Estimation of the multilevel hidden Markov model

Emmeke Aarts

2019-10-30

There are several methods to estimate the parameters of a hidden Markov model (HMM). To be complete, we discuss three methods typically used when assumed that the data is generated by one (uni-level) model. That is, the Maximum Likelihood approach, the Baum Welch algorithm utilizing the forward backward probabilities, and the Bayesian estimation method. Note that all these methods assume that the number of states is known from the context of the application, i.e., specified by the user. The issue of determining the number of states is discussed in the vignette "tutorial-mhmm".
After discussing the simplified case of estimating the parameters where the data consists of only one observed sequence (or, with multiple sequences, assuming that all data is generated by one identical model), we proceed with elaborating on estimating the parameters of a multilevel hidden Markov model.

Estimating the parameters of the HMM

Maximum likelihood (ML)

ML estimation can be used to estimate the parameters of the HMM. The relevant likelihood function has a convenient form:

$$L_T = \delta \mathbf{P}(o_1)\mathbf{\Gamma}\mathbf{P}(o_2)\mathbf{\Gamma}\mathbf{P}(o_3)\dots\mathbf{\Gamma}\mathbf{P}(o_T)\mathbf{1}'. \tag{1}$$

In equation 1, $\mathbf{P}(o_t)$ denotes a diagonal matrix with the state-dependent conditional probabilities of observing $O_t = o$ as entries, δ denotes the distribution of the initial probabilities π_i, $\mathbf{\Gamma}$ denotes the transition probability matrix, and $\mathbf{1}'$ is a column vector consisting of m (i.e., the number of distinct states) elements which all have the value one. See the vignette "tutorial-mhmm" for an explanation of these quantities. Direct maximization of the log-likelihood poses no problems even for very long sequences, provided measures are taken to avoid numerical underflow[1].

Expectation Maximization (EM) or Baum-Welch algorithm

The EM algorithm (Dempster, Laird, and Rubin 1977), in this context also known as the Baum-Welch algorithm (Baum et al. 1970; Rabiner 1989), can also be used to maximize the log-likelihood function. Here, the unobserved latent states are treated as missing data, and quantities known as the forward and the backward probabilities are used to obtain the 'complete-data log-likelihood' of the HMM parameters: the log-likelihood based on both the observed event sequence and the unobserved, or "missing", latent states. The forward probabilities $\boldsymbol{\alpha}_t(i)$ denote the joint probability of the observed event sequence from time point 1 to t and state S at time point t being i:

$$\boldsymbol{\alpha}_t(i) = Pr(O_1 = o_1, O_2 = o_2, \dots, O_t = o_t, S_t = i). \tag{2}$$

The name "forward probabilities" derives from the fact that when computing the forward probabilities $\boldsymbol{\alpha}_t$, one evaluates the sequence of hidden states in the chronological order (i.e., forward in time) until time point t. The backward probabilities $\boldsymbol{\beta}_t(i)$ denote the conditional probability of the observed event sequence after time point t until the end, so from $t+1, t+2, \dots, T$, given that state S at time point t equals i:

$$\boldsymbol{\beta}_t(i) = Pr(O_{t+1} = o_{t+1}, O_{t+2} = o_{t+2}, \dots, O_T = o_T \mid S_t = i). \tag{3}$$

[1] In case of a discrete state-dependent distribution, multiplication of the elements of the likelihood function, being made up of probabilities, results in progressively smaller outcomes as one proceeds in the function from 1 to T, eventually rounding to zero. To avoid this phenomenon, referred to as numerical underflow, a so-called scaling factor is implemented, see e.g., Zucchini and MacDonald (2016)

When computing the backward probabilities $\boldsymbol{\beta}_t$, one evaluates the sequence of hidden states in the reversed order, i.e., from $S_T, S_{T-1}, \ldots, S_{t+1}$. The forward and backward probabilities together cover the complete event sequence from $t = 1$ to T, and combined give the joint probability of the complete event sequence and state S at time point t being i:

$$\boldsymbol{\alpha}_t(i)\boldsymbol{\beta}_t(i) = Pr(O_1 = o_1, O_2 = o_2, \ldots, O_T = o_T, S_t = i). \tag{4}$$

We refer to Cappé (2005) for a discussion on the advantages of combining forward and backward probability information in the EM algorithm over direct maximization of the likelihood for the HMM.

Bayesian estimation

A third approach is to use Bayesian estimation to infer the parameters of the HMM. We refer to e.g., Gelman et al. (2014), Lynch (2007), Rossi, Allenby, and McCulloch (2012) for an in-depth exposition of Bayesian statistics. In general terms, the difference between frequentist and Bayesian estimation is the following. In frequentist estimation, we view the parameters as fixed entities in the population, which are subject only to sampling fluctuation (as quantified in the standard error of the estimate). In Bayesian estimation, however, we assume that each parameter follows a given distribution. The general shape of this distribution is determined beforehand, using a prior distribution. This prior distribution not only determines the shape of the parameter distribution, but also allows for giving some information to the model with respect to the most likely values of the parameter that is estimated. To arrive at the final distribution for the parameter - which is called the posterior distribution -, the prior distribution is combined with the likelihood function of the data using Bayes' theorem. Here, the likelihood function provides us with the probability of the data given the parameters. While any aspect of these distributions may be of interest, the emphasis is usually on the mean (or median) of the posterior distribution, which serves as the point estimate of the parameter of interest (analogous to the frequentist parameter estimates). In the event that one has no or vague expectations about the possible parameter values, one can specify "non-informative" priors (e.g., uniform distribution). That is, one can choose parameters of the prior distributions, so called hyper-parameters, such that the parameters may assume a wide range of possible values. Non-informative priors therefore express a lack of knowledge.
In the implemented hidden Markov model, both the transitions from state i at time point t to any of the other states at time point $t + 1$ and the observed outcomes within state i follow a categorical distribution, with parameter sets $\boldsymbol{\Gamma}_i$ (i.e., the probabilities in row i of the transition probability matrix $\boldsymbol{\Gamma}$) and $\boldsymbol{\theta}_i$ (i.e., the state i dependent probabilities of observing an act). A convenient (conjugate) prior distribution on the parameters of the categorical distribution is a (symmetric) Dirichlet distribution. We assume that the rows of $\boldsymbol{\Gamma}$ and the state-dependent probabilities $\boldsymbol{\theta}_i$ are independent. That is,

$$S_{t=2,\ldots,T} \sim \boldsymbol{\Gamma}_{S_{t-1}} \quad \text{with} \quad \boldsymbol{\Gamma}_i \sim \text{Dir}(\mathbf{a}_{10}) \quad \text{and} \tag{5}$$

$$O_{t=1,\ldots,T} \sim \boldsymbol{\theta}_{S_t} \quad \text{with} \quad \boldsymbol{\theta}_i \sim \text{Dir}(\mathbf{a}_{20}), \tag{6}$$

where the probability distribution of the current state S_t is given by the row of the transition probability matrix $\boldsymbol{\Gamma}$ corresponding to the previous state in the hidden state sequence S_{t-1}. The probability distribution of S_t given by $\boldsymbol{\Gamma}$ holds for states after the first time point, i.e., t starts at 2 as there is no previous state in the hidden state sequence for state S at $t = 1$. The probability of the first state in the hidden state sequence S_1 is given by the initial probabilities of the states π_i. The probability distribution of the observed event O_t is given by state-dependent probabilities $\boldsymbol{\theta}_i$ corresponding to the current state S_t. The hyper-parameter $\mathbf{a}_{10}$ of the prior Dirichlet distribution on $\boldsymbol{\Gamma}_i$ is a vector with length equal to the number of states m, and the hyper-parameter $\mathbf{a}_{20}$ of the prior Dirichlet distribution on $\boldsymbol{\theta}_i$ is a vector with length equal to the number of categorical outcomes q. Note that in this model, the hyper-parameter values are assumed invariant over the states i. The initial probabilities of the states π_i are assumed to coincide with the stationary distribution of $\boldsymbol{\Gamma}$ and are therefore not independent (to-be-estimated) parameters.
Given these distributions, our goal is to construct the joint posterior distribution of the hidden state sequence and the parameter estimates, given the observed event sequence and the hyper-parameters

$$\Pr\left((S_t), \boldsymbol{\Gamma}_i, \boldsymbol{\theta}_i \mid (O_t)\right) \propto \Pr\left((O_t) \mid (S_t), \boldsymbol{\theta}_i\right) \Pr\left((S_t) \mid \boldsymbol{\Gamma}_i\right) \Pr\left(\boldsymbol{\Gamma}_i \mid \mathbf{a}_{10}\right) \Pr\left(\boldsymbol{\theta}_i \mid \mathbf{a}_{20}\right) \tag{7}$$

by drawing samples from the posterior distribution. By applying a Gibbs sampler, we can iteratively sample from the appropriate conditional posterior distributions of S_t, $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\theta}_i$, given the remaining parameters

in the model. In short, the Gibbs sampler iterates between the following two steps: first the hidden state sequence $S_1, S_2, \ldots, S_T$ is sampled, given, the observed event sequence $O_1, O_2, \ldots, O_T$, and the current values of the parameters $\mathbf{\Gamma}$ and $\boldsymbol{\theta}_i$. Subsequently, the remaining parameters in the model ($\mathbf{\Gamma}_i$ and $\boldsymbol{\theta}_i$) are updated by sampling them conditional on the sampled hidden state sequence $S_1, S_2, \ldots, S_T$ and observed event sequence $O_1, O_2, \ldots, O_T$.

Sampling the hidden state sequence of the HMM by means of the Gibbs sampler can be performed in various ways. Here, we use the approach outlined by Scott (2002). That is, we use the forward-backward Gibbs sampler, in which first the forward probabilities $\boldsymbol{\alpha}_t(i)$ (i.e., the joint probability of state $S = i$ at time point t and the observed event sequence from time point 1 to t) as given in equation 2 are obtained, after which the hidden state sequence is sampled in a backward run (i.e., drawing $S_T, S_{T-1}, \ldots, S_1$) using the corresponding forward probabilities $\boldsymbol{\alpha}_{T:1}$. The forward-backward Gibbs sampler produces sampled values that rapidly represent the complete area of the posterior distribution, and produces useful quantities as byproducts, such as the log-likelihood of the observed data given the current draws of the parameters in each iteration (Scott 2002). In the section "*Hybrid Metropolis within Gibbs sampler used to fit the multilevel HMM*" , we provide a more detailed description of how the Gibbs sampler proceeds for the HMM.

As it generally takes a number of iterations before the Gibbs sampler converges to the appropriate region of the posterior distribution, the initial iterations are usually discarded as a 'burn-in' period. The remaining sampled values of $\mathbf{\Gamma}_i$ and $\boldsymbol{\theta}_i$ provide the posterior distributions of their respective parameters.

A problem that can arise when using Bayesian estimation in this context is "label switching", i.e., as the hidden states of the HMM have no a priori ordering or interpretation, their labels (i.e., which state represents what) can switch over the iterations of the Gibbs sampler, without affecting the likelihood of the model (see e.g., Scott 2002; Jasra, Holmes, and Stephens 2005). As a result, the marginal posterior distributions of the parameters are impossible to interpret because they represent the distribution of multiple states. Sometimes, using reasonable starting values (i.e., the user-specified parameter values of the "zero-th" iteration used to start the MCMC sampler) suffices to prevent label switching. Otherwise, possible solutions are to set constraints on the parameters of the state-dependent distribution, or use (weakly) informative priors on the state-dependent distributions (Scott 2002). Hence, before making inferences from the obtained marginal distributions, one should first assess if the problem of label switching is present (e.g., by using plots of the sampled parameter values of the state-dependent distributions over the iterations), and if necessary, take steps to prevent the problem of label switching. In our own experience, the use of reasonable starting values always sufficed to prevent label switching.

Both EM and Bayesian Gibbs sampling are viable inferential procedures for HMMs, but for more complex HMMs such as multilevel HMMs, the Bayesian estimation method has several advantages (e.g., lower computational cost, and less computation time) over the EM algorithm. We refer to Rydén (2008) for a comparison on frequentist (i.e., the EM algorithm) and Bayesian approaches.

Estimating the parameters of the multilevel HMM

Bayesian estimation of multilevel models

Bayesian estimation is particularly suited to model multilevel models. In the multilevel model, we have a multi-layered structure in the parameters. For the HMM, we have subject level parameters at the first level pertaining to the observations within a subject, and group level parameters at the second level that describe the mean and variation within the group, as inferred from the sample of subjects. To illustrate the multilevel model, suppose that we have K subjects for which we have each H observations on their number of cups of coffee consumed per day y, i.e., subject $k \in \{1, 2, \ldots, K\}$ and observation $h \in \{1, 2, \ldots, H\}$. Hence, at the first level, we have daily observations on coffee consumption within subjects: y_{11}, y_{12}, …, y_{1H}, y_{21}, y_{22}, …, y_{2H}, y_{K1}, y_{K2}, …, y_{KH}. Using a multilevel model, the observations of each subject are distributed according to the same distribution Q, but each subject has its own parameter set θ_k. That is:

$$y_{kh} \sim Q(\theta_k). \tag{8}$$

In addition, the subject-specific parameter sets θ_k are realizations of a common group level distribution W with parameter set Λ:

$$\theta_k \sim W(\Lambda). \tag{9}$$

That is, in the multilevel model, the subject level model parameters that pertain to the observations within a subject are assumed to be random draws from a given distribution, and, as such, are denoted as "random", independent of the used estimation method (i.e., Bayesian or classical frequentist estimation). This multi-layered structure fits naturally into a Bayesian paradigm since in Bayesian estimation, model parameters are by definition viewed as random. That is, parameters follow a given distribution, where the prior distribution expresses the prior expectations with respect to the most likely values of the model parameters. In the multilevel model, the prior expectations of the subject level model parameter values are reflected in the group level distribution. Hence, in Bayesian estimation, the prior distribution for the subject level parameters, is given by the group level distribution. The group level distribution provides information on the location (e.g., mean) of the subject level (i.e., subject-specific) parameters, and on the variation in the subject level parameters. As the Normal distribution is a flexible distribution with parameters that easily relate to this interpretation, the group level distribution is often taken to be a normal distribution.

To illustrate the notion of the group level (prior) distribution, suppose we assume a Poisson distribution for the observations on daily coffee consumption within each subject k, and a Normal group level distribution on the Poisson mean. In this case, the set of hyper-parameters (i.e., the parameters of the group level distribution, here the mean (Λ_μ) and variance (Λ_{σ^2}) of the Normal distribution) on the Poisson mean denote the group mean number of cups of coffee consumed per day over subjects, and the variation in the mean number of cups of coffee consumed per day between subjects.

Finally, in fitting the multilevel model using Bayesian estimation, a prior distribution is placed on each of these hyper-parameters. Prior distributions on hyper-parameters are referred to as hyper-priors and allow the hyper-parameters to have a distribution instead of being fixed. That is, as the parameters that characterize the group level prior distribution (i.e., the hyper-parameters) are now also quantities of interest (i.e., to-be-estimated), they are viewed as random in Bayesian estimation methods. The randomness in the hyper-parameters is thus specific to the Bayesian estimation method of the multilevel model, in contrast to the randomness in the subject level parameters.

To continue our example, the hyper-prior on the mean of the Normal prior distribution for the subject level mean of cups of coffee consumed daily denote our prior belief on the mean number of cups of coffee consumed per day in the group. The hyper-prior on the variance of the Normal prior distribution for the subject level mean of cups of coffee consumed per day denote our prior belief on how much this mean number of cups of coffee varies over subjects. Often, the hyper-prior distribution and its values are chosen to be vague (i.e., not informative), like a uniform distribution:

$$\begin{aligned} \Lambda_\mu &\sim U(0, 20), \\ \Lambda_{\sigma^2} &\sim U(0, 500). \end{aligned} \tag{10}$$

See e.g., Gelman et al. (2014), Lynch (2007), Rossi, Allenby, and McCulloch (2012) for an in-depth exposition of various multilevel Bayesian models and e.g. Snijders and Bosker (2011), Hox, Moerbeek, and Schoot (2017), Goldstein (2011) for coverage of the classical, frequentist approach to multilevel (also called hierarchical or random effects) models.

The present implemented multilevel model pertains only to data comprised of (multivariate) categorical observations, and, possibly, time invariant covariates (i.e., for each covariate, we have one value per subject). As such, data is comprised of $O_{d,k,t}$ observations on the categorical outcome(s) for categorical outcome variable $d = 1, 2, \ldots, D$ for subject $k = 1, 2, \ldots, K$ at time point $t = 1, 2, \ldots, T$. In addition, we have a matrix $\boldsymbol{X}$ that consists of k covariate vectors with length $p \times 1$, $\boldsymbol{X}_k = (X_{k1}, X_{k2}, \ldots, X_{kp})$. As yet, the explanation of the estimation procedure in this vignette is restricted to the univariate case for simplicity. That is, to the instance that we only have one observed categorical outcome variable per subject, and our outcome data is comprised of O_{kt} observations. However, the explanation extents quite naturally to the multivariate case.

Given these observations, we construct a multilevel model for each of the parameters in the HMM with q observable categorical outcomes, and m hidden states, possibly predicted by p covariates. Using the multilevel framework, each parameter is assumed to follow a given distribution, and the parameter value of a given subject represents a draw from this common (i.e., group level) distribution. Hence, in the multilevel Bayesian HMM, the parameters are: the subject-specific transition probability matrix $\boldsymbol{\Gamma}_k$ with transition probabilities γ_{kij} and the subject-specific state-dependent probability distribution denoting the subject-specific probabilities $\boldsymbol{\theta}_{ki}$ of categorical outcomes within state i. The initial probabilities of the states $\pi_{k,j}$ are not estimated as π_k is assumed to be the stationary distribution of $\boldsymbol{\Gamma}_k$. Subsequently, the parameter values of the subjects are

assumed to be realizations of a model component and state specific (multivariate) Normal distribution. We discuss the multilevel model for the two components of the HMM ($\boldsymbol{\Gamma}_k$ and $\boldsymbol{\theta}_{ki}$ separately. Table 1 provides an overview of the used symbols in the multilevel models related to the two components of the HMM. We use the subscript 0 to denote values of the hyper-prior distribution parameters.

Multilevel model for the state-dependent probabilities $\boldsymbol{\theta}_{ki}$

In the standard (non-multilevel) Bayesian HMM estimation, we specified a Dirichlet prior distribution on the state-dependent probabilities $\boldsymbol{\theta}_i$. To provide a flexible model that allows for the inclusion of random effects and (time invariant) covariates, we follow Altman (2007) and extend the subject-specific state-dependent probabilities $\boldsymbol{\theta}_{ki}$ to a multinomial logit (MNL) model. Hence, we utilize a linear predictor function to estimate the probability of observing categorical outcome l within state i. The state i specific linear predictor function at the subject level consists of $q-1$ random intercepts (i.e., each subject has its own intercept). That is, each categorical outcome l has its own intercept, with the exception of the first categorical outcome in the set for which the intercept is omitted for reasons of model identification (i.e., not all probabilities can be estimated freely as within subject k and state i, the probabilities need to add up to 1). By making the intercepts random (i.e., each subject has its own intercept), we accommodate heterogeneity between subjects in their state conditional probabilities. Hence, in the MNL model for $\boldsymbol{\theta}_{ki}$, subject k's probabilities of observing categorical outcome $l \in \{1, 2, \ldots, q\}$ within a state $i \in \{1, 2, \ldots, m\}$, θ_{kil}, are modeled using m batches of $q-1$ random intercepts, $\boldsymbol{\alpha}_{(O)ki} = (\alpha_{(O)ki2}, \alpha_{(O)ki3}, \ldots, \alpha_{(O)kiq})$. That is,

$$\theta_{kil} = \frac{\exp(\alpha_{(O)kil})}{1 + \sum_{\bar{l}=2}^{q} \exp(\alpha_{(O)ki\bar{l}})}, \tag{11}$$

where K, m, and q are the number of subjects, states, and categorical outcomes, respectively. The numerator is set equal to one for $l = 1$, making the first categorical outcome in the set the baseline category in every state.

At the group level, these subject-level intercepts are (possibly) partly determined by covariates that differentiate between subjects. Thus, in addition to the subject level random intercepts $\boldsymbol{\alpha}_{(O)ki}$, we have m matrices of $p * (q-1)$ fixed regression coefficients, $\boldsymbol{\beta}_{(O)i}$, where p denotes the number of used covariates. The columns of $\boldsymbol{\beta}_{(O)i}$ are $\boldsymbol{\beta}_{(O)il} = (\beta_{(O)il1}, \beta_{(O)il2}, \ldots, \beta_{(O)ilp})$ to model the random intercepts for state i and categorical outcome l given p covariates. Combining both terms, each batch of random intercepts $\boldsymbol{\alpha}_{(O)ki}$ (i.e., the batch of $q-1$ intercepts for the state i conditional probabilities of a categorical outcome for subject k) come from a state i specific population level multivariate Normal distribution, with mean vector $\bar{\boldsymbol{\alpha}}_{(O)i} + \boldsymbol{X}_{\boldsymbol{k}}^{\top} \boldsymbol{\beta}_{\boldsymbol{(O)il}}$ that has length $q-1$, and covariance Φ_i that denotes the covariance between the $q-1$ state i specific intercepts over subjects and models the dependence of the probabilities of categorical outcomes within state i (i.e., we specify a state specific multivariate Normal prior distribution on the subject-specific $\boldsymbol{\alpha}_{\boldsymbol{(O)ki}}$ parameters). A convenient hyper-prior on the hyper-parameters of the group level prior distribution is a multivariate Normal distribution for the mean vector $\bar{\boldsymbol{\alpha}}_{(O)i}$ and the fixed regression coefficients $\boldsymbol{\beta}_{\boldsymbol{(O)il}}$, and an Inverse Wishart distribution for the covariance Φ_i (see e.g., Gelman et al. 2014). That is,

$$\begin{aligned}
O_{kt} &\sim \boldsymbol{\theta}_{k,S_{kt}} && \text{with} \quad \boldsymbol{\theta}_{ki} \sim \mathrm{MNL}\big(\boldsymbol{\alpha}_{(O)ki}\big), && (12)\\
\boldsymbol{\alpha}_{(O)ki} &\sim \mathrm{N}\big(\bar{\boldsymbol{\alpha}}_{(O)i} + \boldsymbol{X}_{\boldsymbol{k}}^{\top} \boldsymbol{\beta}_{\boldsymbol{(O)il}}, \Phi_i\big) && \text{with} \quad \bar{\boldsymbol{\alpha}}_{(O)i} \sim \mathrm{N}\big(\boldsymbol{\alpha}_{(O)0}, \tfrac{1}{K_0}\Phi_i\big), && (13)\\
& && \text{and} \quad \boldsymbol{\beta}_{(O)i} \sim \mathrm{N}\big(\boldsymbol{\beta}_{(O)0}, \tfrac{1}{K_0}\Phi_i\big), && (14)\\
& && \text{and} \quad \Phi_i \sim \mathrm{IW}\big(\Phi_0, df_0\big), &&
\end{aligned}$$

where the probability distribution of the observed categorical outcomes O_{kt} is given by the subject k specific state-dependent probabilities $\boldsymbol{\theta}_{ki}$ corresponding to the current state S_{kt} (where t indicates the time point, see Table 1). The parameters $\boldsymbol{\alpha}_{(O)0}$, $\boldsymbol{\beta}_{(O)0}$, and K_0 denote the values of the parameters of the hyper-prior on the group (mean) vector $\bar{\boldsymbol{\alpha}}_{(O)i}$ and $\boldsymbol{\beta}_{(O)i}$, respectively. Here, $\boldsymbol{\alpha}_{(O)0}$ and $\boldsymbol{\beta}_{(O)0}$ represent a vector of means and K_0 denotes the number of observations (i.e., the number of hypothetical prior subjects) on which the prior mean vector $\boldsymbol{\alpha}_{(O)0}$ and $\boldsymbol{\beta}_{(O)0}$ are based, i.e., K_0 determines the weight of the prior on $\bar{\boldsymbol{\alpha}}_{(O)i}$ and $\boldsymbol{\beta}_{(O)i}$. The parameters Φ_0 and df_0, respectively, denote the values of the covariance and the degrees of freedom of the hyper-prior Inverse Wishart distribution on the population variance Φ_i of the subject-specific random

Table 1: Elements of the multilevel HMM

Symbol	Description
k	subject $k \in \{1, 2, \ldots, K\}$, where K is the total number of subjects in the dataset.
t	Time point $t \in \{1, 2, \ldots, T\}$, where T is the total length of each sequence of observations. In the current notation, T is assumed equal over subjects. Within the R package mHMMbayes this is not a requirement, however.
q	Number of distinct observation categories.
m	Number of distinct states.
p	Number of (time invariant) covariates.
O_{kt}	Observation on the categorical outcome for subject k at time point t.
S_{kt}	State for subject k at time point t for $S \in \{1, 2, \ldots, m\}$.
i	Realization of the current state S_t, where $i \in \{1, 2, \ldots, m\}$.
$\boldsymbol{X}$	Matrix of k covariate vectors with length $p \times 1$, $\boldsymbol{X}_k = (X_{k1}, X_{k2}, \ldots, X_{kp})$.
$\boldsymbol{\Gamma_k} = [\gamma_{kij}]$	Subject-specific transition probability matrix between states with the probabilities γ_{kij} of transitioning from state $S_{kt} = i$ to state $S_{k(t+1)} = j$.
$\boldsymbol{\alpha}_{(S)ki}$	subject k and state i specific batch of $m-1$ random intercepts that model the transitions from state i to the next state j.
$\bar{\boldsymbol{\alpha}}_{(S)i}$	State i specific group mean vector over the subject k batches of the $m-1$ random intercepts $\boldsymbol{\alpha}_{(S)ki}$.
$\boldsymbol{\beta}_{(S)i}$	State i specific fixed regression coefficients to predict the random intercepts $\boldsymbol{\alpha}_{(S)ki}$.
Ψ_i	State i specific covariance between the subject k batches of the $m-1$ random intercepts $\boldsymbol{\alpha}_{(S)ki}$.
$\boldsymbol{\alpha}_{(S)0}$, $\boldsymbol{\beta}_{(S)0}$, K_0	Values of the parameters of the hyper-prior on the group mean vector $\bar{\boldsymbol{\alpha}}_{(S)i}$ and fixed regression coefficients $\boldsymbol{\beta}_{(S)i}$.
Ψ_0, df_0	Values of the parameters of the hyper-prior on the group covariance Ψ_i.
O_{kt}	Observed event for subject k at time point t for $O \in \{1, 2, \ldots q\}$.
l	Realization of current event O_{kt}, where $l \in \{1, 2, \ldots, q\}$.
$\boldsymbol{\theta}_{ki} = [\theta_{kil}]$	Subject-specific state i categorical conditional distribution, with the probabilities p_{kil} of observing the categorical outcome $O_{kt} = l$ in state $S_{kt} = i$.
$\boldsymbol{\alpha}_{(O)ki}$	Subject k state i specific batch of $q-1$ random intercepts that model the probability of a categorical outcome O_{kt} within state i.
$\bar{\boldsymbol{\alpha}}_{(O)i}$	State i specific group mean vector over the subject k batches of the $q-1$ random intercepts $\boldsymbol{\alpha}_{(O)ki}$.
$\boldsymbol{\beta}_{(O)i}$	State i specific fixed regression coefficients to predict the subject-specific random intercepts $\boldsymbol{\alpha}_{(O)ki}$.
Φ_i	State i specific covariance between the subject k batches of the $q-1$ random intercepts $\boldsymbol{\alpha}_{(O)ki}$.
$\boldsymbol{\alpha}_{(O)0}$, $\boldsymbol{\beta}_{(O)0}$, K_0	Values of the parameters of the hyper-prior on the group mean vector $\bar{\boldsymbol{\alpha}}_{(O)i}$ and the fixed regression coefficients $\boldsymbol{\beta}_{(O)i}$.
Φ_0, df_0	Values of the parameters of the hyper-prior on the group covariance Φ_i.

intercepts $\boldsymbol{\alpha}_{(O)ki}$. Note that we chose the values of the parameters of the hyper-prior distributions that result in uninformative hyper-prior distributions, as such the values of the parameters of the hyper-priors are assumed invariant over the states i.

Multilevel model for the transition probability matrix $\boldsymbol{\Gamma_k}$ with transition probabilities $\boldsymbol{\gamma_{kij}}$

Similar to the state-dependent probabilities $\boldsymbol{\theta}_{ki}$, we extend each set of state i specific state transition probabilities γ_{kij} to a MNL model to allow for the inclusion of random effects and (time invariant) covariates. Hence, we use a linear predictor function to estimate the probability to transition from behavioral state i to state j. The linear predictor function consists of $m-1$ random intercepts to allow for heterogeneity between subjects in their probabilities to switch between states. That is, within row i of the transition probability matrix $\boldsymbol{\Gamma_k}$, each state j has its own intercept, where the intercept that relates to transitioning to the first state in the set is omitted for reasons of model identification (i.e., not all probabilities can be estimated freely as the row-probabilities need to add up to 1). Hence, each subject's probability to transition from behavioral state $i \in \{1, 2, \ldots, m\}$ to state $j \in \{1, 2, \ldots, m\}$ is modeled using m batches of $m-1$ random intercepts, $\boldsymbol{\alpha}_{(S)ki} = (\alpha_{(S)k13}, \ldots, \alpha_{(S)k1m}, \alpha_{(S)k23}, \ldots, \alpha_{(S)k2m}, \ldots, \alpha_{(S)km2}, \ldots, \alpha_{(S)km(m-1)})$. That is,

$$\gamma_{kij} = \frac{\exp(\alpha_{(S)kij})}{1 + \sum_{\bar{j} \in Z} \exp(\alpha_{(S)ki\bar{j}})}, \tag{15}$$

$$\text{where } Z \in \{2, \ldots, m\}$$

where K and m are again the number of subjects in the dataset, and the distinct number of states, respectively. The numerator is set equal to 1 for $j = 1$, making the first state of every row of the transition probability matrix $\boldsymbol{\Gamma}_k$ the baseline category.
At the group level, these subject-level intercepts are (possibly) partly determined by covariates that differentiate between subjects. Thus, in addition to the subject level random intercepts $\boldsymbol{\alpha}_{(S)ki}$, we have m matrices of $p * (q-1)$ fixed regression coefficients, $\boldsymbol{\beta}_{(S)i}$, where p denotes the number of used covariates. The columns of $\boldsymbol{\beta}_{(S)i}$ are $\boldsymbol{\beta}_{(S)ij} = (\beta_{(S)ij1}, \beta_{(S)ij2}, \ldots, \beta_{(S)ijp})$ to model the random intercepts denoting the probability of transitioning from behavioral state i to state j given p covariates. Combining both terms, each batch of random intercepts $\boldsymbol{\alpha}_{(S)ki}$ come from a state i specific population level multivariate Normal distribution, with mean vector $\bar{\boldsymbol{\alpha}}_{(S)i} + \boldsymbol{X_k}^\top \boldsymbol{\beta_{(S)ij}}$ that has length $q-1$, and covariance $\Psi_i i$ that denotes the covariance between the $q-1$ state i specific intercepts over subjects, and models the dependency between the probabilities of states within random intercept batch $\boldsymbol{\alpha}_{(S)ki}$ (i.e., we specify a state specific multivariate Normal prior distribution on the subject-specific $\boldsymbol{\alpha}_{(S)ki}$ parameters). A convenient hyper-prior on the hyper-parameters of the group level prior distribution is a multivariate Normal distribution for the mean vector $\bar{\boldsymbol{\alpha}}_{(S)i}$ and the fixed regression coefficients $\boldsymbol{\beta_{(S)il}}$ and an Inverse Wishart distribution for the covariance Ψ_i. That is,

$$S_{k,t=2,\ldots,T} \sim \boldsymbol{\Gamma}_{k,S_{k,t-1}} \quad \text{with} \quad \boldsymbol{\Gamma}_{k,i} \sim \text{MNL}\big(\boldsymbol{\alpha}_{(S)ki}\big), \tag{16}$$

$$\boldsymbol{\alpha}_{(S)ki} \sim \text{N}\big(\bar{\boldsymbol{\alpha}}_{(S)i} + \boldsymbol{X_k}^\top \boldsymbol{\beta_{(S)il}}, \Psi_i\big) \quad \text{with} \quad \bar{\boldsymbol{\alpha}}_{(S)i} \sim \text{N}\big(\boldsymbol{\alpha}_{(S)0}, \tfrac{1}{K_0}\Psi_i\big), \tag{17}$$

$$\text{and} \quad \boldsymbol{\beta}_{(S)i} \sim \text{N}\big(\boldsymbol{\beta}_{(S)0}, \tfrac{1}{K_0}\Psi_i\big), \tag{18}$$

$$\text{and} \quad \Psi_i \sim \text{IW}\big(\Psi_0, df_0\big),$$

where the subject-specific probability distribution of the current state S_{kt} is given by the row of the transition probability matrix $\boldsymbol{\Gamma}_k$ corresponding to the previous state in the hidden state sequence $S_{k,t-1}$. The probability distribution of S_{kt} given by $\boldsymbol{\Gamma}_k$ holds for states after the first time point, i.e., t starts at 2 as there is no previous state in the hidden state sequence for state S_{kt} at $t = 1$. The probability of the first state in the hidden state sequence $S_{k,1}$ is given by the initial probabilities of the states $\pi_{k,j}$. The parameters $\boldsymbol{\alpha}_{(S)0}$, $\boldsymbol{\beta}_{(S)0}$, and K_0 denote the values of the parameters of the hyper-prior on the group (mean) vector $\bar{\boldsymbol{\alpha}}_{(S)i}$ and $\boldsymbol{\beta}_{(S)i}$, respectively. Here, $\boldsymbol{\alpha}_{(S)0}$ and $\boldsymbol{\beta}_{(S)0}$ represent a vector of means and K_0 denotes the number of observations (i.e., the number of hypothetical prior subjects) on which the prior mean vector $\boldsymbol{\alpha}_{(S)0}$ and $\boldsymbol{\beta}_{(S)0}$ are based. The parameters Ψ_0 and df_0, respectively, denote values of the covariance and the degrees of freedom of the hyper-prior Inverse Wishart distribution on the group variance Ψ_i of the subject-specific random intercepts $\boldsymbol{\alpha}_{(S)ki}$.

Hybrid Metropolis within Gibbs sampler used to fit the multilevel HMM

Given the above distributions, our goal is to construct the joint posterior distribution of the parameters - i.e., the subject-specific hidden state sequences, the subject level (i.e., subject-specific) parameters and the group level parameter estimates - given the observations (i.e., the observed event sequences for all k subjects that are analyzed simultaneously as one group, and the hyper-prior parameter values)

$$\begin{aligned}
&\Pr\left(S_{kt}, \boldsymbol{\Gamma}_{ki}, \boldsymbol{\alpha}_{(S)ki}, \bar{\boldsymbol{\alpha}}_{(S)i}, \boldsymbol{\beta}_{(S)i}, \Psi_i, \boldsymbol{\theta}_{ki}, \boldsymbol{\alpha}_{(O)ki}, \bar{\boldsymbol{\alpha}}_{(O)i}, \boldsymbol{\beta}_{(O)ki}, \Phi_i \mid O_{kt}, \boldsymbol{X}\right) \\
&\quad \propto \Pr\left(O_{kt} \mid S_{kt}, \boldsymbol{\theta}_{ki}\right) \Pr\left(S_{kt} \mid \boldsymbol{\Gamma}_{ki}\right) \Pr\left(\boldsymbol{\theta}_i \mid \boldsymbol{\alpha}_{(O)ki}\right) \Pr\left(\boldsymbol{\Gamma}_i \mid \boldsymbol{\alpha}_{(S)ki}\right) \Pr\left(\boldsymbol{\alpha}_{(O)ki} \mid \bar{\boldsymbol{\alpha}}_{(O)i}, \boldsymbol{X}, \boldsymbol{\beta}_{(O)i}, \Phi_i\right) \\
&\qquad \Pr\left(\boldsymbol{\alpha}_{(S)ki} \mid \bar{\boldsymbol{\alpha}}_{(S)i}, \boldsymbol{X}, \boldsymbol{\beta}_{(S)i}, \Psi_i\right) \Pr\left(\bar{\boldsymbol{\alpha}}_{(O)i} \mid \boldsymbol{\alpha}_{(O)0}, K_0, \Phi_i\right) \Pr\left(\boldsymbol{\beta}_{(O)i} \mid \boldsymbol{\beta}_{(O)0}, K_0, \Phi_i\right) \Pr\left(\Phi_i \mid \Phi_0, df_0\right) \\
&\qquad\quad \Pr\left(\bar{\boldsymbol{\alpha}}_{(S)i} \mid \boldsymbol{\alpha}_{(S)0}, K_0, \Psi_i\right) \Pr\left(\boldsymbol{\beta}_{(S)i} \mid \boldsymbol{\beta}_{(S)0}, K_0, \Psi_i\right) \Pr\left(\Psi_i \mid \Psi_0, df_0\right)
\end{aligned} \tag{19}$$

by drawing samples from the posterior distribution. We follow a MCMC sampler algorithm to iteratively sample from the appropriate conditional posterior distributions of $\boldsymbol{\alpha}_{(O)ki}$, $\boldsymbol{\alpha}_{(S)ki}$, $\bar{\boldsymbol{\alpha}}_{(O)i}$, $\boldsymbol{\beta}_{(O)i}$, Φ_i, $\bar{\boldsymbol{\alpha}}_{(S)i}$, $\boldsymbol{\beta}_{(S)i}$, and Ψ_i given the remaining parameters in the model (see below). The conditional posterior distributions of all parameters are provided in the Section "*Full conditional posterior distributions of the multilevel HMM*".
In Bayesian estimation, it is preferable to use the natural conjugate prior as prior distribution, as this conveniently results in a closed form expression of the (conditional) posterior distribution(s), making Gibbs sampling possible. However, as the non-conjugate Normal prior provides a much more intuitive interpretation of the prior group level distribution compared to using the natural conjugate prior of the MNL model, and since the asymptotic Normal approximation is excellent for the MNL likelihood (Rossi, Allenby, and McCulloch 2012), we opt for the former and do not use the conjugate prior of the MNL model. Therefore, we cannot use a Gibbs sampler to update the parameters of the subject-specific state-dependent distributions and the subject-specific transition probabilities, $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$, respectively. Instead, we use a combination of the Gibbs sampler and the Metropolis algorithm, i.e., a Hybrid Metropolis within Gibbs sampler. That is, we use a Metropolis sampler to update $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$, and we use a Gibbs sampler to update all other model parameters. There are various types of Metropolis algorithms, and each type involves specific choices. Simulation studies showed that, in line with Rossi, Allenby, and McCulloch (2012), the Random Walk (RW) Metropolis sampler outperformed the Independence Metropolis sampler in terms of efficiency for estimating the parameters of the (multilevel) HMM, we chose to use the RW Metropolis sampler to update the parameters of the subject-specific state-dependent distributions ($\boldsymbol{\alpha}_{(O)ki}$) and subject-specific state transition probabilities ($\boldsymbol{\alpha}_{(S)ki}$) in our Hybrid Metropolis within Gibbs sampler.
The Hybrid Metropolis within Gibbs sampler for the multilevel HMM proceeds in a similar fashion as the Gibbs sampler for the HMM: first the hidden state sequences are sampled (for each subject separately), after which the (subject level and group level) parameters are sampled given the observed event sequence (for each subject, O_{kt}), the sampled hidden state sequences (of each subject, S_{kt}), and the current values of the remaining parameters in the model. We provide a stepwise walkthrough of the hybrid Metropolis within Gibbs sampler for the multilevel HMM below.

Stepwise walkthrough of the used hybrid Metropolis within Gibbs sampler

The Hybrid Metropolis within Gibbs sampler used to fit the multilevel HMM proceeds as described below. We use the subscript c to denote the current (i.e., updated using a combination of the value of the hyper-prior and the data) parameters of the conditional posterior distributions.

- Given the observed event sequence for each subject k $O_{k1}, O_{k2}, \ldots, O_{kT}$ and the current values of the parameters $\boldsymbol{\Gamma}$ and $\boldsymbol{\theta}_i$, a hidden state sequence $S_{k1}, S_{k2}, \ldots, S_{kT}$ is sampled for each subject separately, utilizing the forward probabilities. Note that for each subject $k \in \{1, 2, \ldots, K\}$, the subject-specific parameters (i.e., $\boldsymbol{\Gamma}_{ki}$ and $\boldsymbol{\theta}_{ki}$) are used as input for the forward-backward recursions. For the first run of the algorithm, user-specified start values are used for the subject-specific parameters $\boldsymbol{\Gamma}_{ki}$ and $\boldsymbol{\theta}_{ki}$.
- Given the current subject-specific sets of intercepts $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ (related to the subject-specific state-dependent probabilities $\boldsymbol{\theta}_{ki}$ and the subject-specific state transition probability matrix $\boldsymbol{\Gamma}_k$, respectively)

and the observed (time invariant) covariates $\boldsymbol{X}$, new parameter estimates are drawn for the group mean and covariance of the subject-specific sets of intercepts $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ and the fixed regression coefficients $\boldsymbol{\beta}_{(O)i}$ and $\boldsymbol{\beta}_{(S)i}$ from their conditional posterior distributions $\Pr(\bar{\boldsymbol{\alpha}}_{(O)i}, \boldsymbol{\beta}_{(O)i} \mid)$, $\Pr(\Phi_i \mid)$, $\Pr(\bar{\boldsymbol{\alpha}}_{(S)i}, \boldsymbol{\beta}_{(S)i} \mid)$, and $\Pr(\Psi_i \mid)$, respectively.
That is, first the state i specific group variance-covariance matrices Φ_i and Ψ_i (i.e., the covariance between intercepts for the state i and subject k specific intercept vector $\boldsymbol{\alpha}_{(O)ki}$ or $\boldsymbol{\alpha}_{(S)ki}$) are drawn from $\Pr(\Phi_i \mid) \sim \text{IW}(\Phi_{ci}, df_c)$ and $\Pr(\Psi_i \mid) \sim \text{IW}(\Psi_{ci}, df_c)$, where Φ_{ci} and Ψ_{ci} represent a combination of the chosen prior values Φ_0 and Ψ_0 and the state i specific covariance observed over subjects in $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$, respectively, and df_c represent a combination of the chosen prior value df_0 and the number of subjects in the analyzed subject dataset. See Gelman et al. (2014) for details on updating the parameters of an Inverse Wishart distribution.
Next, the state i specific mean group estimates $\bar{\boldsymbol{\alpha}}_{(O)i}$ and $\bar{\boldsymbol{\alpha}}_{(S)i}$ are drawn simultaneously with the state i specific fixed regression coefficient $\boldsymbol{\beta}_{(O)i}$ and $\boldsymbol{\beta}_{(S)i}$ from $\Pr(\bar{\boldsymbol{\alpha}}_{(O)i}, \boldsymbol{\beta}_{\boldsymbol{(O)i}} \mid \ \Phi_i, \boldsymbol{\alpha}_{(S)ki}, \boldsymbol{X}) \sim \text{N}(\boldsymbol{\mu}_{(O)ci}, \frac{1}{K_c}\Phi_i)$ and $\Pr(\bar{\boldsymbol{\alpha}}_{(S)i}, \boldsymbol{\beta}_{\boldsymbol{(S)i}} \mid \ \Psi_i, \boldsymbol{\alpha}_{(S)ki}, \boldsymbol{X}) \sim \text{N}(\boldsymbol{\mu}_{(S)ci}, \frac{1}{K_c}\Psi_i)$, where $\boldsymbol{\mu}_{(O)ci}$ and $\boldsymbol{\mu}_{(S)ci}$ represent a combination of the chosen prior values $\boldsymbol{\alpha}_{(O)0}$ and $\boldsymbol{\beta}_{(O)0}$, and $\boldsymbol{\alpha}_{(S)0}$ and $\boldsymbol{\beta}_{(S)0}$, the observed state i specific mean vector over subjects of the sets of intercepts $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ and the least squares estimators of $\boldsymbol{\beta}_{(O)i}$ and $\boldsymbol{\beta}_{(S)i}$. The parameter K_c represents a combination of the prior value K_0 and the number of subjects in the analyzed dataset.

- Given the observed event sequence for each subject (O_{kt}), the sampled hidden state sequences of each subject (S_{kt}), the group distributions for the subject-specific sets of intercepts $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ parameterized by $\bar{\boldsymbol{\alpha}}_{(O)i}$, $\boldsymbol{\beta}_{(O)i}$ and Φ_i, and $\bar{\boldsymbol{\alpha}}_{(S)i}$, $\boldsymbol{\beta}_{(S)i}$ and Ψ_i, respectively, new estimates of the subject-specific sets of intercepts $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ are drawn from their posterior conditional distribution $\Pr(\boldsymbol{\alpha}_{(O)ki} \mid)$ and $\Pr(\boldsymbol{\alpha}_{(S)ki} \mid)$ using a Random Walk (RW) Metropolis sampler. That is, to draw new estimates for $\boldsymbol{\alpha}_{(O)ki}$, first a candidate vector $\boldsymbol{\alpha}_{(O)ki[candidate]}$ is sampled from a proposal distribution, which we chose to be an asymptotic normal approximation to the conditional posterior distribution (see below). In the RW Metropolis sampler, the vector of means of the proposal distribution is equal to the current estimate of $\boldsymbol{\alpha}_{(O)ki}$, and the scale of the distribution (i.e., here the covariance) has to be specified by the user. To define the scale of the proposal distribution, we followed the method outlined by Rossi, Allenby, and McCulloch (2012), which is described below. In summary, we use a subject-specific scale parameter $\Sigma_{\alpha(O)ki}$, which is a combination between the prior covariance (i.e., the group covariance Φ_i), a covariance matrix that captures the distribution of the data of the subjects (i.e., for $\boldsymbol{\alpha}_{(O)ki}$, this is the covariance in the observed outcomes within state i for subject k), and a scalar s^2. Next, the candidate $\boldsymbol{\alpha}_{(O)ki[candidate]}$ drawn from the proposal distribution $\text{N}(\boldsymbol{\alpha}_{(O)ki}, \Sigma_{\alpha(O)ki})$ is accepted with the probability $min(1, \rho_{\alpha(O)})$, where $\rho_{\alpha(O)}$ is the ratio between the posterior conditional distribution evaluated at the candidate value and the posterior conditional distribution evaluated at the current value:

$$\rho_{\alpha(O)} = \frac{L\big(\boldsymbol{\alpha}_{(O)ki[candidate]} \mid O_{kt}, S_{kt} = i\big) \Pr\big(\boldsymbol{\alpha}_{(O)ki[candidate]} \mid \bar{\boldsymbol{\alpha}}_{(O)i}, \boldsymbol{X}, \boldsymbol{\beta}_{\boldsymbol{(O)i}}, \Phi_i\big)}{L\big(\boldsymbol{\alpha}_{(O)ki[current]} \mid O_{kt}, S_{kt} = i\big) \Pr\big(\boldsymbol{\alpha}_{(O)ki[current]} \mid \bar{\boldsymbol{\alpha}}_{(O)i}, \boldsymbol{X}, \boldsymbol{\beta}_{\boldsymbol{(O)i}}, \Phi_i\big)}. \tag{20}$$

If the candidate $\boldsymbol{\alpha}_{(O)ki[candidate]}$ is accepted, the candidate represents the new estimate for $\boldsymbol{\alpha}_{(O)ki}$. If the candidate is not accepted, the estimate for $\boldsymbol{\alpha}_{(O)ki}$ remains unchanged.
The new estimates for $\boldsymbol{\alpha}_{(S)ki}$ are drawn in a similar fashion: a candidate vector $\boldsymbol{\alpha}_{(S)ki[candidate]}$ is drawn from the proposal distribution $\text{N}(\boldsymbol{\alpha}_{(S)ki}, \Sigma_{\alpha(S)ki})$, and accepted with the probability $\min(1, \rho_{\alpha(S)})$:

$$\rho_{\alpha(S)} = \frac{L\big(\boldsymbol{\alpha}_{(S)ki[candidate]} \mid S_{k,n}, S_{k,n-1} = i\big) \Pr\big(\boldsymbol{\alpha}_{(S)ki[candidate]} \mid \bar{\boldsymbol{\alpha}}_{(S)i}, \boldsymbol{X}, \boldsymbol{\beta}_{\boldsymbol{(S)i}}, \Psi_i\big)}{L\big(\boldsymbol{\alpha}_{(S)ki[current]} \mid S_{k,n}, S_{k,n-1} = i\big) \Pr\big(\boldsymbol{\alpha}_{(S)ki[current]} \mid \bar{\boldsymbol{\alpha}}_{(S)i}, \boldsymbol{X}, \boldsymbol{\beta}_{\boldsymbol{(S)i}}, \Psi_i\big)}. \tag{21}$$

Note that the RW Metropolis sampler is repeated for each subject $k \in \{1, 2, \ldots, K\}$.

These steps are repeated for a large number of iterations, and, after discarding the first iterations as a "burn-in" period, the sampled parameter estimates provide the empirical posterior distribution of the model parameters.
Regarding the acceptance rate of the RW Metropolis sampler for the subject-specific sets of intercepts $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ (i.e., related to the subject-specific state-dependent probabilities $\boldsymbol{\theta}_{ki}$ and the subject-specific state

transition probability matrix $\mathbf{\Gamma}_k$, respectively), an acceptance rate of $\sim 23\%$ is considered optimal when many parameters are being updated at once (Gelman et al. 2014). Within the R package mHMMbayes, the number of accepted draws of a model are stored in emiss_naccept and gamma_naccept for the conditional distributions and the transition probabilities, respectively.

Scaling the proposal distribution of the RW Metropolis sampler

To obtain the scale parameter $\Sigma_{\alpha(O)ki}$ and $\Sigma_{\alpha(S)ki}$ of the proposal distributions of the RW Metropolis sampler for $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$, respectively, we followed the method outlined by Rossi, Allenby, and McCulloch (2012), which has several advantages as discussed below.

The general challenge of the RW Metropolis sampler is that it has to be "tuned" by choosing the scale of the symmetric proposal distribution (e.g., the variance or covariance of a Normal or multivariate Normal proposal distribution, respectively). The scale of the proposal distribution is composed of a covariance matrix Σ, which is then tuned by multiplying it by a scaling factor s^2. Hence we denote the scale of the proposal distribution by $s^2\Sigma$. The scale $s^2\Sigma$ has to be set such that the drawn parameter estimates cover the entire area of the posterior distribution (i.e., the scale Σ should not be set too narrow because then only candidate parameters in close proximity of the current parameter will be drawn), but remains reasonably efficient (i.e., the scale Σ should not be set too wide because then many candidate parameters will be rejected resulting in a slowly progressing chain of drawn parameter estimates).

There are various options for the covariance matrix Σ. Often, the covariance matrix Σ is set such that it resembles the covariance matrix of the actual posterior distribution. To capture the curvature of each subject's conditional posterior distribution, the scale of the RW Metropolis proposal distribution should be customized to each subject. This also facilitates the possibility to let the amount of information available within the data of a subject for a parameter determine to which degree the group level distribution dominates the estimation of the subject-specific parameters. Hence, to approximate the conditional posterior distribution of each subject, the covariance matrix is set to be a combination of the covariance matrix obtained from the subject data and the group level covariance matrix Φ_i or Ψ_i. To estimate the covariance matrix from the subject data, which is only used for the proposal distribution of the RW Metropolis sampler, we simply use a Maximum Likelihood Estimate (MLE), as this quantity is only used for the purpose of scaling the proposal distribution and is not part of the estimated parameter values that constitute the posterior distribution. The MLE estimate of the covariance matrix is obtained by maximizing the likelihood of the Multinomial Logit (MNL) model on the data, and retrieving the Hessian matrix H_{ki} (i.e., the second order partial derivatives of the likelihood function with respect to the parameters). The covariance matrix of the parameters is the inverse of the Hessian matrix, H_{ki}^{-1}. Hence, the covariance matrices for $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$, are defined by $\Sigma_{\boldsymbol{\alpha}_{(O)ki}} = (H_{\alpha(O)ki} + \Phi_i^{-1})^{-1}$ and $\Sigma_{\boldsymbol{\alpha}_{(S)ki}} = (H_{\alpha(S)ki} + \Psi_i^{-1})^{-1}$, respectively. For $\boldsymbol{\alpha}_{(O)ki}$, the data on which the Hessian is obtained is the frequency with which a categorical outcome is observed in state i of subject k. For $\boldsymbol{\alpha}_{(S)ki}$, this data is the frequency with which state i transitions to another state within subject k. Hence, the subject-specific covariance matrix (i.e., the inverse of the Hessian matrix) is based on the sampled hidden state sequence. Therefore, the MLE estimates of the subject-specific covariance matrices that are used for the RW Metropolis proposal distributions have to be obtained in each iteration, as the sampled hidden state sequence changes in each iteration.

A potential problem with maximizing the log-likelihood of each subject's data, is that a certain state might not be sampled for a subject. To circumvent this problem, we modify the subject likelihood function by adding a so-called regularizing likelihood function that has a defined maximum to the subject-level likelihood function. We maximize the resulting pooled likelihood function in order to obtain the MLE estimates. Here, we use the likelihood function of the combined data over all subjects that are considered to be part of one group as the regularizing likelihood function. The pooled likelihood function is scaled by $1 - w \times$ subject-level likelihood $+ w \times$ overall likelihood$^{n.obs_k/N.obs}$, so that the overall likelihood function does not dominate the subject-level likelihood function, and where $n.obs_k$ is the number of data observations for subject k and $N.obs$ is the total number of data observations over all subjects in a group.

Now that we defined the covariance matrix Σ for the scale of the RW Metropolis sampler proposal distribution, we have to define the scalar factor s^2 to obtain the scale $s^2\Sigma$ of the proposal distribution. As in Rossi, Allenby, and McCulloch (2012)}, we adopt the scaling proposal of Roberts, Rosenthal, and others (2001), and set scaling to $s^2 = (2.93/\sqrt{\text{n.param}})^2$, where $n.param$ is the number of parameters to be estimated for $\boldsymbol{\alpha}_{(S)ki}$ or

$\boldsymbol{\alpha}_{(O)ki}$ in the RW Metropolis sampler, which equals $m-1$ in case of $\boldsymbol{\alpha}_{(S)ki}$ (where m denotes the number of states) and $q-1$ in case of $\boldsymbol{\alpha}_{(O)ki}$ (where q denotes the number of categorical outcomes).
In summary, the scale parameter $s^2\Sigma_{\alpha(O)ki}$ and $s^2\Sigma_{\alpha(S)ki}$ of the proposal distributions of the RW Metropolis sampler for $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ are defined as:

$$s^2\Sigma_{\alpha(O)ki} = (2.93/\sqrt{q-1})^2 \times (H_{\alpha(O)ki} + \Phi_i^{-1})^{-1}, \text{ and} \tag{22}$$

$$s^2\Sigma_{\alpha(S)ki} = (2.93/\sqrt{m-1})^2 \times (H_{\alpha(S)ki} + \Psi_i^{-1})^{-1}, \tag{23}$$

where $H_{\alpha(O)ki}$ is the Hessian of the k^{th} subject's data of the frequency with which a categorical outcome is observed within state i evaluated at the MLE of the pooled likelihood, $H_{\alpha(S)ki}$ is the Hessian of the k^{th} subject's data of the frequency with which state i transitions to another state evaluated at the MLE of the pooled likelihood, and Φ_i^{-1} and Ψ_i^{-1} are the inverses of the group level covariance matrices. This provides us with m pairs of scale parameters that closely resemble the scale of the subject-level conditional posterior distribution, and that 1) are automatically tuned (i.e., we do not require experimentation to determine s^2 to tune the covariance matrix), 2) allow the amount of information available within the data of a specific subject to determine the degree to which the group level distribution dominates the estimation of that subject's level parameters, and 3) do not require each state to be sampled in the hidden state sequence as not each subject-level likelihood is required to have a maximum.

Full conditional posterior distributions of the multilevel HMM

In the hybrid Metropolis within Gibbs sampler, all level 2 model parameters are directly sampled from their full conditional posterior distributions. The full conditional posterior distributions are obtained by applying Bayes theorem, combining the (hyper-)prior distribution of the model parameter and the likelihood function. Direct sampling from the conditional posterior distributions for these model parameters is possible, as the choice of the (hyper-)prior distribution results in a closed form expression of the full conditional posterior distribution. That is:

- The full conditional posterior distributions of Φ_i and Ψ_i (i.e., the state i specific group covariance between the subject k batches of the $q-1$ random intercepts $\boldsymbol{\alpha}_{(O)ki}$ pertaining to the subject-specific state-dependent probabilities $\boldsymbol{\theta}_{ki}$, and the state i specific group covariance between the subject k batches of the $m-1$ random intercepts $\boldsymbol{\alpha}_{(S)ki}$ pertaining to the subject-specific state transition probabilities, $\boldsymbol{\Gamma}_k$) are:

$$\begin{aligned}
\Pr(\Phi_i \mid) &\sim \text{IW}(\Phi_{ci}, df_c) \\
\Phi_{ci} &= \Phi_0 + (\boldsymbol{\alpha}_{(O)ki} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{(O)ci}})^\top(\boldsymbol{\alpha}_{(O)ki} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{(O)ci}}) + \\
&\quad \Big(\boldsymbol{\mu}_{\boldsymbol{(O)ci}} - \begin{bmatrix}\boldsymbol{\alpha}_{(O)0}\\ \boldsymbol{\beta}_{(O)0}\end{bmatrix}\Big)^\top \boldsymbol{K_0}\Big(\boldsymbol{\mu}_{\boldsymbol{(O)ci}} - \begin{bmatrix}\boldsymbol{\alpha}_{(O)0}\\ \boldsymbol{\beta}_{(O)0}\end{bmatrix}\Big) \\
\boldsymbol{\mu}_{\boldsymbol{(O)ci}} &= (\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{K_0})^{-1}\Big(X^\top\boldsymbol{\alpha}_{(O)ki} + \boldsymbol{K_0}\begin{bmatrix}\boldsymbol{\alpha}_{(O)0}\\ \boldsymbol{\beta}_{(O)0}\end{bmatrix}\Big) \\
df_c &= df_0 + K
\end{aligned} \tag{24}$$

$$\begin{aligned}
\Pr(\Psi_i \mid) &\sim \text{IW}(\Psi_{ci}, df_c) \\
\Psi_{ci} &= \Psi_0 + (\boldsymbol{\alpha}_{(S)ki} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{(S)ci}})^\top(\boldsymbol{\alpha}_{(S)ki} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{(S)ci}}) + \\
&\quad \Big(\boldsymbol{\mu}_{\boldsymbol{(S)ci}} - \begin{bmatrix}\boldsymbol{\alpha}_{(S)0}\\ \boldsymbol{\beta}_{(S)0}\end{bmatrix}\Big)^\top \boldsymbol{K_0}\Big(\boldsymbol{\mu}_{\boldsymbol{(S)ci}} - \begin{bmatrix}\boldsymbol{\alpha}_{(S)0}\\ \boldsymbol{\beta}_{(S)0}\end{bmatrix}\Big) \\
\boldsymbol{\mu}_{\boldsymbol{(S)ci}} &= (\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{K_0})^{-1}\Big(X^\top\boldsymbol{\alpha}_{(S)ki} + \boldsymbol{K_0}\begin{bmatrix}\boldsymbol{\alpha}_{(S)0}\\ \boldsymbol{\beta}_{(S)0}\end{bmatrix}\Big) \\
df_c &= df_0 + K
\end{aligned} \tag{25}$$

where x^T is the transpose of x, $\boldsymbol{\alpha}_{(O)0}$ and $\boldsymbol{\alpha}_{(S)0}$ denote a vector of chosen mean values of the Normal hyper-prior distribution on the group mean vector $\bar{\boldsymbol{\alpha}}_{(O)i}$ and $\bar{\boldsymbol{\alpha}}_{(S)i}$, respectively, $\boldsymbol{\beta}_{(O)0}$ and $\boldsymbol{\beta}_{(S)0}$ denote

a vector of chosen values of the Normal hyper-prior distribution on fixed regression parameters $\boldsymbol{\beta}_{(O)i}$ and $\boldsymbol{\beta}_{(S)i}$, respectively, $\boldsymbol{K_0}$ denotes a diagonal matrix with on the diagonal the number of observations (i.e., the number of hypothetical prior subjects) on which the prior values $\bar{\boldsymbol{\alpha}}_{(O)i}$, $\bar{\boldsymbol{\alpha}}_{(S)0}$, $\boldsymbol{\beta}_{(O)i}$, and $\boldsymbol{\beta}_{(S)i}$ are based, K denotes the total number of subjects in the dataset, Φ_0 and Ψ_0 denote the chosen prior covariance values of the Inverse Wishart hyper-prior distribution on the group covariance Φ_i and Ψ_i, respectively, and df_0 denotes the prior specified degrees of freedom of the Inverse Wishart hyper-prior distribution on the group covariance Φ_i and Ψ_i.

- The full conditional posterior distributions of $\bar{\boldsymbol{\alpha}}_{(O)i}$ and $\boldsymbol{\beta}_{(O)i}$, and $\bar{\boldsymbol{\alpha}}_{(S)i}$ and $\boldsymbol{\beta}_{(S)i}$ (i.e., the state i specific group mean vector of the subject-specific batches of intercepts $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ pertaining to the state-dependent probabilities and state transition probabilities, respectively, and the fixed regression coefficients predicting the subject-specific batches of intercepts $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$) are:

$$\begin{aligned} \Pr\Big(\begin{bmatrix} \bar{\boldsymbol{\alpha}}_{(O)i} \\ \boldsymbol{\beta}_{(O)i} \end{bmatrix} \mid \ \Big) &\sim \mathrm{N}\Big(\boldsymbol{\mu_{(O)ci}}, \tfrac{1}{K_c}\Phi_i\Big) \\ \boldsymbol{\mu_{(O)ci}} &= \Big(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{K_0}\Big)^{-1}\Big(X^\top \boldsymbol{\alpha}_{(O)ki} + \boldsymbol{K_0}\begin{bmatrix} \boldsymbol{\alpha}_{(O)0} \\ \boldsymbol{\beta}_{(O)0} \end{bmatrix}\Big) \\ K_c &= K + K_0 \end{aligned} \tag{26}$$

$$\begin{aligned} \Pr\Big(\begin{bmatrix} \bar{\boldsymbol{\alpha}}_{(S)i} \\ \boldsymbol{\beta}_{(S)i} \end{bmatrix} \mid \ \Big) &\sim \mathrm{N}\Big(\boldsymbol{\mu_{(S)ci}}, \tfrac{1}{K_c}\Psi_i\Big) \\ \boldsymbol{\mu_{(S)ci}} &= \Big(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{K_0}\Big)^{-1}\Big(X^\top \boldsymbol{\alpha}_{(S)ki} + \boldsymbol{K_0}\begin{bmatrix} \boldsymbol{\alpha}_{(S)0} \\ \boldsymbol{\beta}_{(S)0} \end{bmatrix}\Big) \\ K_c &= K + K_0 \end{aligned} \tag{27}$$

where $\boldsymbol{\alpha}_{(O)0}$ and $\boldsymbol{\alpha}_{(S)0}$ denote the chosen mean values of the Normal hyper-prior distribution on the group mean vector $\bar{\boldsymbol{\alpha}}_{(O)i}$ and $\bar{\boldsymbol{\alpha}}_{(S)i}$, respectively, $\boldsymbol{\beta}_{(O)0}$ and $\boldsymbol{\beta}_{(S)0}$ denote a vector of chosen values of the Normal hyper-prior distribution on fixed regression parameters $\boldsymbol{\beta}_{(O)i}$ and $\boldsymbol{\beta}_{(S)i}$, K_0 denotes the number of observations (i.e., the number of hypothetical prior subjects) on which the prior values $\boldsymbol{\alpha}_{(O)0}$, $\boldsymbol{\alpha}_{(S)0}$, $\boldsymbol{\beta}_{(O)0}$, and $\boldsymbol{\beta}_{(S)0}$ are based, and K denotes the total number of subjects in the dataset.

For the random intercepts $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$, related to the subject-specific state-dependent probabilities of observing a categorical outcome $\boldsymbol{\theta}_{ki}$ and the subject-specific state transition probability matrix $\boldsymbol{\Gamma}_k$, respectively, the choice of prior distributions does not result in closed form expressions of the full conditional posterior distributions. That is, for the subject-specific sets of intercepts $\boldsymbol{\alpha}_{(O)ki}$ related to the subject-specific state-dependent probabilities of observing a categorical outcome within state i, the full conditional posterior distribution when we assess a standard multivariate normal prior is:

$$\begin{aligned} Pr(\boldsymbol{\alpha}_{(O)ki} \mid) &\propto L\big(\boldsymbol{\alpha}_{(O)ki} \mid O_{kt}, S_{kt} = i\big) \Pr\big(\boldsymbol{\alpha}_{(O)ki} \mid \bar{\boldsymbol{\alpha}}_{(O)i}, \boldsymbol{\beta_{(O)i}}, \Phi_i\big), \\ \Pr\big(\boldsymbol{\alpha}_{(O)ki} \mid \bar{\boldsymbol{\alpha}}_{(O)i}, \boldsymbol{\beta_{(O)i}}, \Phi_i\big) &\sim \mathrm{N}(\bar{\boldsymbol{\alpha}}_{(O)i} + \boldsymbol{X}^\top \boldsymbol{\beta_{(O)i}}, \Phi_i)), \end{aligned} \tag{28}$$

and the likelihood is the product of the probabilities of the observed outcomes $O_{kt} = l \in \{1, 2, \ldots, q\}$ within sampled states $S = i$ in subject k over time points t:

$$\begin{aligned} L\big(\boldsymbol{\alpha}_{(O)ki} \mid O_{kt}, S_{kt} = i\big) &= \prod_t Pr(O_{k,t} = l \mid S_{kt} = i, \boldsymbol{\alpha}_{(O)ki}), \\ Pr(O_{k,t} = l \mid S_{kt} = i, \boldsymbol{\alpha}_{(O)ki}) &= \frac{\exp(\alpha_{(O)kil})}{1 + \sum_{\bar{l}=2}^{q} \exp(\alpha_{(O)ki\bar{l}})}, \end{aligned} \tag{29}$$

where the product is restricted to the set of time points that coincide with the sampled state S for subject k at time point t being i, and q is the number of categorical outcomes. The numerator is set equal to one for $l = 1$, making the first categorical outcome in the set the baseline category in every state.
For the subject-specific sets of intercepts $\boldsymbol{\alpha}_{(S)ki}$ related to the state-transition probabilities to transition from

state i to any of the other states $j \in \{1, 2, \ldots, m\}$, the full conditional posterior distribution when we assess a standard multivariate normal prior is:

$$\begin{aligned} Pr(\boldsymbol{\alpha}_{(S)ki} \mid) &\propto L\big(\boldsymbol{\alpha}_{(S)ki} \mid S_{kt}, S_{k(t-1)} = i\big) \Pr\big(\boldsymbol{\alpha}_{(S)ki} \mid \bar{\boldsymbol{\alpha}}_{(S)i}, \boldsymbol{\beta_{(S)i}}, \Psi_i\big), \\ \Pr\big(\boldsymbol{\alpha}_{(S)ki} \mid \bar{\boldsymbol{\alpha}}_{(S)i}, \boldsymbol{\beta_{(S)i}}, \Psi_i\big) &\sim \mathrm{N}(\bar{\boldsymbol{\alpha}}_{(S)i} + \boldsymbol{X}^\top \boldsymbol{\beta_{(S)i}}, \Psi_i), \end{aligned} \tag{30}$$

and the likelihood is the product of the probabilities of the observed transitions from state i in the previous time point $t-1$ to any of the other states $S_{kt} = j$ over time points t in subject k:

$$\begin{aligned} L\big(\boldsymbol{\alpha}_{(S)ki} \mid S_{kt}, S_{k(t-1)} = i\big) &= \prod_n Pr(S_{k,t} = j \mid S_{k(t-1)} = i, \boldsymbol{\alpha}_{(S)ki}), \\ Pr(S_{k,t} = j \mid S_{k(t-1)} = i, \boldsymbol{\alpha}_{(S)ki}) &= \frac{\exp(\alpha_{(S)kij})}{1 + \sum_{\bar{j} \in Z} \exp(\alpha_{(S)ki\bar{j}})}, \\ &\text{where } Z \in \{1, 2, \ldots, m,\ Z \neq 1\} \end{aligned} \tag{31}$$

where the product is restricted to the set of time points that coincide with the sampled state S in the previous time point $t-1$ being i for subject k, and m is the number of states. The numerator is set equal to 1 for $j = 1$, making the first state of every row of the transition probability matrix $\boldsymbol{\Gamma}_k$ the baseline category.
As the conditional posterior distributions for $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ do not result in a closed form expression of a know distribution, we cannot directly sample values of $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ from their conditional posterior distributions with pre-defined equations on how to obtain the current (i.e., updated using a combination of the value of the hyper-prior and the data) parameters of the conditional posterior distributions. Instead, new values for $\boldsymbol{\alpha}_{(O)ki}$ and $\boldsymbol{\alpha}_{(S)ki}$ are sampled using a RW Metropolis sampler, as described above.

References

Altman, Rachel MacKay. 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). Taylor & Francis: 201–10.

Baum, Leonard E, Ted Petrie, George Soules, and Norman Weiss. 1970. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *The Annals of Mathematical Statistics*. JSTOR, 164–71.

Cappé, E. AND Rydén, O. AND Moulines. 2005. *Inference in Hidden Markov Models*. New York: Springer.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin. 1977. "Maximum Likelihood from Incomplete Data via the Em Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 1–38.

Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin. 2014. *Bayesian Data Analysis*. Vol. 2. London: Taylor & Francis.

Goldstein, Harvey. 2011. *Multilevel Statistical Models*. 4th ed. West Sussex: John Wiley & Sons.

Hox, Joop J, Mirjam Moerbeek, and Rens van de Schoot. 2017. *Multilevel Analysis: Techniques and Applications. 3rd Edn*. New York, NY: Routledge.

Jasra, Ajay, CC Holmes, and DA Stephens. 2005. "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling." *Statistical Science*. JSTOR, 50–67.

Lynch, Scott M. 2007. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer Science & Business Media.

Rabiner, Lawrence R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77 (2). IEEE: 257–86.

Roberts, Gareth O, Jeffrey S Rosenthal, and others. 2001. "Optimal Scaling for Various Metropolis-Hastings Algorithms." *Statistical Science* 16 (4). Institute of Mathematical Statistics: 351–67.

Rossi, Peter E, Greg M Allenby, and Rob McCulloch. 2012. *Bayesian Statistics and Marketing*. West Sussex: John Wiley & Sons.

Rydén, Tobias. 2008. "EM Versus Markov Chain Monte Carlo for Estimation of Hidden Markov Models: A Computational Perspective." *Bayesian Analysis* 3 (4). International Society for Bayesian Analysis: 659–88.

Scott, Steven L. 2002. "Bayesian Methods for Hidden Markov Models." *Journal of the American Statistical Association* 97 (457).

Snijders, Tom, and Roel Bosker. 2011. *Multilevel Analysis: An Introduction to Basic and Applied Multilevel Analysis*. London: Sage.

Zucchini, Walter, Iain L MacDonald, and Roland Langrock. 2016. *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton: CRC Press.