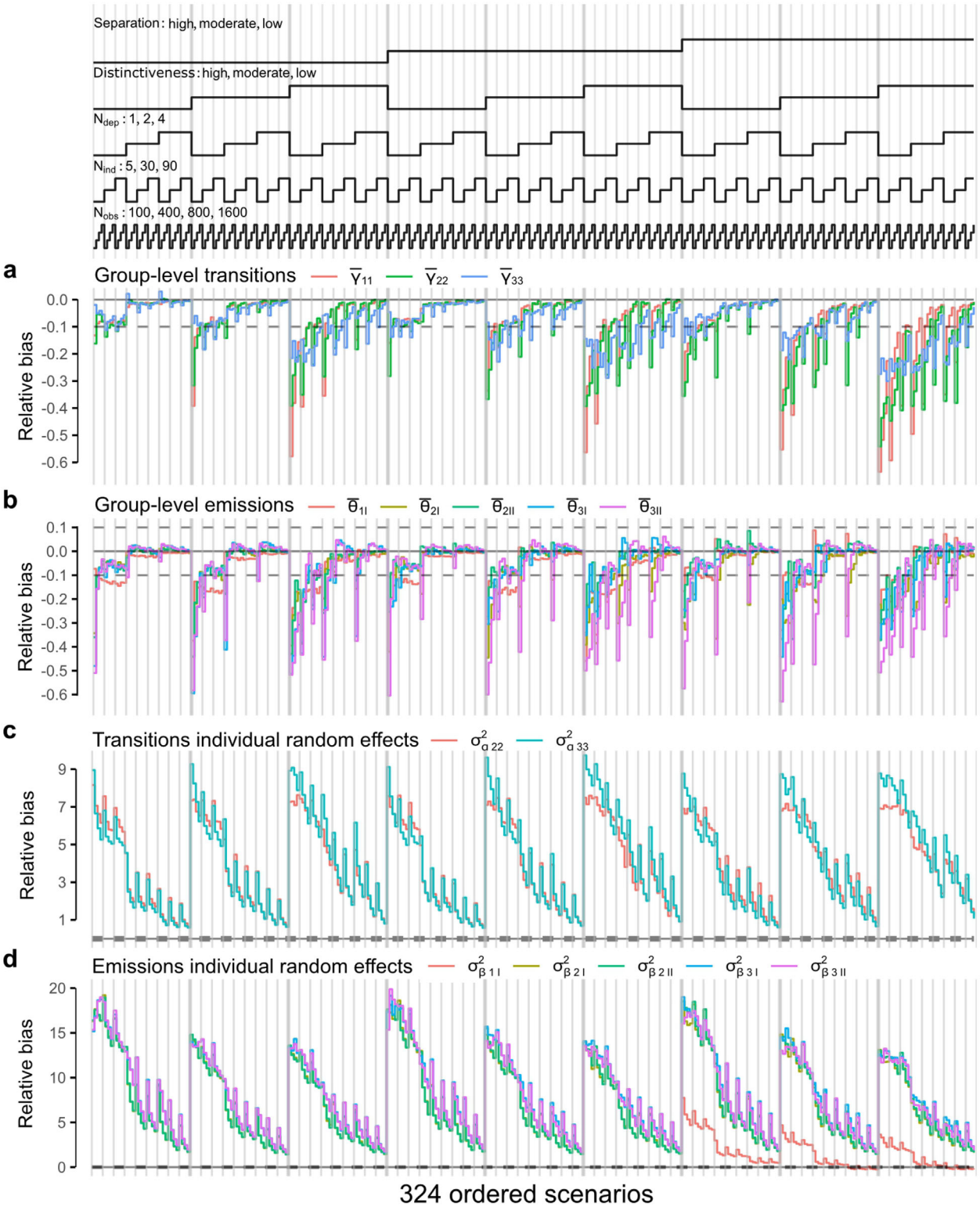

Figure 5. Nested loop plot. Mean relative bias in the point estimation of the group-level parameter estimates (panels a and b) and variance of the individual level random effects (panels c and d) over the 324 scenarios. Scenarios are ordered from outer to inner loops by scenarios values for the state *separation* ("high," "moderate," "low"), state *distinctiveness* ("high," "moderate," "low"), number of dependent variables $N_{dep} = (1, 2, 4)$, number of individuals $N_{ind} = (5, 30, 90)$, and number of observations per individual $N_{obs} = (100, 400, 800, 1600)$. To avoid overcrowding of the plot, displayed results are limited to the parameters corresponding to the diagonal of the transition distribution and the defining categories of the emission distribution. The horizontal dashed lines indicate a relative bias of $\pm 10\%$ on the estimation of parameters. Notice that because the defining categories of the emission distribution change between levels of *overlap*, a special notation (e.g., $\bar{\theta}_{1I}$ and $\bar{\theta}_{1II}$) was used. Note that the results for scenarios with $N_{dep} = 8$ were left our of this Figure, since they were only tested for the extreme conditions of high and low state distinctiveness and separation.

S2–S7). In data scenarios in which states are well-defined and well-separated, all empirical SE values of the transition parameters are below 0.07, which rises to a value of maximum 0.15 in scenarios with low state distinctiveness and separation. In addition, in data scenarios in which states are well-defined and well-separated, the precision of the group-level emission parameters is worse for the defining than for the non-defining categories. This pattern is not seen when state distinctiveness and separation is low. For the variance of the transition parameters' individual random effects, decrements in state distinctiveness and separation resulted in decreased precision. A direct comparison on precision between high and low levels of state distinctiveness and separation and the emission parameters' individual random effects is difficult, as the true parameter values differ over the scenarios.

Coverage

In scenarios in which states are well-defined and well-separated, average 95%CI coverage of the group-level model parameters of both transitions and emissions was generally acceptable (range: 80–100%; Tables S2 and S3). However, in scenarios with a low state distinctiveness and separation, average 95%CI coverage was lower (range: 22–99%; see Figure 7) and required more data in the form of more dependent variables or more observations as detailed below. Under-coverage was more severe for group-level parameters of the transition distribution and for parameters with probabilities close to 0 and 1 (e.g., transitions from the first state; see Tables S4, S5, and S7). As described above, individual random effects were generally overestimated by the model (see Figure 5c,d), leading to under-coverage in the majority of simulation scenarios with data of any level of distinctiveness and separation of the states (Tables S2–S7). Extreme under-coverage occurred for individual random effects with smaller true parameter values (e.g., $\sigma^2_{[\alpha]2j} = 0.16$ and $\sigma^2_{[\beta]2q} = 0.11$). However, for individual random effects with larger true parameter values (e.g., $\sigma^2_{[\alpha]2j} = 1.12$ and $\sigma^2_{[\beta]2q} = 1.01$), 95% coverage was generally better.

State decoding

Overall, the percentage of correct state assignment aggregated over all the main scenarios was 76.4% (range: 32.3–98.8%), with an accuracy of at least 80% obtained for 50.3% of the scenarios. The level of distinctiveness and separation of the states' distributions had an profound effect on state decoding, with the distinctiveness having a larger effect (see Figure 8 and Figure S7). In scenarios in which states are well-defined and well-separated, the average percentage of correct state assignment aggregated over all other factors was 87.4% (range: 54.0–97.8%), and an accuracy of at least 80% was obtained for 97.2% of the scenarios. In comparison, in data with low state distinctiveness and separation, the average percentage of correct state assignment aggregated over all other factors was 23.7% (range: 1.1–53.2%).

Interestingly, the state forward probabilities appear to capture the uncertainty on the state decoding (Figure 8), as on average 88.7% (range: 63.3–99.4%) of the times the correct state probability took a value of at least 0.2. The decoding accuracy varied over the states, and was better for states with a longer persistence (i.e., higher self-transitions). Aggregated over all simulation scenarios, the percentage of correct decoding for state 1 ($\bar{\gamma}_{11} = 0.9$) was 84.3% (range: 29.0–99.2%); for state 2 ($\bar{\gamma}_{22} = 0.7$), 66.5% (range: 20.4–98.4%); and state 3, with the lowest self-transitions ($\bar{\gamma}_{33} = 0.5$), 58.1% (range: 13.2–97.6%).

Number of dependent variables

Bias

In scenarios with a low level of state distinctiveness and separation, including multivariate data substantially reduces bias in the point estimates of the group-level transitions' and emissions' parameters, especially when N_{ind} and N_{obs} are low (see Figure 6). Moving from $N_{dep} = 4$ to $N_{dep} = 8$ does not appear to have a substantial impact on the parameter bias of the group-level parameters (see also Tables S3–S7). In the variance of the individual random effects, expanding the data beyond univariate observations only had a negligible effect on model performance over all used levels of state distinctiveness and separation (Tables S2 and S3 and Figure S4).

In comparison to the baseline scenarios, the models with additional random noise dependent variables did not present substantial differences with respect to the bias of the group-level parameters. That is, a slight increase was observed in the bias of the group-level transitions (Δ mean abs. bias $= 0.01$ over parameters) and emissions (Δ mean abs. bias $= 0.02$ over parameters) for data of both high and low levels of state distinctiveness and separation (see Figure S3). The increase in bias was the largest for the scenarios with the smaller sample size ($N_{dep} = 1$ and $N_{obs} = 100$) in data of both high and low levels of state distinctiveness and separation. The opposite happened for the estimation of individual random effects, with a moderate decrease in bias of (Δ mean abs. bias $= -0.14$)

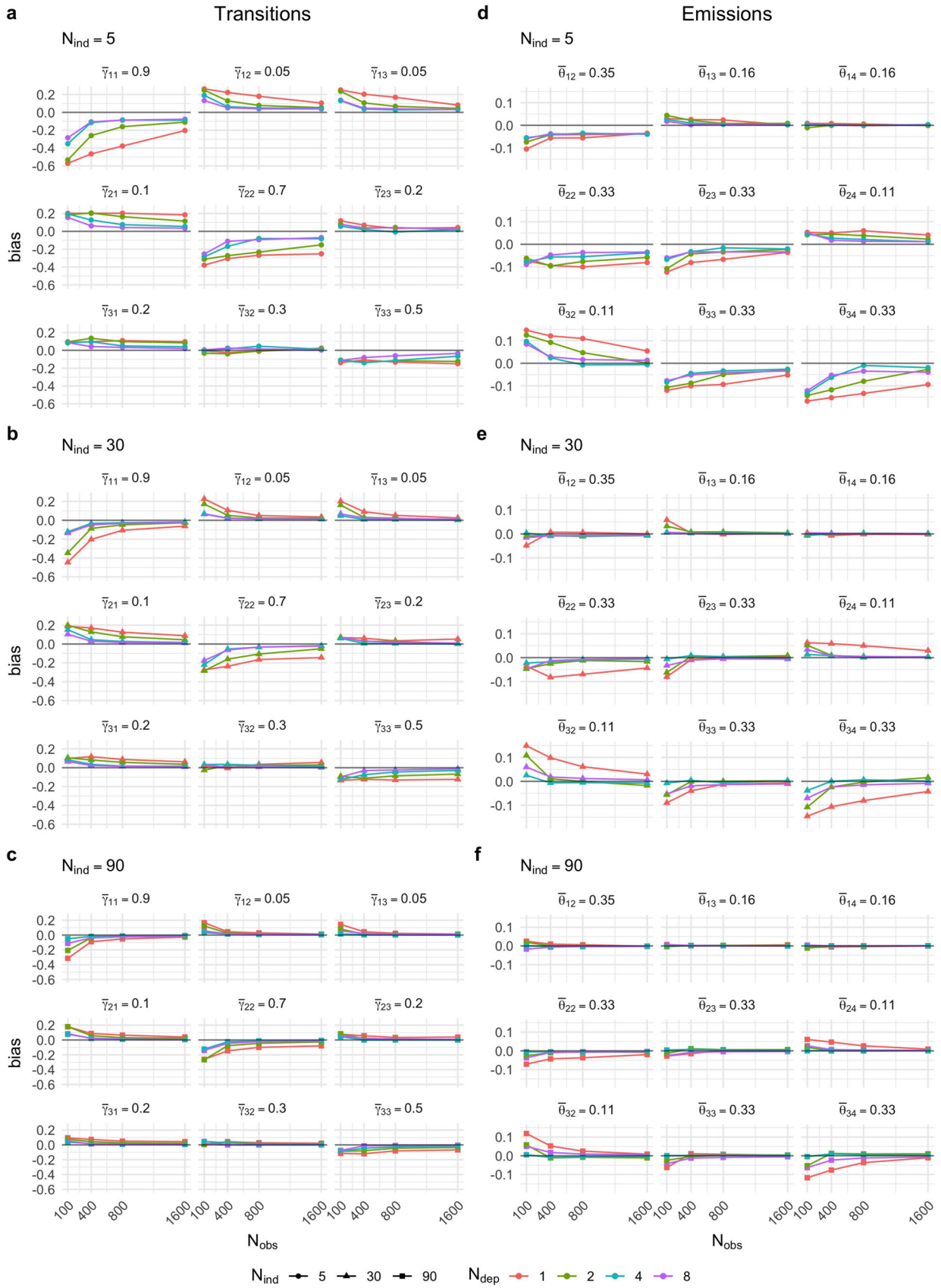

Figure 6. Trellis plot. Mean bias on the estimation of group-level parameters of the transition distribution (panels a–c) and the emission distribution (panels d–f) for the subset of scenarios with low state *distinctiveness* and *separation* over levels of number of dependent variable N_{dep}, number of individuals N_{ind} and number of observations per individual N_{obs}. Line color indicates the value for the number of dependent variables N_{dep} of the scenario; item shape indicates the value for the number of individuals N_{ind} of the scenario. Parameters are displayed on the probability domain to aid interpretation. For the emission distribution, only parameters $\bar{\theta}_{i2}$, $\bar{\theta}_{i3}$, and $\bar{\theta}_{i4}$ are shown, as parameters $\bar{\theta}_{i1}$ and $\bar{\theta}_{i5}$ relate to noise categories within the scenarios concerning low state distinctiveness and separation.

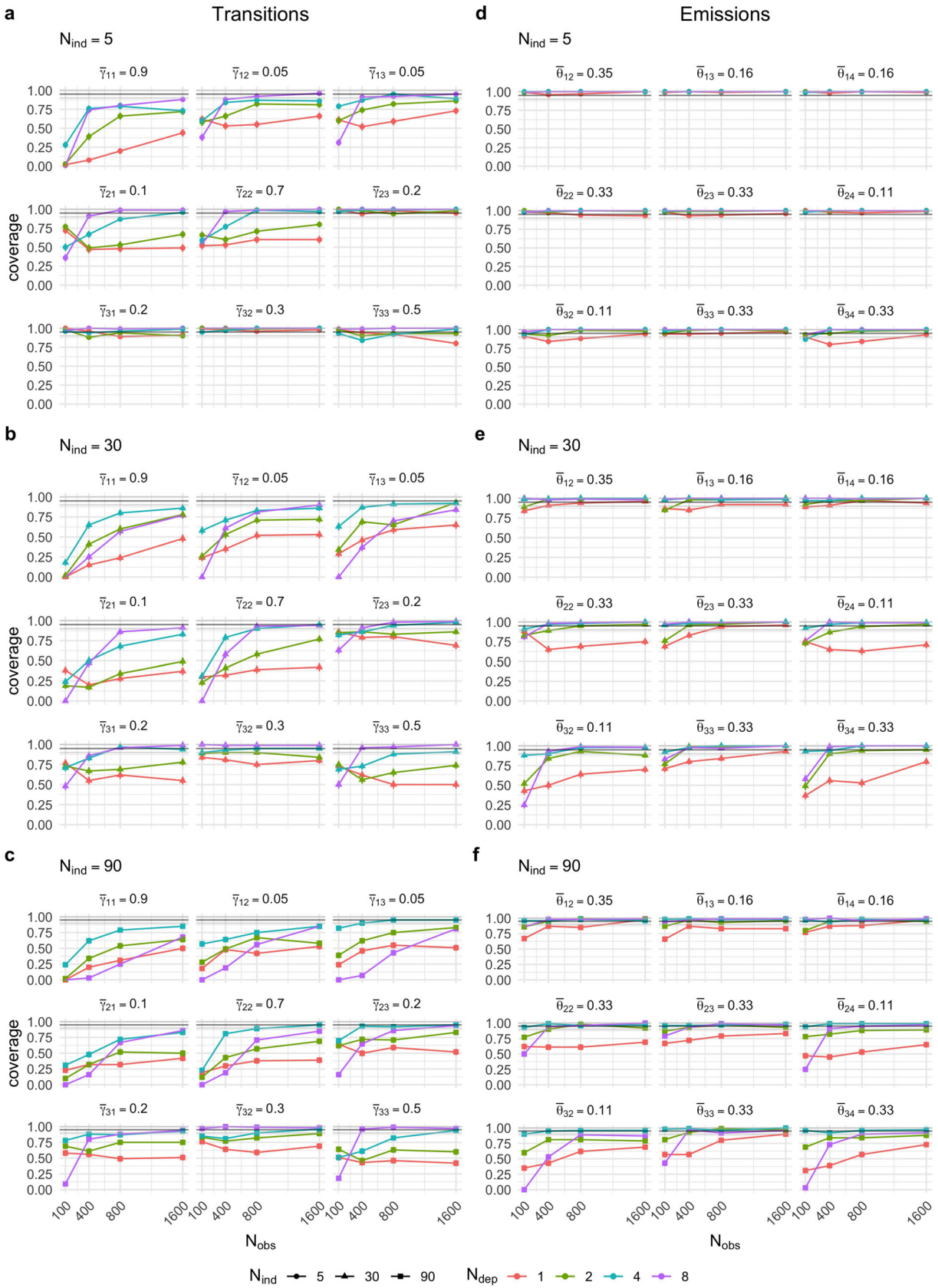

Figure 7. Trellis plot. Mean coverage for group-level parameters of the transition distribution (panels a–c) and the emission distribution (panels d–f) for the subset of scenarios with a low level of state *distinctiveness* and *separation* in the emission distributions, over levels of number of dependent variable N_{dep}, number of individuals N_{ind} and number of observations per individual N_{obs}. Line color indicates the value for N_{dep} of the scenario; item shape indicates the value for the number of individuals N_{ind} of the scenario. Coverage corresponding to parameters on the probability domain to aid interpretation. For the emission distribution, only parameters $\bar{\theta}_{i2}$, $\bar{\theta}_{i3}$, and $\bar{\theta}_{i4}$ are shown, as parameters $\bar{\theta}_{i1}$ and $\bar{\theta}_{i5}$ relate to noise categories within the scenario concerning low state distinctiveness and separation.

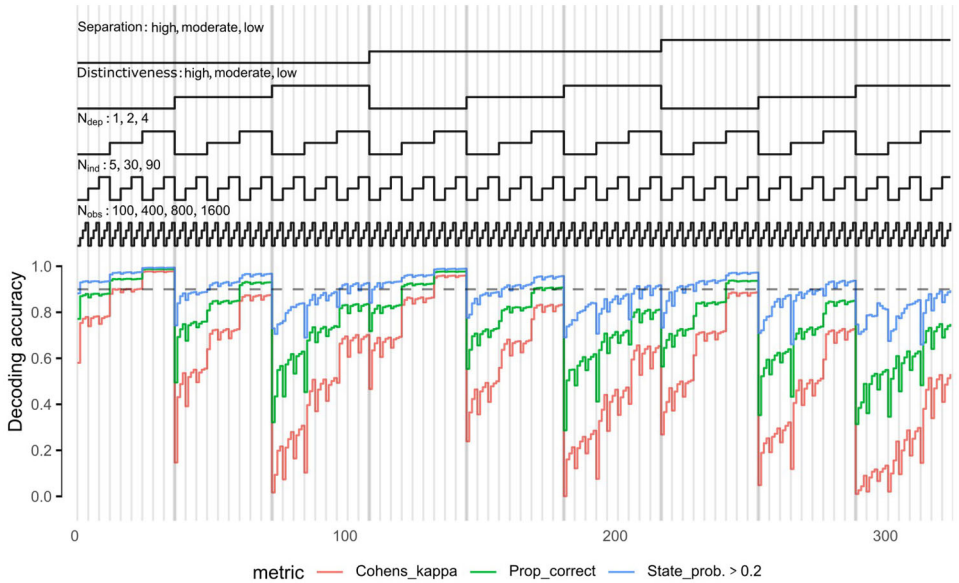

Figure 8. Nested loop plot. Three metrics of accuracy of global state decoding with Viterbi algorithm averaged over individuals and repetitions of a scenario, shown over the 324 scenarios. Line color indicates: mean Cohen's κ (in red), mean proportion of states correctly assigned (in green), and mean proportion of true states with a state probability ≥ 0.2 (in blue). Scenarios are ordered from outer to inner loops by scenarios values for the state *separation* ("high," "moderate," "low"), state *distinctiveness* ("high," "moderate," "low"), number of dependent variables $N_{dep} = (1, 2, 4)$, number of individuals $N_{ind} = (5, 30, 90)$, and number of observation per individual $N_{obs} = (100, 400, 800, 1600)$. The horizontal dashed line indicates an accuracy of 90%. Note that the results for scenarios with $N_{dep} = 8$ were left our of this Figure, since they were only tested for the extreme conditions of a low and high level of state distinctiveness and separation.

for the transitions and (Δ mean abs. bias $= -0.25$) for the emissions of models with additional noise variables compared to the baseline (see Figure S4).

Empirical standard error

Using multivariate data has a significant effect on the precision of the estimation of group-level parameters, especially in case of data with low state distinctiveness and separation (Tables S4 and S5). However, increasing N_{dep} beyond 4 to $N_{dep} = 8$ only slightly further reduces the empirical SE of the group-level parameters (see Tables S5 and S7). Regarding the empirical SE values for the variance of the individual level random effects, N_{dep} effects the precision of the estimation to a lesser extent. In comparison to the baseline scenarios, the models with additional random noise dependent variables did not present substantial differences with respect to the empirical SE of the group-level parameters. For the individual random effects, a more substantial decrease in the empirical SE of (Δ mean abs. bias $= -0.13$) was observed for the transitions and (Δ mean abs. bias $= -0.14$) for the emissions of models with additional noise variables.

Coverage

Increments on N_{dep} were most effective at increasing the coverage of group-level parameters of all varied design factors. In scenarios in which states are well-defined and well-separated, average 95%CI coverage of the group-level model parameters of both transitions and emissions increased from a range of 80–98% with univariate data to 92–100% with four dependent variables (Tables S2 and S3). In scenarios with a low level of state distinctiveness and separation, average 95%CI coverage of the group-level model parameters increased from a range of 22–98% with univariate data to 63–99% with four dependent variables; see Figure 7. For the individual random effects, 95%CI coverage tended to improve with increments in the number of dependent variables as well (Tables S2 and S6).

The models with $N_{dep} = 8$ generally presented lower coverage on both the group-level transition and emission parameters compared to the models with $N_{dep} = 4$, for both the group-level effects as well as for the individual random effects (see Tables S6 and S7). This possibly occurred as a result of the reduced empirical SE of the model parameters which leads to narrower, more restrictive confidence intervals. A similar pattern of decline in the coverage was seen for the scenarios with additional random noise dependent variables, with respect to the baseline (see Figures S6 and S7).

State decoding

State decoding accuracy improved with increasing N_{dep} (see Figure 8 and Figure S7). In scenarios in

which states are well-defined and well-separated, having $N_{dep} = 1, 2, 4,$ or 8 increased accuracy from 86.3% to 94.3% to 98.7% to 99.8%. In scenarios with a with a low level of state distinctiveness and separation, accuracy increased from 46.3% to 55.2% to 69.4% and slightly decreased to 67.5% with $N_{dep} = 1, 2, 4,$ and 8, respectively.

The scenarios with additional non-informative (noise) dependent variables did not fare worse in terms of decoding accuracy, with the single exception of the scenarios with $N_{obs} = 100$ and single informative dependent variable (Figure S5). In scenarios with a low level of state distinctiveness and separation, the mean percent Cohen′ sκ averaged over all the other factors improved slightly with added noise variables from 23.7% to 32.8% (range: 0.4–68.9%).

Number of individuals and number of observations

Bias

In scenarios containing a low distinctiveness and separation between states, increments in N_{ind} beyond 30 only appeared to reduce bias slightly in the group-level parameters. Increments on the N_{obs} result in better model performance over all specified levels (albeit improvement was limited for the higher end of N_{obs}; see Figure 6). However, in some of the transition parameters, no combination of number of individuals or number of observations produced acceptable bias in case of univariate data, while in multivariate data ($N_{dep} \geq 2$) an acceptable level of bias in all transitions (and emissions) was observed with $N_{ind} \geq 30$ in combination with $N_{obs} \geq 400$. For the emissions, $N_{ind} \geq 30$ with $N_{obs} \geq 1600$ appeared to be sufficient for univariate data.

For accurate estimation of the variance of the individual random effects, model performance was most affected by the number of individuals N_{ind} compared with the number of observations N_{obs}. That is, model performance improved with increments in N_{ind} for both the transition probability parameters $\sigma^2_{[\alpha]ij}$ and emission distribution parameters $\sigma^2_{[\beta]iq}$. For data with well-defined and well-separated states, increments over a minimum of 400 observations had a negligible effect on the estimation performance (Tables S2 and S3). For data with a low state distinctiveness and separation, the positive effect of increments in all levels N_{obs} translated to improvements on the estimation performance (Tables S4 and S5), albeit less pronounced compared to the effect of the number of individuals N_{ind}.

Empirical standard error

For the precision of the group-level estimations, N_{ind} was a more influential factor when the states were well-defined and well-separated (Tables S2 and S3). Here, including $N_{ind} \geq 30$ and $N_{obs} \geq 400$ resulted in empirical SE values to dive below 0.04 and 0.02 for the transition and emission parameters, respectively. In scenarios with a low distinctiveness and low separation between states, the empirical SE declined with increments in N_{ind} and N_{obs} to values below 0.05.

Regarding the empirical SE for the variance of the individual level random effects, results supported the notion that N_{ind} is the primary driver of accurate and reliable estimation of random effects in the model. In contrast, the effect of N_{obs} appeared to be negligible. Empirical SE values declined noticeably with an increment in the number of individuals from $N_{ind} = 5$ to $N_{ind} = 30$ for both transition and emission probability parameters for both low and high levels of distinctiveness and separation between states (Tables S2–S7). An additional increment in N_{ind} only showed a negligible effect on the empirical SE.

Coverage

After the number of dependent variables N_{dep}, increments in N_{obs} were most effective at increasing the coverage of group-level parameters. Coverage declined with increments in N_{ind}, possibly due to narrower arms on the 95% confidence intervals. Using at least 800 observations per individual and bivariate data ensured an average coverage of 74% (range: 51–91%) for transition group-level parameters. Meanwhile, even 400 observations per individual and univariate data appeared to be sufficient to ensure a coverage of 86% (range: 73–96%) on emission group-level parameters.

State decoding

The decreased correct state assignment and accuracy obtained in data with a low level of state distinctiveness and separation could partially be alleviated by increments in N_{obs} (see Figure 8 and Figure S7). Meanwhile, increments in N_{ind} only had a marginal effect on state decoding accuracy, except for settings in which $N_{obs} = 100$.

Empirical application

The empirical example is part of a study by Hale (n.d.) on nonverbal communication between adolescents diagnosed with a mood disorder and their therapists. The data consists of five categorical variables

observed for both the patient and the therapist, for $N_{ind} = 39$ patient-therapist dyads with $N_{obs} = 900$ (that is, behavior was recorded for 15 min at a frequency of 1 observation per second for each couple). We focus on two of these variables, namely: vocalizing (three categories: speaking, silent, backchanneling) and looking (binary: looking and not looking at the other). To keep our illustration simple, no covariates to explain possible heterogeneity between dyads were included in the model.

We analyzed the data by fitting MHMMs with two to six hidden states using the R package *mHMMbayes* (Aarts, 2019; for R code, see Mildiner Moraga & Aarts, 2022). For each model we fit four chains of 4000 iterations with a burn-in period of 2000, and different starting values based on the maximal likelihood estimates of the corresponding parameters on a single-level HMM fit with the R package *depmixS4* (Visser et al., 2009). Only models with two and three hidden states passed the convergence criterion (MPSRF ≤ 1.05 on all the group-level model parameters; Brooks & Gelman, 1998). We selected the model with three hidden states based on its lower AIC ($AIC_{M=2} = 3333.7$ and $AIC_{M=3} = 3315.6$).

In our model, states 1 and 3 describe situations where either the therapist or the patient speak while the other is looking at the speaker, respectively (see Figure 9). In state 2, both therapist and patient have a moderate probability of speaking, whether with overlapping vocalizations or not. In addition, we also observe a significant proportion of back channeling by the patient in state two. Visual inspection of the individual specific posterior densities of vocalizing (Figure 10) reveals that the composition of the mixed state (state two) varies considerably across dyads. On the contrary, patient and therapist vocalizing within the state that that either of them dominates (states one and three) are quite uniform across dyads.

Group-level point estimates for the probabilities of transitioning between the states are presented in Table 2, along with the range of dyadic-specific TPM over the 39 dyads. The self-transitions took higher values than the off-diagonal transitions, signaling that behavioral states tend to persist in time once they start. States one and three, in which a single member of the dyad has a high probability of taking the word tend to be more persistent in time. The mean expected duration[2] of these two states was 14.3 and 14.5 s, respectively, while state two is expected to last only 3.6 s. In addition, transitions from the second state presented a larger variability between dyads than transitions from the other two states (see ranges in Table 2). As such, state two not only has a more heterogeneous composition (see Figure 10), but also its dynamics and persistence are relatively heterogeneous between dyads. On the contrary, dynamics, persistence and composition of the other two states appear to be relatively homogeneous between dyads.

Posterior predictive checks (PPCs) generally indicate an adequate fit of the three-state MHMM to the patient-therapist data. In the PPCs, $N_{sim} = 500$ new data sets were simulated using the parameter estimates obtained in the fitted three-state MHMM, and the extent to which a set of summary statistics over the simulated data sets recapture the values in the empirical data was assessed. For a more detailed explanation, we refer the reader to Lynch (2007), Gelman et al. (2013), and Gelman and Carlin (2014). The first posterior predictive check (PPC1) aimed to assess whether the model could adequately reproduce the mean proportions of observation categories (over dyads) for the four dependent variables. Figure 11 shows the mean proportion of each category of observation on the empirical data (bars), and the simulated data sets (medians and 95% credibility intervals). Overall, Figure 11 and the corresponding posterior predictive values (PPV) show that the MHMM is able to recover the general trend of observed proportions of the empirical set, which is a indicator of adequate fit to the data (i.e., $.025 \geq ppv[rep > true] \leq .975$ for all but two of the observation categories). However, the model overestimates the proportion of events of "backchanneling" for both patient and therapist (producing extreme posterior predictive values of $ppv[rep > true] = .976$ and $ppv[rep > true] = .994$ for the patient and the therapist, respectively).

The second PPC (PPC2) aimed to explore whether the model was able to reproduce the proportion of events in which one, both, or neither therapist and patient vocalized in the same time. We labeled the data according to who displayed a vocalizing behavior at each t on each dyad. As labels, we used the categories "both," "patient," "therapist," and "none" when both were silent. The visual representation of PPC2 on Figure 12 indicates that the model adequately reproduces the pattern of sources of vocalizations in the data (i.e., $.025 \geq ppv[rep > true] \leq .975$), although the proportion of observation in which neither is speaking tended to be underestimated ($ppv[rep > true] = .046$). The later is likely a

[2]The MHMM implicitly assumes a geometric distribution for the state duration. To ease interpretation, we can approximate the geometric distribution with the exponential distribution using $\lambda = \bar{\gamma}_{ii}$, and obtain the expected duration of a state by $1/(1 - \bar{\gamma}_{ii})$.

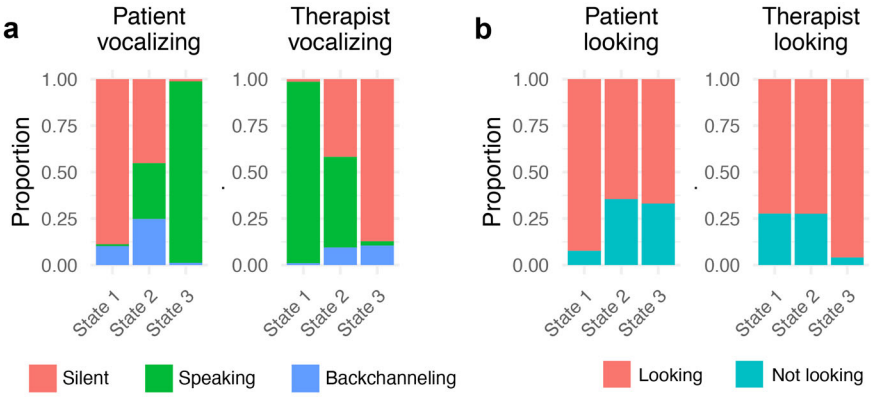

Figure 9. Mosaic plot. Composition of hidden states: maximum a posteriori estimations of the group-level emission probabilities over the states for the four dependent variables. Panel (a) displays the emission probabilities for "patient vocalizing" and "therapist vocalizing"; panel (b), for "patient looking" and "therapist looking." Color fill indicates the observation category of the corresponding dependent variables. State 1 and state 3 are dominated by either patient or therapist (correspondingly), while for state 2 both have a moderate probability of speaking (mixed state).

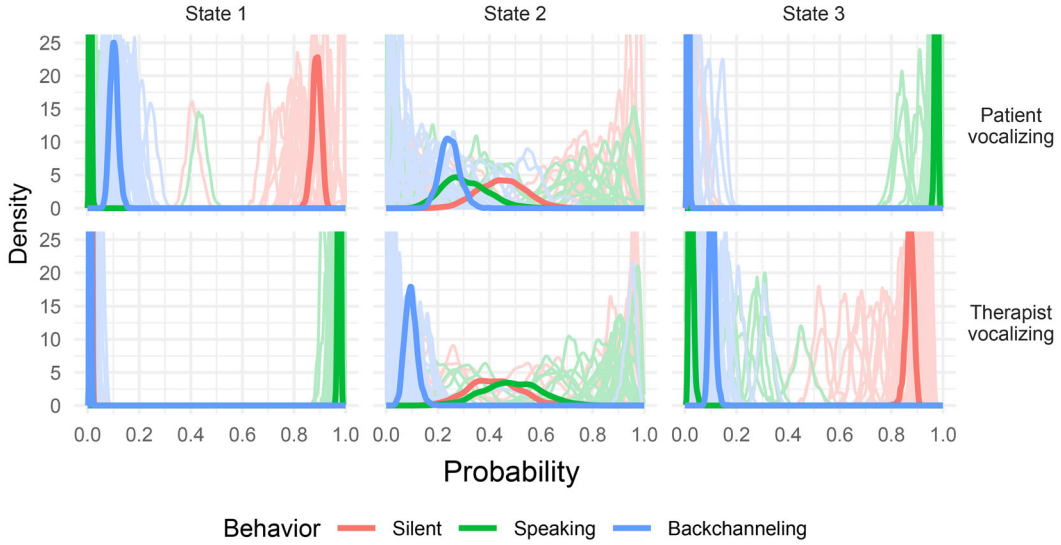

Figure 10. Density plot. Posterior distributions of the emission probabilities of "patient vocalizing" and "therapist vocalizing" over the three hidden states. Thick lines indicate group-level emission probabilities, and thin lines, individual-specific emission probabilities. Color fill indicates the observation category of the corresponding dependent variables. The composition of state 2 (mixed state) is more variable across dyads than the composition of the state 1 and 3.

consequence of underestimating the proportion of times the patient is silent during a session (see PPC1 in Figure 11, results for the variable "patient vocalizing").

This empirical application provides an example of how the dynamics over time in behavioral data are summarized using the MHMM. Multivariate data with four dependent variables is collapsed in a data-driven manner into three non-verbal communication states over time, were the composition of the obtained non-verbal communication states have a clear behavioral interpretation. Dynamics over time are captured in the transition probabilities, which also shed light on the time persistence of each of the non-verbal communication states. Variability between dyads is captured in the model, and can be visualized intuitively. In addition, the PPCs show the versatility of the Bayesian method to assess the goodness of fit of the model.

Discussion

In the current study, we have investigated factors that influence parameter estimation performance of the MHMM for categorical data. Investigating parameter estimation performance in the MHMM is crucial to ensure reliable and replicable results in applying the MHMM and a first step toward defining a set of data-

Table 2. Maximum a posteriori estimates of the transition group-level parameters on the patient-therapist data (range of the dyad-specific transition probabilities indicated in parentheses).

		Transition to		
		State 1	State 2	State 3
From	State 1	0.93 (0.86–0.97)	0.05 (0.01–0.12)	0.02 (0.01–0.09)
	State 2	0.16 (0.05–0.28)	0.73 (0.52–0.92)	0.11 (0.03–0.27)
	State 3	0.03 (0.01–0.15)	0.04 (0.01–0.10)	0.93 (0.81–0.96)

requirement guidelines. Simulation scenarios included varying levels of number of dependent variables, sample sizes, and level of state distinctiveness and separation in the emission distributions. The fitted MHMM included individual random effects that allow for the quantification of heterogeneity between individuals on both the parameters of the transition probabilities and the parameters of the emission distribution. Our results provide insights about the data requirements of the model and show that it is feasible to obtain accurate group-level estimates with the MHMM even on conditions of low state distinctiveness and separation.

In the following, we summarize our findings, give guidelines for social and behavioral researchers who would like to apply the MHMMs, and discuss how further research could shed more light on the potential of using MHMMs.

Main findings

Using multivariate data

In general, including multivariate data ($N_{dep} \geq 2$) appears to benefit the convergence of the model parameters, the estimation of group-level parameters and individual random effects in both transition and emission distribution, and the accuracy of hidden state decoding. The beneficial effect of multivariate data is especially noticeable when dealing with states whose distribution is not so well separated. The model can use any complementary information encoded in the variables to map different aspects of the latent process, which leads to a better identification of hidden states, generally improving model performance. This also explains why, for data containing low levels of state distinctiveness and separation in the emission distributions, using multiple dependent variables produced a better model performance than an equivalent increase in the number of observations per individual for a single variable (i.e., models with bivariate data and 400 observations per individual produced better estimations than models with 800 univariate observations per individual).

These results go in line with previous research on the number of indicators (here, dependent variables) on latent class analysis (Wurpts & Geiser, 2014), structural equation modeling with continuous latent variables (Marsh et al., 1998), and latent transition analysis (Collins & Wugalter, 1992), showing that the benefits over the estimation performance of the model of including additional indicators appear to overweight any detrimental effects due to increments in the complexity of the model. This is more so the case for the HMM, because of the assumption of conditional independence between dependent variables (indicators). As a result, no variance-covariance matrix is modeled between them, which avoids sparseness issues on response patterns. Also in line with these studies, our results indicate that the beneficial effect of multivariate data is more pronounced for data sets with smaller sample sizes, where only limited information is available for the identification of the hidden states.

Thus, multivariate data can alleviate the negative effect of low separation between states, a relevant finding considering that sometimes the N_{obs} is constrained by the type of empirical application (i.e., behavioral processes such as therapy sessions often have a maximal duration). Notwithstanding, when the states are well separated, and a sufficient N_{obs} is given, using multivariate data does not have a significant effect over the group-level parameters. Only when N_{obs} is insufficient –for instance, $N_{obs} = 100$ in our simulation– using multivariate data appears to improve the estimation of group-level parameters through compensation.

Interestingly, the scenarios with additional random noise dependent variables revealed that the model is relatively robust to the inclusion of variables that do not contain information for mapping to the hidden states. Overall, the inclusion of these variables only had a moderate detrimental effect on parameter coverage, and which also extended to bias in scenarios with smaller sample size ($N_{dep} = 1$ and $N_{obs} = 100$). Moreover, the inclusion of the additional random noise variables had no substantial impact on the accuracy of state decoding. Even though it is reassuring that the model can generally handle this type of miss-specification, in practice these variables would rarely be included in the model, as one can use theory and model selection techniques to choose which variables to include or leave out. Although our results did not reveal a tipping point after which including additional dependent variables has a marked detrimental effect over the stability of the results, it is possible

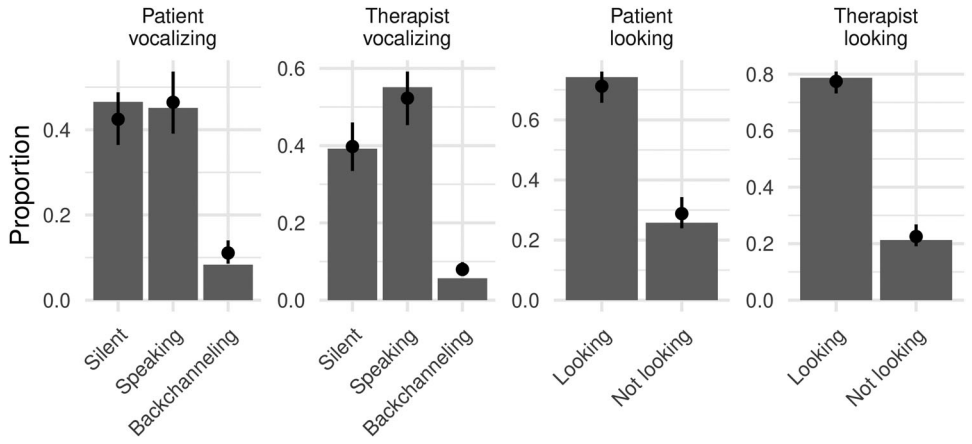

Figure 11. Bar plot. Results of posterior predictive check 1 (PPC1) for the group-level emission parameters. The bars represent the mean proportion of each category of observation (over dyads) on the empirical data. The point and range, the median and 95% credibility interval (95%CI) of the mean proportions (over dyads) aggregated over the $N_{rep} = 500$ simulated data sets (model predictions). Overlapping of the 95%CI arms with the empirical means indicate good fit to the data. The model is generally able to capture the group-level emission parameters, although it overestimates the incidence of "backchanneling" in the data.

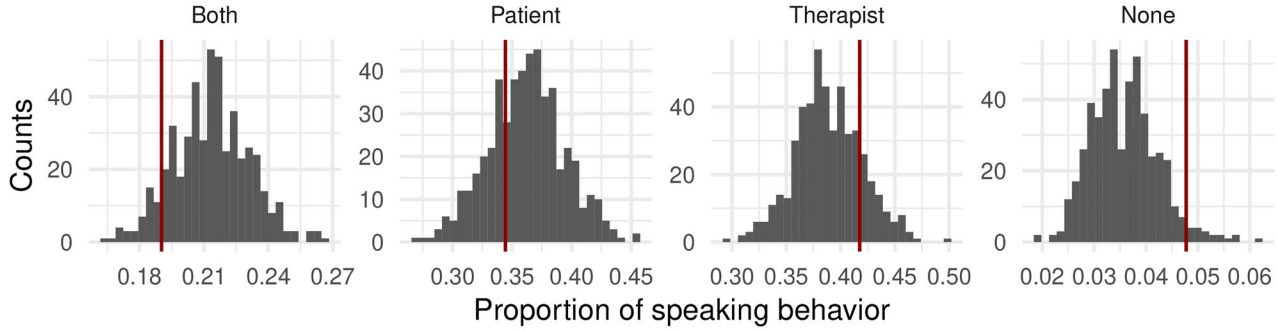

Figure 12. Histogram. Who is talking? Results of posterior predictive check 2 (PPC2) on the group-level emission probabilities of the dyads. The red lines represent the empirical mean of the proportion of observations over sessions in which "patient," "therapist," "both," or "none" vocalized. The histograms represent the model predictions for the mean proportions over the $N_{rep} = 500$ simulated data sets. Overlapping of the histogram with the vertical line indicates good fit. The model is generally able to capture the frequency of each category of observation, slightly overestimating the incidence of instances of "both" speaking, and underestimating the incidence of "none" speaking.

that under certain conditions the estimation cost of additional variables outweighs their benefits (e.g., dependent variables not related to the latent process under study, variables with sources of variability unaccounted for in the model, or that present high levels of multicollinearity to the point of harming parameter estimation).

Number of individuals and number of observations

In line with previous simulations in multilevel latent variable models, both the number of individuals and the number of observations per individual play an important part in the estimation of group-level parameters on the MHMM, albeit their effect and relevance are not the same (e.g., Finch & French, 2014; McClintock, 2021; Schultzberg & Muthén, 2018). In particular, having a small number of observations per individual ($N_{obs} = 100$) appears to be more detrimental than a small number of individuals ($N_{ind} = 5$). Also in line with these studiers, our results show that a large number of observations per individual or number of individuals can partly compensate for a small number on the other (e.g., Schultzberg & Muthén, 2018). A large number of observations per individual improves the precision of the individual-level parameters, which results in a better estimation of group-level parameters. A large number of individuals implies that even if the individual-level estimations were not so accurate, the group-level average would approach the population value. However, for accurate estimation of the group-level parameters, both the number of observations per individual and the number of individuals need to be of a sufficient amount ($N_{obs} > 100$ and $N_{ind} > 5$). Regarding the estimation of individual random effects, it is clear that the number of individuals is the most important

factor, and cannot be compensated by the number of observations per individual. This result goes in line with previous research on multilevel models: the number of individuals is usually considered the main driver on the estimation of random effects in multilevel models (e.g., Finch & French, 2014; McNeish, 2019; Schultzberg & Muthén, 2018). For decoding of the sequence of hidden states, our results show that ensuring a sufficient number of observations per individual appears to be more relevant than increasing the number of individuals in the data. When data for more individuals is available, the group-level parameters can be estimated more accurately.

Hence, whether a large number of observations per individual or a large number of individuals is to be preferred depends both on the costs related to sampling on either level, and whether one is mainly interested in the overall, group-level temporal dynamics, in the between-individual variation (measured by the individual random effects), or in accurately uncovering each individuals states sequences.

General findings in parameter estimation

Our results revealed a few additional aspects about of the parameter estimation in the MHMM.

Generally, estimating parameters closer to the boundary of the parameter space (i.e., probabilities close to zero and one) is more challenging than estimating parameters that take intermediate values (Beyer et al., 2013; Jonsen, 2016; McClintock, 2021). Our results confirmed this notion: the state with the lowest self-transition in these results led to the most accurate results in both self-transitions and off-diagonal transitions. This result indicates that data sets with a transition distribution that is not so extreme require a smaller amount of data to deliver accurate estimation. Notwithstanding, higher self-transitions (longer persistence) also lead to more accurate decoding of the sequence hidden states. Because states with a longer persistence occur more frequently in the data, the model counts with more information to correctly identify them.

In addition, estimation of group-level parameters of the transition probability matrix was downwardly biased for the self-transitions and upwardly biased for the off-diagonal transitions. That is, the MHMM generally underestimated parameters in the probability scale with a value close to one, and it overestimated low-probability parameters whose values lye close to zero. The bias size was comparable for the two types of transitions departing from the same hidden state, but naturally resulted in a larger relative bias in the estimation of off-diagonal transitions. Off-diagonal transitions were likely harder to estimate because of their relative infrequent occurrence in the data.

Furthermore, the estimation of group-level parameters was more accurate compared to estimation of random effects. This result goes in line with the previous evidence for multilevel models and multilevel time-series (see e.g., Asparouhov et al., 2018; Hox et al., 2018; Landau & Stahl, 2013; McClintock, 2021). In particular, the model tended to overestimate the transition probabilities' random effects across scenarios. The overestimation is likely a consequence of using a scaled Inverse-Wishart distribution to sample the variance-covariance matrices of random effects, which has a low density around zero (Gelman, 2006; Lemoine, 2019). The estimation of variances can perform poorly when variances are small relative to means, constraining variances upwards and correlations downwards (Alvarez et al., 2014). As a result, sampling from the Inverse-Wishart can lead to slight to moderate overestimation of the posterior variance-covariance matrix (Alvarez et al., 2014; Lemoine, 2019). As such, researchers interested in the between-individual variability can expect to require a larger N_{ind} than those mainly interested in group-level parameters. However, the number of individuals needed can partly be alleviated by the use multivariate data.

Using MHMM for modeling intensive longitudinal data: guidelines for applied researchers

We base the following guidelines on the variety of scenarios that applied researchers may encounter depending on their interests and variations in the outcome variable (that is, in the levels of state distinctiveness and separation in the emission distributions). We remind the readers that the guidelines are based on our simulation, which assumed data with relatively high self-transitions, and in which all dependent variables contribute partly unique information. For data sets with lower self-transitions, the model would likely require less data to produce accurate estimates.

When the focus is on the group-level parameters

Research focusing on the group-level parameters of any of the two components of the model should prioritize using multivariate data and ensuring a sufficient number of observations per individual. As such, expanding the number of individuals in the data is of a lesser importance. Here, it is important to note that the temporal design (e.g., measurement timing,

frequency, and spacing) should be based on the timeframe at which the underlying process under study is assumed to change, rather than the required number of observations per individual, as a miss-match between the sampling frequency and the process under study can lead to misleading results (Ariens et al., 2020; Collins, 2006; Hamaker & Wichers, 2017). Researchers should also note that, as mentioned before, the relative estimation performance is better for self-transitions compared to off-diagonal transitions. Upward bias can be expected on the off-diagonal transitions in data sets with very high self-transitions, which is characteristic in data sets with high measurement frequency.

The specific guidelines on the N_{ind}, N_{obs}, and N_{dep}, depend on the amount of state distinctiveness and separation in the emission distributions. For instance, in a data set where the states are distinctive and well separated, estimating the transition probabilities reliably would minimally require: (a) 400 observations for two dependent variables on five individuals, (b) 800 observations for one dependent variable on five individuals, or (c) 400 observations for one dependent variable in 30 individuals (see Figure 5). However, obtaining a given level of accuracy with a simpler model (e.g., two hidden states) or less complex data (e.g., states well separated, low between-individual variability, continuous data) will likely require less data.

In contrast, for a data set on the lower end of level of state distinctiveness and separation on the emission distributions, the acceptable estimation accuracy of the group-level parameters would require at least: (a) 800 observations for four dependent variables on five individuals, (b) 1600 observations for two dependent variables on five individuals, or (c) 1600 observations for one dependent variable in 30 individuals (see Figure 5).

When the focus is on the individual random effects

Researchers focusing on the random effects should prioritize increasing the number of individuals N_{ind} over the number of observations per individual N_{obs}, as the former is the main factor driving the estimation of these effects. In addition, researchers are encouraged to include multivariate data whenever possible. As previously discussed, we note that researchers can expect an overestimation of the individual random effects. In our setting, almost none of the combinations of researcher-controlled factors were sufficient to produce acceptable levels of bias for these parameters. However, we notice that a $N_{ind} = 30$ combined with $N_{dep} = 2$ offers a significant improvement on both bias and precision for data over all levels of state distinctiveness and separation in the emission distributions. The estimation improves even further with $N_{ind} = 90$ and $N_{dep} = 4$, but with a decreasing marginal gain. Based on our results, they can also expect estimation of individual random effects to become more reliable in data sets with a larger between-individual variability.

When the focus is on the state decoding

Researchers especially interested on the decoding of the sequence of hidden states on the data are encouraged to prioritize multivariate data, and to a lesser extent, increasing the number of observations per individual. For data in which states are well-defined and well-separated, a decoding accuracy $\geq 90\%$ was achieved with multivariate data ($N_{dep} \geq 2$) and 100 observations per individual in our results. For data with a low level of state distinctiveness and separation, four dependent variables and 400 observations per individual were sufficient to achieve a decoding accuracy $\geq 70\%$ in our results. While these guidelines offer a general perspective on the sample size required, researchers can expect some variability on the data requirements as a consequence of the characteristics of each specific data set. For instance, for data with continuous or count observations, researchers can expect to obtain the same level of decoding accuracy with a smaller data requirement. Finally, we note that a more accurate decoding is to be expected for states with higher self-transitions.

Conclusion

Our study demonstrates that the MHMM can effectively handle the data requirements needed to model complex data using current standards. By utilizing multivariate data, we found that it is possible to mitigate the negative effects of data containing states with varying degrees of distinctiveness and separation. Additionally, the number of individuals and the number of observations per individual can complement each other when estimating group-level parameters. By implementing efficient data usage in analysis design, researchers can extract valuable insights on complex processes using the MHMM in a reliable and replicable manner. These findings can assist researchers in making informed decisions when analyzing their intensive longitudinal data.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work made use of the Dutch national e-infrastructure Snellius with the support of the SURF Cooperative using grant no. EINF-2570, which is (partly) financed by the Dutch Research Council (NWO).

Role of the funders: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Aarts, E. (2019). *mHMMbayes: Multilevel hidden Markov models using Bayesian estimation*. https://CRAN.R-project.org/package=mHMMbayes

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, *37*(6A), 3099–3132. https://doi.org/10.1214/09-AOS689

Altman, R. M. K. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, *102*(477), 201–210. https://doi.org/10.1198/016214506000001086

Alvarez, I., Niemi, J., & Simpson, M. (2014). Bayesian inference for a covariance matrix. In *Conference on Applied Statistics in Agriculture*.

Ariens, S., Ceulemans, E., & Adolf, J. K. (2020). Time series analysis of intensive longitudinal data in psychosomatic research: A methodological overview. *Journal of Psychosomatic Research*, *137*, 110191. https://doi.org/10.1016/j.jpsychores.2020.110191

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Bartolucci, F., & Farcomeni, A. (2015). A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. *Biometrics*, *71*(1), 80–89. https://doi.org/10.1111/biom.12224

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2012). *Latent Markov models for longitudinal data*. CRC Press.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, *37*(6), 1554–1563. https://doi.org/10.1214/aoms/1177699147

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171. https://doi.org/10.1214/aoms/1177697196

Beyer, H. L., Morales, J. M., Murray, D., & Fortin, M.-J. (2013). The effectiveness of Bayesian state-space models for estimating behavioural states from movement paths. *Methods in Ecology and Evolution*, *4*(5), 433–441. https://doi.org/10.1111/2041-210X.12026

Bode, N. W., & Seitz, M. J. (2018). Using hidden Markov models to characterise intermittent social behaviour in fish shoals. *The Science of Nature*, *105*(1–2), 7. https://doi.org/10.1007/s00114-017-1534-9

Brekkan, A., Jönsson, S., Karlsson, M. O., & Plan, E. L. (2019). Handling underlying discrete variables with bivariate mixed hidden Markov models in NONMEM. *Journal of Pharmacokinetics and Pharmacodynamics*, *46*(6), 591–604. https://doi.org/10.1007/s10928-019-09658-z

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.2307/1390675

Cabrera-Quiros, L., Demetriou, A., Gedik, E., van der Meij, L., & Hung, H. (2021). The MatchNMingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing*, *12*(1), 113–130. https://doi.org/10.1109/TAFFC.2018.2848914

Cappé, O., Moulines, E., & Ryden, T. (2005). *Inference in hidden Markov models*. Springer.

Chiang, S., Vannucci, M., Goldenholz, D. M., Moss, R., & Stern, J. M. (2018). Epilepsy as a dynamic disease: A Bayesian model for differentiating seizure risk from natural variability. *Epilepsia Open*, *3*(2), 236–246. https://doi.org/10.1002/epi4.12112

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, *57*(1), 505–528. https://doi.org/10.1146/annurev.psych.57.102904.190146

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*(1), 131–157. https://doi.org/10.1207/s15327906mbr2701_8

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., & Hooten, M. B. (2018). A guide to Bayesian model checking for ecologists. *Ecological Monographs*, *88*(4), 526–542. https://doi.org/10.1002/ecm.1314

de Haan-Rietdijk, S., Kuppens, P., Bergeman, C. S., Sheeber, L. B., Allen, N. B., & Hamaker, E. L. (2017). On the use of mixed Markov models for intensive longitudinal data. *Multivariate Behavioral Research*, *52*(6), 747–767. https://doi.org/10.1080/00273171.2017.1370364

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

DeRuiter, S. L., Langrock, R., Skirbutas, T., Goldbogen, J. A., Calambokidis, J., Friedlaender, A. S., & Southall, B. L. (2017). A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *The Annals of Applied Statistics*, *11*(1), 362–392. https://doi.org/10.1214/16-AOAS1008

Ebner-Priemer, U. W., & Trull, T. J. (2009). Ecological momentary assessment of mood disorders and mood dysregulation. *Psychological Assessment*, *21*(4), 463–475. https://doi.org/10.1037/a0017075

Finch, W. H., & French, B. F. (2014). Multilevel latent class analysis: Parametric and nonparametric models. *The Journal of Experimental Education*, *82*(3), 307–333. https://doi.org/10.1080/00220973.2013.813361

Galindo Garre, F., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by bayesian posterior mode estimation. *Behaviormetrika*, *33*(1), 43–59. https://doi.org/10.2333/bhmk.33.43

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.

Hale, W. W., III (n.d.). *Therapy improvement and non-specific factors (human-ethological observations of adolescents and therapists)*.

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177/0963721416666518

Hecht, M., & Zitzmann, S. (2021). Sample size recommendations for continuous-time models: compensating shorter time series with larger numbers of persons and vice versa. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 229–236. https://doi.org/10.1080/10705511.2020.1779069

Holsclaw, T., Greene, A. M., Robertson, A. W., & Smyth, P. (2017). Bayesian nonhomogeneous Markov models via Pólya-Gamma data augmentation with applications to rainfall modeling. *The Annals of Applied Statistics*, *11*(1), 393–426. https://doi.org/10.1214/16-AOAS1009

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Inaba, Y. (2017). Mixed hidden Markov models for clinical research with discrete repeated measurements. *American Journal of Theoretical and Applied Statistics*, *6*(6), 290. https://doi.org/10.11648/j.ajtas.20170606.15

Jackson, J. C., Albert, P. S., & Zhang, Z. (2015). A two-state mixed hidden Markov model for risky teenage driving behavior. *The Annals of Applied Statistics*, *9*(2), 849–865. https://doi.org/10.1214/14-AOAS765

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, *20*(1), 50–67. https://doi.org/10.1214/088342305000000016

Jonsen, I. (2016). Joint estimation over multiple individuals improves behavioural state inference from animal movement data. *Scientific Reports*, *6*(1), 20625. https://doi.org/10.1038/srep20625

Kang, K., Song, X., Hu, X. J., & Zhu, H. (2019). Bayesian adaptive group lasso with semiparametric hidden Markov models. *Statistics in Medicine*, *38*(9), 1634–1650. https://doi.org/10.1002/sim.8051

Landau, S., & Stahl, D. (2013). Sample size and power calculations for medical studies by simulation when closed form expressions are not available. *Statistical Methods in Medical Research*, *22*(3), 324–345. https://doi.org/10.1177/0962280212439578

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., & Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data: Hidden Markov models and extensions. *Ecology*, *93*(11), 2336–2342. https://doi.org/10.1890/11-2241.1

Lemaignan, S., Edmunds, C. E., Senft, E., & Belpaeme, T. (2018). The PInSoRo dataset: Supporting the datadriven study of child-child and child-robot social dynamics.

PLoS One, *13*(10), e0205999. https://doi.org/10.1371/journal.pone.0205999

Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, *128*(7), 912–928. https://doi.org/10.1111/oik.05985

Lin, X., Mermelstein, R., & Hedeker, D. (2020). Mixed location scale hidden Markov model for the analysis of intensive longitudinal data. *Health Services and Outcomes Research Methodology*, *20*(4), 222–236. https://doi.org/10.1007/s10742-020-00217-5

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer

Lynch, S. M., & Western, B. (2004). Bayesian posterior predictive checks for complex models. *Sociological Methods & Research*, *32*(3), 301–335. https://doi.org/10.1177/0049124103257303

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*(2), 181–220. https://doi.org/10.1207/s15327906mbr3302_1

Maruotti, A. (2011). Mixed hidden Markov models for longitudinal data: An overview. *International Statistical Review*, *79*(3), 427–454. https://doi.org/10.1111/j.1751-5823.2011.00160.x

Maruotti, A., Fabbri, M., & Rizzolli, M. (2022). Multilevel hidden Markov models for behavioral data: A hawk-and-dove experiment. *Multivariate Behavioral Research*, *57*(5), 825–839. https://doi.org/10.1080/00273171.2021.1912583

Maruotti, A., & Rocci, R. (2012). A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. *Statistics in Medicine*, *31*(9), 871–886. https://doi.org/10.1002/sim.4478

Maruotti, A., & Rydén, T. (2009). A semiparametric approach to hidden Markov models under longitudinal observations. *Statistics and Computing*, *19*(4), 381–393. https://doi.org/10.1007/s11222-008-9099-2

McClintock, B. T. (2021). Worth the effort? A practical examination of random effects in hidden Markov models for animal telemetry data. *Methods in Ecology and Evolution*, *12*(8), 1475–1497. https://doi.org/10.1111/2041-210X.13619

McClintock, B. T., Langrock, R., Gimenez, O., Cam, E., Borchers, D. L., Glennie, R., & Patterson, T. A. (2020). Uncovering ecological state dynamics with hidden Markov models. *Ecology Letters*, *23*(12), 1878–1903. https://doi.org/10.1111/ele.13610

McClintock, B. T., & Michelot, T. (2018). momentuHMM: R package for generalized hidden Markov models of animal movement. *Methods in Ecology and Evolution*, *9*(6), 1518–1530. https://doi.org/10.1111/2041-210X.12995

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Chapman and Hall.

McKellar, A. E., Langrock, R., Walters, J. R., & Kesler, D. C. (2015). Using mixed hidden Markov models to examine behavioral states in a cooperatively breeding bird. *Behavioral Ecology*, *26*(1), 148–157. https://doi.org/10.1093/beheco/aru171

McNeish, D. (2019). Poisson multilevel models with small samples. *Multivariate Behavioral Research*, *54*(3), 444–455. https://doi.org/10.1080/00273171.2018.1545630

Mehl, M. R., & Conner, T. S. (Eds.). (2012). *Handbook of research methods for studying daily life*. The Guilford Press.

Mildiner Moraga, S., & Aarts, E. (2022). Accompanying code for "Go multivariate: recommendations on multilevel hidden Markov models with categorical data of varying complexity" (1.0.0). *Zenodo*. https://doi.org/10.5281/zenodo.6385219

Nylund-Gibson, K., Garber, A. C., Carter, D. B., Chan, M., Arch, D. A. N., Simon, O., Whaling, K., Tartt, E., & Lawrie, S. I. (2022). Ten frequently asked questions about latent transition analysis. *Psychological Methods*. https://doi.org/10.1037/met0000486

Orfanos, S., Akther, S. F., Abdul-Basit, M., McCabe, R., & Priebe, S. (2017). Using video-annotation software to identify interactions in group therapies for schizophrenia: Assessing reliability and associations with outcomes. *BMC Psychiatry*, *17*(1), 65. https://doi.org/10.1186/s12888-017-1217-2

Park, J. H. (2012). A unified method for dynamic and cross-sectional heterogeneity: Introducing hidden Markov panel models. *American Journal of Political Science*, *56*(4), 1040–1054. https://doi.org/10.1111/j.1540-5907.2012.00590.x

Pohle, J., Langrock, R., van Beest, F. M., & Schmidt, N. M. (2017). Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, *22*(3), 270–293. https://doi.org/10.1007/s13253-017-0283-8

R Core Team. (2021). *R: A language and environment for statistical computing*. R Development Core Team.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. https://doi.org/10.1109/5.18626

Raffa, J. D., & Dubin, J. A. (2015). Multivariate longitudinal data analysis with mixed effects hidden Markov models. *Biometrics*, *71*(3), 821–831. https://doi.org/10.1111/biom.12296

Ronao, C. A., & Cho, S. B. (2017). Recognizing human activities from smartphone sensors using hierarchical continuous hidden Markov models. *International Journal of Distributed Sensor Networks*, *13*(1), 155014771668368. https://doi.org/10.1177/1550147716683687

Rücker, G., & Schwarzer, G. (2014). Presenting simulation results in a nested loop plot. *BMC Medical Research Methodology*, *14*(1), 129. https://doi.org/10.1186/1471-2288-14-129

Rueda, O. M., Rueda, C., & Diaz-Uriarte, R. (2013). A Bayesian HMM with random effects and an unknown number of states for DNA copy number analysis. *Journal of Statistical Computation and Simulation*, *83*(1), 82–96. https://doi.org/10.1080/00949655.2011.609818

Ruiz-Suarez, S., Leos-Barajas, V., & Morales, J. M. (2022). Hidden Markov and semi-Markov models when and why are these models useful for classifying states in time series data? *Journal of Agricultural, Biological and*

Environmental Statistics, *27*(2), 339–363. https://doi.org/10.1007/s13253-021-00483-x

Rydén, T. (2008). EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Analysis*, *3*(4), 659–688. https://doi.org/10.1214/08-BA326

Schafer, T. L. J., Wikle, C. K., Vonbank, J. A., Ballard, B. M., & Weegman, M. D. (2020). A Bayesian Markov model with Pólya-Gamma sampling for estimating individual behavior transition probabilities from accelerometer classifications. *Journal of Agricultural, Biological and Environmental Statistics*, *25*(3), 365–382. https://doi.org/10.1007/s13253-020-00399-y

Schultzberg, M., & Muthén, B. (2018). Number of subjects and time points needed for multilevel time-series analysis: A simulation study of dynamic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 495–515. https://doi.org/10.1080/10705511.2017.1392862

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, *97*(457), 337–351. https://doi.org/10.1198/016214502753479464

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shirley, K. E., Small, D. S., Lynch, K. G., Maisto, S. A., & Oslin, D. W. (2012). Hidden Markov models for alcoholism treatment trial data. *The Annals of Applied Statistics*, *4*(1), 366–395. https://doi.org/10.1214/09-AOAS282

Towner, A. V., Leos-Barajas, V., Langrock, R., Schick, R. S., Smale, M. J., Kaschke, T., Jewell, O. J. D., & Papastamatiou, Y. P. (2016). Sex-specific and individual preferences for hunting strategies in white sharks. *Functional Ecology*, *30*(8), 1397–1407. https://doi.org/10.1111/1365-2435.12613

van de Schoot, R., & Miočević, M. (Eds.). (2020). *Small sample size solutions: A guide for applied researchers and practitioners* (1st ed.). Routledge.

Vermunt, J. K., Langeheine, R., & Bockenholt, U. (1999). Discrete-time discrete-state latent Markov Models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, *24*(2), 179–207. https://doi.org/10.3102/10769986024002179

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. In *The SAGE encyclopedia of social sciences research methods* (Vol. 2, pp. 549–553). SAGE Publishing.

Visser, I. (2011). Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. *Journal of Mathematical Psychology*, *55*(6), 403–415. https://doi.org/10.1016/j.jmp.2011.08.002

Visser, I., Raijmakers, M. E., & Van Der Maas, H. L. (2009). Hidden Markov models for individual time series. In *Dynamic process methodology in the social and developmental sciences* (pp. 269–289). Springer. https://doi.org/10.1007/978-0-387-95922-1

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269. https://doi.org/10.1109/TIT.1967.1054010

Walls, T. A., Jung, H., & Schwartz, J. E. (2006). Multilevel models for intensive longitudinal data. In *Models for intensive longitudinal data* (pp. 3–37). Oxford University Press.

Walls, T. A., & Schafer, J. L. (2012). *Models for intensive longitudinal data*. Oxford University Press.

Wiggins, L. M. (1973). *Panel analysis: Latent probability models for attitude and behavior processes*. Jossey-Bass.

Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods*, *14*(3), 183–201. https://doi.org/10.1037/a0015858

Wurpts, I. C., & Geiser, C. (2014). Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. *Frontiers in Psychology*, *5*, 920. https://doi.org/10.3389/fpsyg.2014.00920

Xia, Y. M., & Tang, N. S. (2019). Bayesian analysis for mixture of latent variable hidden Markov models with multivariate longitudinal data. *Computational Statistics & Data Analysis*, *132*, 190–211. https://doi.org/10.1016/j.csda.2018.08.004

Xia, Y. M., Tang, N. S., & Gou, J. W. (2016). Generalized linear latent models for multivariate longitudinal measurements mixed with hidden Markov models. *Journal of Multivariate Analysis*, *152*, 259–275. https://doi.org/10.1016/j.jmva.2016.09.001

Zhang, Q., Jones, A. S., Rijmen, F., & Ip, E. H. (2010). Multivariate discrete hidden Markov models for domain-based measurements and assessment of risk factors in child development. *Journal of Computational and Graphical Statistics*, *19*(3), 746–765. https://doi.org/10.1198/jcgs.2010.09015

Zhang, X., Zhao, X., & Rong, J. (2014). A study of individual characteristics of driving behavior based on hidden Markov model. *Sensors & Transducers*, *167*(3), 194.

Zucchini, W., Macdonald, I. L., & Langrock, R. (2017). *Hidden Markov models for time series: An introduction using R* (2nd ed.). CRC Press.