Examining the effect of observation sequence variables on hidden Markov model Gibbs sampler inference

J.G. Simons
Supervised by dr. E. Aarts
28 June 2019

Department of Methodology and Statistics, Utrecht University

Abstract

In the preceding two decades, Hidden Markov models have become the method of choice for obtaining novel information from intensive longitudinal data sequences. One of the fundamental problems in hidden Markov modelling pertains to retrieving the structure of a hidden process phenomenon. Retrieving such structures enables researchers to formulate models which best describe unobserved real-world process phenomena. These types of learning problems are typically adressed with the Gibbs sampler. Methodological guidelines on fitting and optimal input specifications for the Gibbs sampler are however sparse. This study seeks to identify the general and optimal relations between the Gibbs sampler and two of its input variables: the number of event types inherent to, and the length of, the event observation sequence. In doing so it seeks to establish specification references for the appropriate and optimal use of the Gibbs sampler in single sequence HMM learning. Results indicate four event types and a sequence input length of 8000 to result in superior Gibbs sampler estimates. This study's conclusions consequently coincide with, and add to the extant literature. Future research avenues in regards to extending current work, and incorporating additional observation variabes are discussed.

Introduction

Over the course of the preceding decades, hidden Markov models (HMMs) have become the method of choice for describing and explicating latent process dynamics (Eddy, 1996; Eddy, 1998; Rabiner & Juang, 1986; Rabiner, 1989). By examining the temporal associations and transitions between each successive observation and its associated underlying hidden state, Markovian modelling techniques enable researchers to extract novel information from intensive longitudinal data (ILD) series (Aarts, 2016; Rabiner, 1989). Conventional statistical models are typically ill-suited to adequately address such time-series data, in that the information contained within the ILD is severely abridged or even completely discarded in their application (Aarts, 2016). Vis-à-vis such conventional models, and due to the flexibility of their mathematical structure, HMMs present a universal and actionable method for ILD analysis, facilitating the reformulation and extension of scientific theory in a wide variety of research areas. These types of models have been utilized to study speech and handwriting recognition (Chen, Kundu, & Zhou, 1994; Rabiner, 1989), human action and shape classification (He, & Kundu, 1991; Yamato, Ohya, & Ishii, 1992), computational molecular biology (Eddy, 2011; Fine, Singer & Tishby, 1998), brain MR image segmentation (Zhang, Brady, & Smith, 2001), precipitation occurence (Hughes, Guttorp, Charles, 1999) and stock market forecasting (Hassan, & Nath, 2005).

Due to increases in the general availability and quality of ILD, coupled with advances in computing power and estimation procedures, HMMs have become an increasingly relevant and accessible modelling strategy (Visser, 2011; Gagniuc, 2017). The expanding significance of the HMM in the contemporary analysis of ILD necessitates the formulation of a set of guidelines to aid current and future researchers in its correct and efficient application. One of the fundamental problems in hidden Markov modelling relates to the identification of the underlying latent structure of the process phenomenon of interest (Rabiner, 1989). Uncovering this structure is central to most HMM applications, because it enables researchers to optimally adapt the parameters of the HMM to ILD sequences, i.e., formulate HMMs which best represent real-world process phenomena (Visser, 2011; Rabiner, 1989). However, since the latent configuration of the system is hidden and therefore apriori unknown, its optimal configuration needs to be approximated with the use of Markov Chain Monte Carlo (MCMC) sampling algorithms (Rydén, 2008; Scott, 2002). In short, MCMC sampling algorithms enable researchers to calculate numerical approximations of multi-dimensional integrals (Lynch, 2007). These types of methods approximate the distribution of a particular parameter by sampling from a multi-dimensional random variable, resulting in an ensemble distribution from which summary statistics such as the mean and variance can consequently be extracted (Lynch, 2007).

For this particular investigation, the Gibbs sampler MCMC algorithm will be utilized to approximate the latent structure of the HMM, on account of it being the contemporary MCMC algorithm of choice for single sequence HMM learning (Rydén, 2008; Scott, 2002).

Ideally, the Gibbs samplers approximation of the latent structure of the HMM will constitute a one-to-one representation of the true parameter values of the latent system. Accurate Gibbs sampling estimator values are however dependent on a diverse set of input variables, such as the length of the ILD sequence, the number of event types relative to the number of states, and the shape of the probability distributions of the transition and conditional probability matrices (Rabiner, 1989; Rydén, 2008; Scott, 2002; Chudova & Smith, 2002). This input specific accuracy variability neccesitates the existence of a reference set which delineates to researchers how the Gibbs samplers input values can be adapted to provide an optimal and efficient description of the system of interest. Such reference documentation is however generally lacking in the extant literature on HMM Gibbs sampling input specification (Brooks, Gelman, Jones, & Meng, 2011; Cappé, Moulines, & Rydén, 2009). As such, the central objective of this study pertains to adressing and ameliorating this knowledge gap. It seeks to do so by establishing references on the degree to which a subset of the total set of relevant input variables affect the Gibbs samplers capacity to produce accurate HMM parameter estimates. It additionally seeks to identify the single and combined value input ranges for which these variables enable the Gibbs sampler to produce optimal system approximations. Catalogueing such specification references will consequently allow for the tentative formulation of a set of guidelines on the topic of if and when the Gibbs samplers estimates can be expected to accurately reflect the structure of the system of interest. This inquiry will proceed by providing the reader with a stagewise description of the HMM, and the Gibbs sampler MCMC algorithm. Consequently, based on the extant literature on Gibbs sampler input specification, the variables of interest and their hypothesized relations to the functioning of the Gibbs sampler are discussed.

Model description

The Markov chain

The hidden Markov model is an augmented version of the Markov chain (MC) model (Jurafsky & Martin, 2014). The MC is a discrete-time stochastic model describing a system which at any moment in time is in one of a set of N events $E = \{E_1, E_2, ..., E_n\}$ (Rabiner, 1989). At regularly spaced, discrete time points, the system transitions from one event to another based on a probability set that is exclusively determined by the current event (Jurafsky & Martin, 2014; Rabiner, 1989). The time points associated with the event transitions are denoted by t = 1, 2, . . . , T, with the current event at time t having the denotation E_t (Rabiner, 1989). The assumption that the probability for each consecutive event is dependent solely on the event that precedes it is the central premise of the Markovian modelling framework (Gagniuc, 2017; Jurafsky & Martin, 2014). This so-called Markov property defines the MC model to be memoryless; conditional on the present event displayed by the system, its future and past events are independent (Gagniuc, 2017; Jurafsky & Martin, 2014).

In formal terms, the MC model provides a description of a sequence of time-ordered events $\{E_t : t = 1, 2, ..., T\}$, where the values for each event E_t originate from the set of countable natural numbers $\mathbb{N}$, i.e., $E_t \in \{1, 2, ..., \mathbb{N}\}$ (Aarts, 2016). The Markov property dependency assumption is defined by the argument

$$P(E_{t+1} \mid E_t, E_{t+1}....E_1) = P(E_{t+1} \mid E_t) \qquad (1)$$

where the probability of switching to the next event E_{t+1} is exclusively dependent on the state of the current event E_t (Aarts, 2016). See figure 1 below for a graphical depiction of transitions between events conform the Markov property.

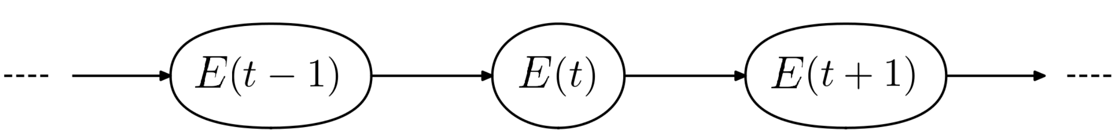


Figure 1: Directed graph illustrating memoryless transitions between events.

Transitions between events in the sequence are represented by a transition probability matrix Γ, in which the element γ_{ij} denotes the probability of transitioning from event i at time t to event j at time t+1 (Rabiner, 1989):

$$\gamma_{ij} = P(E_{t+1} = j \mid E_t = i) \text{ with } \sum_{j}^{n} \gamma_{ij} = 1 \quad \forall i. \qquad (2)$$

The transition probability matrix contains the complete set of probabilities to transition from event i to event j, with $j \in \{1, 2, ..., \mathbb{N}\}$, including the self-transition probability i to i (Rabiner, 1989). Note that Γ is a right-stochastic matrix, i.e., the elements of each row sum to 1 (Gagniuc, 2017).

Also note that the transition probabilities γ_{ij} are assumed to be time-homogeneous, i.e., remain constant over the duration of the finite span of the event sequence (Rabiner, 1989). See figure 2 below for a graphical representation of the transition probability matrix.

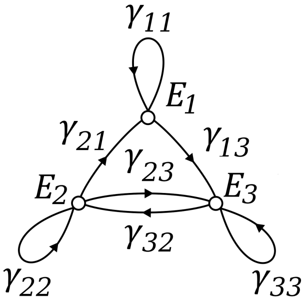


Figure 2: Transition probability diagram for a system with an event set of three. Note that in this particular example, the transition probabilities γ_{12} and γ_{31} equal zero.

The second and final component of the MC is its initial probability distribution - given as a stochastic row vector $\pi = (\pi_1, \pi_2, ..., \pi_i)$ - which represents where the event sequence might start out initially, and with what probabilities (Jurafsky & Martin, 2014). Some events j may have $\pi_j = 0$, meaning that they cannot be initial sequence events (Jurafsky & Martin, 2014). Furthermore, like the transition probability matrix, the elements of the row vector π_i sum to 1, i.e., $\sum_{i=1}^{n} \pi_i = 1$ (Jurafsky & Martin, 2014).

To set ideas, two key applications of the MC model will summarily be discussed. Consider a Markov chain model describing a generic process phenomenon with event set $E = \{E_1, E_2, E_3\}$, for a sequence t = 1, 2, ..., 8 of observed time-ordered events O $= \{E_2, E_3, E_1, E_3, E_1, E_3, E_1, E_2\}$. Assume that the phenomenon is adequately described by the event set at any point t, that the transition probability matrix Γ is of the form

$$\Gamma = \gamma_{ij} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

and that the initial probability distribution row vector equals $\pi = \begin{bmatrix} 0.1 & 0.7 & 0.2 \end{bmatrix}$. Given the specified MC model λ, what is the probability of the observed event sequence O? Analogous to Rabiner (1989), writing

$$\begin{aligned} P(O \mid \lambda) &= [P(E_2, E_3, E_1, E_3, E_1, E_3, E_1, E_2 \mid \lambda] \\ &= P[E_2] \cdot P[E_3 \mid E_2] \cdot P[E_1 \mid E_3] \cdot P[E_3 \mid E_1] \cdot \\ &\quad P[E_1 \mid E_3] \cdot P[E_3 \mid E_1] \cdot P[E_1 \mid E_3] \cdot P[E_2 \mid E_1] \\ &= \pi_2 \cdot \gamma_{23} \cdot \gamma_{31} \cdot \gamma_{13} \cdot \gamma_{31} \cdot \gamma_{13} \cdot \gamma_{31} \cdot \gamma_{12} \\ &= (0.7)(0.2)(0.1)(0.3)(0.1)(0.3)(0.1)(0.3) \\ &= 3.78 \cdot 10^{-6} \end{aligned}$$

results in a probability of $3.78 \cdot 10^{-6}$ for the sequence occurence O given the MC model.

A second query one can address using the MC model concerns that circumstance wherein the model is in a known event, and one seeks to determine the probability that the model will remain in that event for an exact amount of t time points. Analogous to Rabiner (1989), the probability of the observed event sequence

$$O = \{\underset{1}{E_i}, \underset{2}{E_i}, \underset{3}{E_i}, ..., \underset{t}{E_i}, \underset{t+1}{E_j} \neq E_i\}$$

given the MC model λ equals

$$P(O \mid \lambda, E_1 = E_i) = (\gamma_{ii^{t-1}})(1 - \gamma ii) = p_i t \qquad (3)$$

where $p_i t$ is the probability density function of duration t in event i. Based on $p_i t$, the expected number of subsequent observations for a certain event - conditional on it being the initial event in the observed event sequence - is provided by the argument:

$$\bar{t_i} = \sum_{t=1}^{\infty} t p_i(t) = \sum_{t=1}^{\infty} t(\gamma_{ii})^{t-1}(1 - \gamma_{ii}) = \frac{1}{1 - \gamma_{ii}}. \qquad (4)$$

Given the MC model and the self-transition probabilities contained in the transition probability matrix Γ that was defined on the previous page, the number of expected subsequent observations would equal 1 / 0.6 = 1.67 for event E_1, 1 / 0.4 = 2.5 for event E_2 and 1 / 0.2 = 5 for event E_3.

The hidden Markov model

Although the MC model constitutes an effectual method for determining observed event sequence probabilities, it is unable to properly adress those types of process phenomena where the event sequence of interest is not observable (Jurafsky & Martin, 2014). The hidden Markov model (HMM) extends the concept of the Markov chain so that the likelihood of occurence for an event is a probabilistic function of an unobserved underlying state (Rabiner, 1989). The HMM is a doubly embedded stochastic process, with an underlying stochastic process that is not observable, but which can be observed through another set of stochastic processes that produce the observed event sequence (Rabiner, 1989). HMMs thus enable researchers to analyze both observed and hidden underlying events that are thought of as causal factors in the probabilistic model (Jurafsky & Martin, 2014).

In order to explicate the HMM framework to the reader, Visser's (2011) topology of the three fundamental characteristics of the discrete time HMM is presented and expounded on. The first of Visser's (2011) defining characteristics relates to the fact that the marginal distribution of the intensive longitudinal data (ILD) sequence has a mixture distribution, i.e., the collapsed set of event observations are drawn from two or more distributions with different parameter values.

In formal terms, consider an observed event sequence $\{E_t : t = 1, 2, ..., T\}$, which can have either a discrete or a continuous distribution, and an associated underlying HMM consisting of the state variables $\{S_t : t = 1, 2, ..., T\}$. Throughout this particular consideration, the state variables are defined to be discrete; they are elements of a finite set $s = \{1, 2, ..., m\}$, so that $S_t = i, i \in s$. The set s is the finite state space of the HMM - the set of all possible configurations of a system - while m represents the total number of states in the model. Since the HMM considered here is a discrete time model, i.e., for each time point t, there exists at most a single hidden state that can function to generate an observable event, the probability of observing the current event E_t is exclusively determined by the current latent state S_t (Rabiner, 1989):

$$P(E_t \mid E_{t-1}, E_{t-2}, ..., E_1, S_t, S_{t-1}, ..., S_1) = P(E_t \mid S_t). \qquad (5)$$

Put differently, the observations E_t are dependent on the state variables S_t such that the distribution of E_t can be written as $f_i(E_t) := f(E_t \mid S_t = i)$ (Visser, 2011). Because the set s is finite, the marginal distribution of the ILD is a mixture distribution with m states

$$f(E_t) = \sum_{i=1}^{m} p_i f_i(E_t) \qquad (6)$$

where p_i are the state proportions with the constraint that $\sum_{i=1}^{m} p_i = 1$, $p_i \geq 0$ and $f_i(\cdot)$ is the conditional distribution of the data in state i (Visser, 2011). In sum, HMMs are characterized by discrete, hidden states, which can be interpreted as states in a process that generate typical observations for that particular moment in time (Visser, 2011). An intuitive example of this characteristic is provided by the analysis of sleep stages with use of the HMM framework (Flexerand, Dorffner, Sykacekand, & Rezek, 2002). Although sleep stages such as REM sleep, deep sleep and wakefulness are not directly observable, they each generate characteristic continuous EEG measurements - which, for the sake of argument, are subsequently discretizised - that communicate their relative presence to researchers (Flexerand et al., 2002).

The second defining characteristic of the HMM relates to the temporal associations and transitions between the underlying states, which like in the MC model, conform to the Markov property dependency assumption, but now for states instead of events. Write

$$P(S_{t+1} \mid S_t, S_{t+1}, ..., S_1) = P(S_{t+1} \mid S_t) \qquad (7)$$

so that the probability of transitioning to the next state S_{t+1} depends only on the current state S_t. See figure 3 on the next page for a graphical representation of the temporal evolution of the HMM. Note that each state produces a single unique observation, drawn from a distribution distinct to the active state, as per the first defining characteristic, and that the probability for each consecutive state is exclusively determined by the state that precedes it, as per the second.

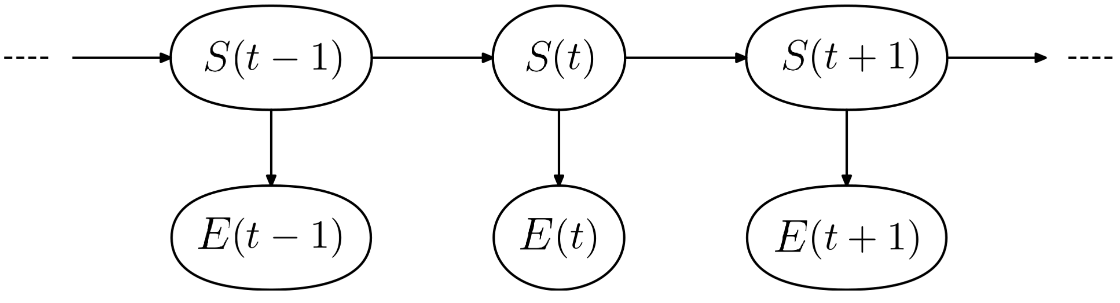


Figure 3: Temporal evolution of the hidden Markov model

Based on the definition of the HMM so far, the model can be said to contain three sets of parameters: the initial probabilities of the states π_i, the transition probability matrix Γ, and the state-dependent probability distribution of observing E_t given S_t with parameter set θ_i. The first two of these parameters sets, namely π and Γ, are analogous to the MC model, but - as stated earlier - now relate to the hidden states as opposed to the observed events. The initial probability distribution is given by stochastic row vector $\pi = (\pi_1, \pi_2, ..., \pi_i)$, and represents where the state sequence might start out initially, and with what probabilities (Jurafsky & Martin, 2014). Some states j may have $\pi_j = 0$, meaning that they cannot be initial sequence states (Jurafsky & Martin, 2014). Furthermore, the elements of the row vector π_i sum to 1, i.e., $\sum_{i=1}^{n} \pi_1 = 1$ (Jurafsky & Martin, 2014). Alternatively, denote the probability that the first state in the state sequence, S_i, equals i with (Rabiner, 1989):

$$\pi_i = P(S_1 = i) \quad \text{with} \quad \pi_i = 1. \tag{8}$$

The transition probability matrix Γ with transition probabilities γ_{ij} denotes the probability of transitioning from state i at time t to state j at time $t+1$ (Rabiner, 1989):

$$\gamma_{ij} = P(S_{t+1} = j \mid S_t = i) \quad \text{with} \quad \sum_{j}^{n} \gamma_{ij} = 1 \quad \forall i. \tag{9}$$

The transition probability matrix contains the complete set of probabilities to transition from state i to state j, with $j \in s, s \in \{1, 2, ..., m\}$, including the self-transition probability i to i (Rabiner, 1989). Note that Γ is a right-stochastic matrix, i.e., the elements of each row sum to 1 (Gagniuc, 2017). Also note that the transition probabilities γ_{ij} are assumed to be time-homogeneous, i.e., remain constant over the duration of the finite span of the state sequence (Gagniuc, 2017). The third parameter set is the state-dependent probability distribution, which denotes the probability of observing E_t given S_t with parameter set θ_i (Jurafsky & Martin, 2014). In this particular consideration, the state-dependent probability distribution is given by a categorical distribution, where the parameter set θ_i is the set of state-dependent probabilities of observing certain events.

This consequently translates into

$$P(E_t = E \mid S_t = i) \sim \mathrm{Cat}(\theta_i) \qquad (10)$$

for the observed event outcome set $E = \{1, 2, ..., n\}$, where $\theta_i = (\theta_1, \theta_2, ..., \theta_n)$ is a vector of probabilities for each state $S = \{i, ..., m\}$ with $\theta_i = 1$, i.e., the set of possible outcome probabilities sum to 1 within each state vector (Aarts, 2016; Gagniuc, 2017).

The third and final characteristic of the HMM relates to the fact that the underlying discrete states S_t are hidden. The distribution function $f(E_t \mid S_t = i)$ or $f_i(Y_t)$ is not a deterministic function but a probability density function (Visser, 2011). It would otherwise simply present a mapping of the states S_t into the event observations E_t, and reduce from an HMM to an MC model since S_t is now observed (Visser, 2011). Relating this third characteristic to the sleep stage example, a probabilistic relationship is defined to exist between the set of discretizised EEG readings and the set of discrete sleep stages. Although some EEG readings might represent uncharacteristic output for one particular sleep stage, while being more typical for the other, all states have a defined chance of generating them.

To summarily set ideas in regards to how HMMs function to generate observed event sequences, consider a discrete time HMM (λ) describing sleep pattern data. The HMM consists of two sets: a state set $S = \{S_1, S_2, S_3\}$, and an event set $E = \{E_1, E_2, E_3, E_4, E_5, E_6\}$. The elements of the state set correspond to the concepts of wakefulness, REM sleep and deep sleep respectively, where the event set refers to different types of discrete EEG measurements. It generates sleep pattern data on the basis of the following parameter sets:

$$\pi = \begin{bmatrix} 0.90 & 0.08 & 0.02 \end{bmatrix}, \theta = \begin{bmatrix} 0.70 & 0.20 & 0.02 & 0.05 & 0.02 & 0.01 \\ 0.10 & 0.02 & 0.50 & 0.30 & 0.03 & 0.05 \\ 0.02 & 0.08 & 0.05 & 0.04 & 0.51 & 0.30 \end{bmatrix}, \Gamma = \begin{bmatrix} 0.3 & 0.6 & 0.1 \\ 0.1 & 0.5 & 0.3 \\ 0.2 & 0.3 & 0.5 \end{bmatrix}.$$

Analogous to Visser (2011), in order to generate data from the HMM framework, first the value for the initial state variable $S_{t=1}$ needs to be determined by drawing from the initial state probability vector π. In this particular instance, it is assumed that the observation period is initiated when the test subject has just entered bed, which naturally corresponds to a high probability for wakefulness as being the initial state in the hidden state sequence. Given the initial state, consequently draw an observation from the appropriate row of the state-dependent distribution θ. Given that S_1 is the active underlying state at time $t = 1$, inspection of the θ matrix shows that the EEG measurement type E_1 has the highest likelihood of occurence among all events in the event set. Note that the remaining EEG measurement types have a small but nonetheless defined probability of occurence, as per the third defining characteristic of the HMM. For this example, assume that the event with the highest likelihood is drawn for $t = 1$, i.e., E_1. Subsequently, generate a transition from the appropriate row of the transition matrix Γ which provides the next value of the state variable S_{t+1}.

Given S_1, there exists a moderate to high probability for self-transitioning S_1, a high probability for transitioning to the REM sleep stage S_2, and a low probability for transitioning to the deep sleep stage S_3. For the sake of argument, assume that the subject has entered REM sleep at time t = 2, and has therefore transitioned from state S_1 to S_2. Repeat the process of alternately drawing event observations from θ for each current state, and generating a transition from Γ for determining the subsequent state, until t = T - 1. See figure 4 below for a graphical reference on this process of data generation.

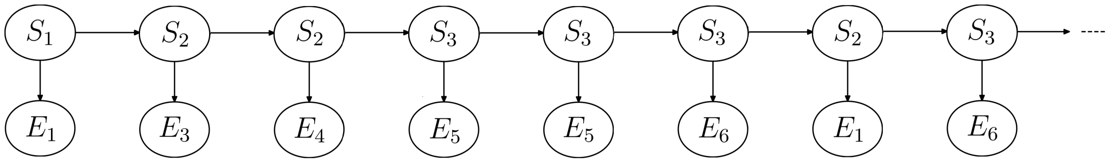


Figure 4: HMM data generation process for sleep stage cycles, extended to t = 8 for illustration purposes.

Model estimation

Having as such defined and explicated the structure of the HMM, three basic problems need to be adressed in order for the model to be applicable in real-world applications (Rabiner, 1989):

Problem 1 (Likelihood / Filtering): Given an observed event sequence $E = \{E_1, E_2, ..., E_T\}$, and an HMM model $\lambda = (\Gamma, \theta, \pi)$, how does one efficiently compute $P(E \mid \lambda)$?
Problem 2 (Decoding / Smoothing): Given the observed event sequence $E = \{E_1, E_2, ..., E_t\}$, and an HMM model $\lambda = (\Gamma, \theta, \pi)$, how does one choose a corresponding state sequence $S = \{S_1, S_2, ..., S_T\}$ which is optimal in some meaningful sense?
Problem 3 (Learning / Training): How does one adjust the model parameters $\lambda = (\Gamma, \theta, \pi)$, so as to maximize $P(E \mid \lambda)$?

For the sake of brevity and structure, the reader is referred to appendix A for a discussion on how to adress the first two problems with the forward and Viterbi algorithms. The third query relates to optimizing the HMM model parameters so that it provides an optimal description of a given observation sequence (Visser, 2011). As stated earlier, effectively adressing this question is key to most HMM applications, because it allows researchers to optimally adapt model parameters to observed event sequences, i.e., create optimal models for real-world process phenomena (Visser, 2011). A number of methods can be used to provide an answer to this problem, such as Maximum Likelihood, Expectation Maximization or the Baum-Welch algorithm (Aarts, 2016). The method that will be explicated and utilized here is Bayesian estimation, on account of its methodological flexibility (Aarts, 2016). The Bayesian approach utilizes a forward-recursion, backwards-sampling Gibbs sampler Markov Chain Monte Carlo (MCMC) algorithm to estimate the parameters of the HMM (Aarts, 2016; Rydén, 2008; Scott, 2002). In order to properly explicate this technique to the reader, the concepts of Bayesian statistics and diffuse priors are summarily discussed, on the basis of which the forward-recursion, backwards-sampling Gibbs sampler is thereafter explicated.

The Bayesian statistical framework postulates that probability expresses a certain degree of belief about the likelihood of a particular event (Lynch, 2007). This belief can either be informed by prior evidence for the event, in the form of pre-existing experimental results for example, or by novel evidence that is yet to be incorporated in the belief assesment, say recently acquired observational data (Lynch, 2007). Combining this set of prior and novel evidence results in posterior evidence, which is an updated or reinformed version of the degree of belief one has about the event (Lynch, 2007).

In a mathematical sense, given two events A and B, the conditional probability of A given that B is true is expressed by Bayes' theorem

$$\text{Posterior} \propto \text{Prior} \cdot \text{Likelihood} \longrightarrow P(A \mid B) \propto P(A) \cdot P(B \mid A)$$

where A represents the proposition about the event and B represents the evidence for A that is to be incorporated in the belief assessment (Lynch, 2007). The $P(A)$ argument represents the prior probability of the proposition A, i.e., the belief about A before any additional evidence is considered. $P(B \mid A)$ represents the likelihood function of the presented evidence given the prior belief, i.e., the probability of B given that A is true (Lynch, 2007). The product of these two terms is proportional to the posterior distribution $P(A \mid B)$, the updated degree of belief about the event (Lynch, 2007). This posterior distribution can consequently be analyzed and summarized in order to acquire an improved understanding of the nature of the event (Lynch, 2007).

In order to be able to apply the Bayesian methodology to learning problems in the HMM framework, appropriate priors for each of the model parameters (Γ, θ) of interest must first be defined. In other words, a degree of belief about their prior form has to be established. Since the true distribution values of these parameters are unknown apriori, their prior specification has to reflect this fact by deliberately specifying them in as vague a manner as possible. As such, the parameters of the prior distribution, which are also called hyperparameters, need to be chosen so that they may assume a wide range of possible values (Lynch, 2007). It is standard practice in the literature to utilize a uniform Dirichlet distribution when the parameters of interest follow a categorical distribution and are most accurately described by a set of diffuse hyperparameters (Lynch, 2007). The uniform Dirichlet distribution is a multivariate probability distribution that describes $k \geq 2$ variables $X_1, ..., X_i$, such that each $x_i \in (0,1)$ and $\sum_{i=1}^{N} x_i = 1$, that is parametrized by a vector of positive-valued parameters $\alpha = (\alpha_1, ..., \alpha_k)$ (Lynch, 2007)

$$\{x_1, ..., x_k\} \sim \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}. \tag{11}$$

with $\alpha_1 = ... = \alpha_k = 1$. The reader is referred to Appendix B for an in-depth exposition of the uniform Dirichlet prior.

Given the uniform Dirichlet prior, assume that the rows of the transition probability matrix Γ and the state-dependent probabilities θ_i are independent (Aarts, 2016):

$$S_{t=2,...,T} \sim \Gamma_{S_{t-1}} \quad \text{with} \quad \Gamma_i \sim \text{Dir}(a_{10}) \quad \text{and} \tag{12}$$

$$E_{t=1,...,T} \sim \theta_{S_t} \quad \text{with} \quad \theta_i \sim \text{Dir}(a_{20}). \tag{13}$$

The argument in Equation (12) posits that the probability distribution for the current state S_t is given by the row in Γ corresponding to the previous state in the hidden state sequence S_{t-1} (Aarts, 2016). This argument is only valid for states after the first time point, since there exists no previous state in the hidden state sequence for state S_1. The probability distribution for S_1 is instead determined by the initial probabilities of the states π_i (Aarts, 2016). Per the argument in Equation (13), the probability distribution of the observed event E_t is given by the appropriate row in θ corresponding to the form of the current state S_t (Aarts, 2016). The hyper-parameter a_{10} of the diffuse prior Dirichlet distribution on Γ_i is a row vector with length equal to the number of states m, so that $a_{10} = \alpha_1, \alpha_2 ..., \alpha_m = 1$ (Aarts, 2016). The hyper-parameter a_{20} of the diffuse prior Dirichlet distribution on θ_i is a vector with length equal to the number of observed events n, so that $a_{20} = \alpha_1, \alpha_2 ..., \alpha_n = 1$ (Aarts, 2016). It is furthermore assumed that π_i is a dependent parameter, i.e., is invariant by the stationary distribution of Γ, so that $\pi = \pi\Gamma$ (Aarts, 2016). To summarily elucidate the set of distribution specifications, consider the set of parameters for a system with $S = \{S_1, S_2, S_3\}$ and $E = \{E_1, E_2, E_3, E_4\}$, with associated diffuse hyper-parameters $a_{10} = (1, 1, 1)$ for each Γ_i and $a_{20} = (1, 1, 1, 1)$ for each θ_i. The priors for the parameters of interest Γ and θ will than be of the form

$$\theta = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}, \Gamma = \begin{bmatrix} 0.33 & 0.33 & 0.33 \\ 0.33 & 0.33 & 0.33 \\ 0.33 & 0.33 & 0.33 \end{bmatrix}.$$

where the probability distribution for each θ_i is given by $E_{t=1,...,T} \sim \theta_{S_t}$ and that of each Γ_i by $S_{t=2,...,T} \sim \Gamma_{S_{t-1}}$. Given this set of distributions, the objective is to construct the joint posterior distribution of the hidden state sequence (S) and the parameter estimates (Γ, θ), given the observed event sequence (E) and the hyper-parameters (a_{10}, a_{20}) (Aarts, 2016):

$$P((S_t), \Gamma_i, \theta_i \mid (E_t)) \propto P((E_t) \mid (S_t), \theta_i) \cdot P((S_t) \mid \Gamma_i) \cdot P(\Gamma_i \mid a_{10}) \cdot P(\theta_i \mid a_{20}).$$

The forward-recursion, backward-sampling Gibbs sampler MCMC algorithm is consequently introduced to approximate $P((S_t), \Gamma_i, \theta_i \mid (E_t))$ (Aarts, 2016; Rydén, 2008; Scott, 2002). MCMC models posit that although the actual hidden state sequence underlying the event observation sequence is unobserved and therefore unidentifiable, any statistic of its posterior distribution can be approximated given a sufficiently large set of obtained simulated samples N from its distribution (Yildirim, 2012). The Gibbs sampler is one MCMC technique that enables generating such a sufficient set of simulated samples (Yildirim, 2012). It does so by iteratively sampling from the conditional posterior distributions of S_t, Γ_i and θ_i, given the remaining parameters in the model (Aarts, 2016). It identifies a sample sequence S on the basis of E, from which it infers Γ_i and θ_i conditional on S and E (Aarts, 2016).

The component of the Gibbs sampling algorithm that adresses the sampling of S is the forward-recursion, backwards-sampling procedure, which obtains the forward probabilities $\alpha_t(i)$, i.e., the joint probability of state $S = i$ at time point t and $E_{t=1,...,T}$, given the current values of Γ and θ (Aarts, 2016; Jurafsky & Martin, 2014). See the discussion of the forward algorithm in appendix A for reference on its formal definition and procedural description in order to solve for $P(O \mid \lambda)$. On the basis of the forward probabilities $\alpha_t(i)$, the procedure consequently generates a hidden state sequence S through backward sampling of $\alpha_{T:1}$, drawing values $(S_T, S_{T-1}, ..., S_1)$ (Aarts, 2016). Conditional on this sampled hidden state sequence $S = \{S_1, S_2, ..., S_T\}$ and the observed event sequence $E = \{E_1, E_2, ..., E_T\}$, the parameters of interest Γ_i and θ_i can consequently be drawn from their conditional posterior distributions $P(\Gamma_i \mid)$ and $P(\theta_i \mid)$ (Aarts, 2016). Specifically, the i^{th} row of Γ is drawn from its conditional posterior distribution $P(\Gamma_i \mid) \sim \text{Dir}(a_{1mi})$, where a_{1mi} represents the sum of the prior Dirichlet values a_{10} and the number of transition counts from state i to state $i + 1$ in the sampled hidden state sequence (Aarts, 2016). The i^{th} row of θ is drawn from its conditional posterior distribution $(P\theta_i \mid) \sim \text{Dir}(a_{2mi})$, where a_{2mi} represents the sum of the prior Dirichlet values a_{20} and the number of observed event counts for each state i (Aarts, 2016). In essence, the Gibbs sampler combines a current Dirichlet prior with a generated evidence likelihood function regarding the parameters of interest Γ and θ, to construct the conditional posterior distributions $P(\Gamma_i \mid)$ and $(P\theta_i \mid)$ from which their updated values are drawn. This process is generally referred to as a single iteration of the Gibbs sampler, where the updated posteriors for Γ_i and θ_i subsequently function as the prior Dirichlet distribution input for the next iteration of the Gibbs sampler (Aarts, 2016; Rydén, 2008). This iterative process continues until convergence, meaning that the sample values of the parameters of interest have the same distribution as if they were sampled from the true posterior joint distribution (Yildirim, 2012). Note that because the algorithm is initialized with diffuse values, the posterior distribution samples may not necessarily be representative of the actual posterior distribution at early iterations in the sequence (Yildirim, 2012). Since MCMC theory guarantees that the samples generated under the Gibbs sampler will ultimately approximate the joint posterior of interest, they are typically run for a large number of iterations in order to achieve convergence. Because samples from early iterations can be assumed to not originate from this posterior, it is common to discard them; this initial iteration period is also referred to as the "burn-in" period (Yildirim, 2012).

To summarily set ideas in regards to the functioning of the Gibbs sampler, suppose that one wishes to find a more optimal parameterization set for the system introduced on the previous page, with $S = \{S_1, S_2, S_3\}$ and $E = \{E_1, E_2, E_3, E_4\}$, and associated diffuse hyper-parameters $a_{10} = (1, 1, 1)$ for each Γ_i and $a_{20} = (1, 1, 1, 1)$ for each θ_i.

Recall that the diffuse prior Dirichlets for the parameters Γ and θ were of the form:

$$\theta = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}, \Gamma = \begin{bmatrix} 0.33 & 0.33 & 0.33 \\ 0.33 & 0.33 & 0.33 \\ 0.33 & 0.33 & 0.33 \end{bmatrix}.$$

Assume that, given the priors and the event observation sequence $E = \{E_3, E_1, E_3, E_2, E_2, E_4, E_3, E_1\}$, the forward-recursion, backward-sampling algorithm has sampled the underlying hidden state sequence $S = \{S_2, S_1, S_1, S_3, S_2, S_3, S_1, S_2\}$. See figure 9 below for a graphical representation of this system.

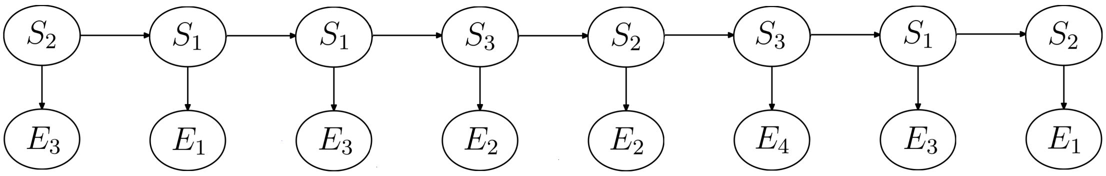


Figure 5: Visualization of the characterized system

Based on the system, index the counts for the cells in each of the respective rows of Γ and θ, where the probability distribution for each θ_i is given by $E_{t=1,\dots,T} \sim \theta_{S_t}$ and that of each Γ_i by $S_{t=2,\dots,T} \sim \Gamma_{S_{t-1}}$:

$$\theta_1 = \begin{bmatrix} 1 & 0 & 2 & 0 \end{bmatrix}, \theta_2 = \begin{bmatrix} 1 & 1 & 1 & 0 \end{bmatrix}, \theta_3 = \begin{bmatrix} 0 & 1 & 0 & 1 \end{bmatrix},$$
$$\Gamma_1 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, \Gamma_2 = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}, \Gamma_3 = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}.$$

Given the respective prior specifications for Γ and θ, the sampled hidden state sequence S and the observation sequence E, and the respective counts for θ_i and Γ_i, the conditional posterior distribution $P(\Gamma_i \mid)$ and $P(\theta_i \mid)$ can be constructed as

$$\begin{aligned} \mathrm{Dir}(a_{11}) &= (1+1, 1+0, 1+2, 1+0) = (2,1,3,1), \\ \mathrm{Dir}(a_{12}) &= (1+1, 1+1, 1+1, 1+0) = (2,2,2,1), \\ \mathrm{Dir}(a_{13}) &= (1+0, 1+1, 1+0, 1+1) = (1,2,1,2). \end{aligned}$$

$$\begin{aligned} \mathrm{Dir}(a_{21}) &= (1+1,\ 1+1,\ 1+1) = (2,2,2), \\ \mathrm{Dir}(a_{22}) &= (1+1,\ 1+0,\ 1+1) = (2,1,2), \\ \mathrm{Dir}(a_{23}) &= (1+1,\ 1+1,\ 1+0) = (2,2,1). \end{aligned}$$

Drawing values from $P(\Gamma_i \mid)$ and $P(\theta_i \mid)$ results in the posterior parameter set

$$\theta = \begin{bmatrix} 2/7 & 1/7 & 3/7 & 1/7 \\ 2/7 & 2/7 & 2/7 & 1/7 \\ 1/6 & 2/6 & 1/6 & 2/6 \end{bmatrix}, \Gamma = \begin{bmatrix} 2/6 & 2/6 & 2/6 \\ 2/5 & 1/5 & 2/5 \\ 2/5 & 2/5 & 1/5 \end{bmatrix}$$

which function as the parameter sets for consequent sampling of S and as the updated prior for the subsequent iteration in the Gibbs sampling procedure.

Problem definition

To re-iterate, the central objective of this study pertains to establishing references on the topic of how certain input variables affect the Gibbs samplers capacity to produce accurate estimates. It additionally seeks to explicate for which single and combined value input ranges these variables enable the Gibbs sampler to produce optimal system approximations. By doing so it seeks to adress current knowledge gaps on the topic of appropriate Gibbs sampling input specification for single sequence HMM learning. Based on the extant literature, and given the structure of the Gibbs sampler, two variables are presently identified as influencing the Gibbs sampler's estimation performance: The ratio between the number of states and event types, and the length of the event observation sequence. Note that this is not an exhaustive list of the variables that can be expected to influence the Gibbs samplers performance; the shape of the probability distributions of both the transition and conditional matrix would for instance have constituted an additional intuitive assessment variable. However, time constraints and processing power limitations require this inquiry to limit the scope of its investigation. Furthermore, given the relative scarcity of the available literature on Gibbs sampler input specification, assessing the role of these two variables is a worthwhile first step in establishing reference material on the subject. As such, the following two reasearch questions are formulated: How do increases in the number of event types, and the length of the event observation sequence affect the accuracy of the Gibbs sampler's estimates? And for which individual and combined value ranges are these estimates optimal? Hypotheses regarding the relation between the Gibbs sampler and the input variables of interest are consequently formulated on the basis of statistical reasoning, and the extant literature on optimal Gibbs sampling input specification.

With respect to the ratio between the number of states and event types, the literature suggests that introducing additional event types will, ceteris paribus, enable the Gibbs sampler to more effectively adress learning problems in the HMM framework. This argument derives from the notion that isolating pattern occurences is more difficult in contexts where the event sequence of interest displays a high pattern periodicity (Chudova & Smyth, 2002). It is a more strenuous task for the Gibbs sampling algorithm to identify periodic pattern boundaries for such event sequence types, because they inherently comprise of unlabeled data structures (Chudova & Smyth, 2002). The absense of clear data label patterns makes the Gibbs procedure more prone to event-state misclassification, which consequently introduces error into its estimates regarding the latent structure of the HMM (Chudova & Smyth, 2002).

In a study on the subject, Chudova & Smyth (2002) examined the hypothesis that high pattern periodicity complicates HMM learning, by positing that an increase in the autocorrelation of periodic observation patterns would result in a decrease in the Gibbs samplers ability to effectively approximate the parameters of the HMM. In other words, observation pattern sequences with a high or even uniform autocorrelation, such as AAAAAA, were expected to represent a more difficult learning task than sequences with a relatively moderate autocorrelation, such as ABABABAB or ABCABCAB (Chudova & Smyth, 2002). As per the expectation, the authors found that increasing pattern structure periodicity resulted in higher estimation error probabilities for the parameters of the HMM (Chudova & Smyth, 2002). See figure 6 for a graphical illustration on how estimation error will increase given a higher sequence pattern periodicity.

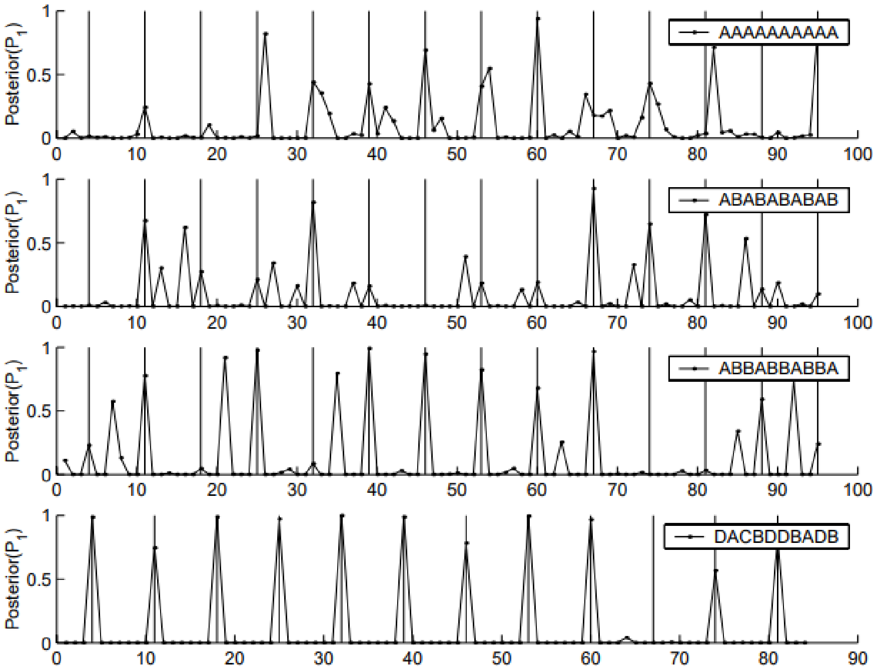


Figure 6: Posterior probabilities for the first state in the system as presented by Chudova & Smyth (2002), given increasingly random pattern periodicities. The X-axis represents position in the sequence. Reprinted from Chudova, & Smyth (2002).

Chudova & Smyth (2002) consequently conclude that ”in general, the detection of structured patterns in a Markov context presents a more difficult learning problem than the detection of random patterns”. Another study on the subject by Van Helden, André, & Collado-Vides (1998) reported that only patterns with a clear periodic structure complicated learning inference, reinforcing the notion that the boundaries of periodic patterns are harder to determine than those of non-periodic patterns. Barring these two studies, work on the influence of pattern periodicity on the accuracy of the Gibbs samplers estimates is generally sparse.

Analogous to Chudova & Smyth, this study therefore reasons that an increase in the number of event types will induce randomization into the patterns of the observation sequence, thereby lowering autocorrelation and increasing pattern variabllity. Increased pattern variability will consequently better enable the Gibbs sampler to identify the periodic boundaries of the patterns in the sequence, and infer from it the underlying latent structure of the HMM. As such, it is here hypothesized that the introduction of additional event types will function to improve the estimates of the Gibbs sampler. Given the hypothesized relationship between the number of event types and the functioning of the Gibbs sampler, this study consequently seeks to identify the value input range for which this variable enables the Gibbs sampler to optimize its HMM parameter approximations. Optimization is defined here as selecting the best input value from the set of available alternatives on the basis of some selection criterion. This criterion is characterized as being that value where the marginal cost of obtaining one estimator accuracy increment outweighs its marginal benefit. Its definition is based on the notion that in general, studies seek to obtain estimates which are as optimal as possible, while minimizing the time and computing power required to produce it. As such, this study hypothesizes the optimal number of event types to be situated in the lower, but not the lowest bounds of the event type variable. This expectation is based on findings by Chudova & Smyth (2002), who report that initial increases in the number of event types resulted in drastic increases in the accuracy of the Gibbs samplers estimates, but that the effect of subsequent increases were generally redundant. The reader is once more referred to figure 6 for a visual demonstration of this finding.

With respect to the length of the observed event sequence E, additional observational data will enable the forward recursion - backward sampling algorithm to produce a larger sample of the hidden state sequence, which consequently provides additional information from which to infer its latent structure (Jurafsky & Martin, 2014; Lynch, 2007). One would therefore intuitively expect a longer event sequence to result in improved Gibbs sampling estimates for the parameters of the HMM. This argument is intuitive because, by the theorem of large numbers, the sample average $\bar{A} = \frac{1}{n}(X_1 + ... + X_N)$ will converge to the expected value $\bar{A} \to \mu$ for $n \to \infty$ (Hsu & Robbins, 1947). More formally, assume that $X_1, ..., X_N$ are independent and identically distributed random variables with mean μ. Let $\bar{A}$ represent the average of n variables. Then, for any $\epsilon \geq 0$, the following must hold (Hsu & Robbins, 1947):

$$\lim_{n \to \infty} P(|\bar{A} - \mu| \leq \epsilon) = 1.$$

An illustration of the phenomenon of the law of large numbers is provided in the discussion on updating priors in appendix B. A graphical illustration of the phenomenon is furthermore provided in figure 7 on the next page. The notion that additional sequence length is beneficial to the Gibbs samplers performance is moreover supported by empirical studies on the subject.

For instance, Chen and Schmeiser (1993) recommend the use of a single long run as opposed to an aggregate of multiple short runs to minimize point-estimator bias. In a comment on implementation strategies for MCMC sampling techniques, Raftery & Lewis (1992) similarly ”recommend that inference ultimately be based on a single long run, but that this be monitored using carefully chosen diagnostics”.

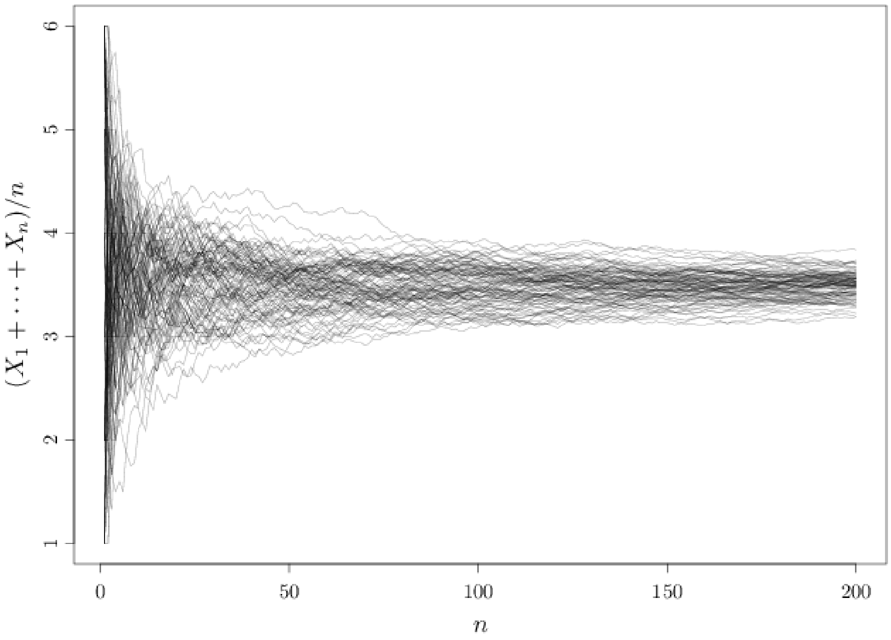


Figure 7: Case example of the mechanism of the law of large numbers for a six-sided die. The figure shows that as the number of rolls n increases, the averages values of the die converge to the expected value $\mu = 3.5$.

Furthermore, a study on the topic of applying Bayesian techniques to natural language processing (NLP) problems found that the accuracy of Gibbs sampler HMM parameter estimates increased when the total input data did as well (Gao & Johnson, 2008). Chudova & Smyth (2002) similarly report that, ceteris paribus, additional input data resulted in improved Gibbs sampler HMM parameter estimates. Although these studies are relatively dated, they nonetheless lend empirical credence to the argument that, in general, additional data input and single long runs will enable the Gibbs sampler to improve on its parameter estimates. This study consequently hypothesizes that, ceteris paribus, an increase in the sequence length will result in improved Gibbs sampler estimates. With regards to the optimization query, both the law of large numbers and literature suggest that increases in the length of the event observation sequence will continuously result in improved estimates of the system, albeit with increasingly decreasing marginal returns. As such, this study hypothesizes that the marginal estimator accuray rate for this variable follows a logarithmic growth curve, where optimal value input ranges are situated in the upper quartile of the variable.

Having thus characterized the general and optimal input relations between the variables of interest and the Gibbs sampler on the individual variable level, this study subsequently seeks to identify optimal such input relations on the combined variable level. The objective in regards to this query is twofold. Analogous to the individual level optimization hypotheses, it first seeks to identify the combined value specification ranges for which the Gibbs samplers approximation of the systems latent structure is optimized. It secondly seeks to identify constrained combined optimal value ranges, i.e., variable specifications which produce optimized Gibbs sampler estimates given that one of the two variables is constrained to a particular value. Establishing references with regards to this second query is especially significant, because it enables researchers to adapt the input specifications of the Gibbs sampler to the conditions of the context application. Put differently, it maximizes Gibbs sampler estimates in application contexts where specifying optimal input values is, for one reason or another, unfeasible. An example of a constrained context could pertain to the topic of exoplanet discovery, where the length of the event observation sequence is determined by the time interval in which the exoplanet of interest is situated in front of its host star. With regards to the unconstrained optimization query, this study hypothesizes optimal value ranges to be situated in the intersection between the lower bound values of the number of event types, and the upper quartile values for the length of the observed event sequence. With regards to the constrained optimization query, this study hypothesizes that, given the constraint, estimates will be maximized in accordance with the optimal individual variable input values. In other words, this study hypothesizes that, given a constrained value for the number of event types, optimal Gibbs sampler input values for the length of the sequence will be situated in the upper quartile. Inversely, given an imposed constraint on the length of the observation sequence, optimal Gibbs sampling inputs for the number of event types will be situated in the the lower bounds of the variable. This inquiry will proceed by explicating the design of the simulation study that is utilized to investigate the sum of the here defined hypotheses. The results of this study are consequently reported and discussed, on the basis of which conclusions are drawn with respect to the here formulated hypotheses.

Simulation design

The proposed relations between the variables of interest and the functioning of the Gibbs sampler are assessed by conducting and subsequently interpreting the results of a simulation study. Simulations can be understood to be approximate imitations of the functioning of a particular process or system (Banks, Carson, Nelson, & Nicol, 2010). They consist of a model, which provides a description of the process or system of interest, and the simulation itself, which describes how the model operates in a particular scenario (Banks, et al., 2010). Simulation studies comprise of a set of multiple such simulations, which describe the operations of the model of interest in a number of different scenarios (Banks, et al., 2010). Empirical results regarding these operations can consequently be analyzed to evaluate the models performance in each scenario (Morris, White, & Crowther, 2019).

In this particular study, the Gibbs samplers capacity to retrieve latent HMM structures is the system of interest. The model used to represent the Gibbs sampler is specified in R, a language and environment for statistical computing, developed at Bell Laboratories by John Chambers and colleagues. The specific R function that prescribes how the Gibbs sampler model operates is authored by Emmeke Aarts, who is an assistant professor based at the department of methodology and statistics at Utrecht University. Given the Gibbs sampler model, the objective of this simulation study is to describe the procedure's operational capacity in retrieving the latent parameter values for a number of varyingly specified HMM data sequences. These sequences have been generated with an R-function that emulates the functioning of a generic HMM. Like the model function, this function was originally authored by Aarts. The input values for the HMM function were able to vary with respect to the number of states, the number of event types, the length of the input sequence, the initial state probabilities and the probability matrices for Γ and θ. The number of states and its associated transition probability matrix Γ were kept constant over all HMMs. Specifically, a three state HMM system was specified, with associated transition probability matrix:

$$\Gamma = \begin{bmatrix} 0.87 & 0.06 & 0.07 \\ 0.03 & 0.92 & 0.05 \\ 0.07 & 0.01 & 0.92 \end{bmatrix}.$$

The initial state probability vector was similarly kept constant over all HMMs, with:

$$\pi = \begin{bmatrix} 0.40 & 0.30 & 0.30 \end{bmatrix}.$$

The values for the length of the observation sequence were able to vary over the elements of the set $L = \{1000, 2000, 4000, 6000, 8000\}$, whereas the values for the number of event types were able to vary over the elements of the set $N = \{3, 4, 6, 8, 10\}$. The ranges for both these variables were formulated to represent a comprehensive span of their typical real-world values.

Note that the lower value intervals are somewhat more narrow as opposed to those in the upper ranges. This design choice seeks to reflect the consensus in the literature that initial increases in N and L will have a larger impact on the accuracy of the Gibbs samplers estimates. The respective conditional probability matrices for each element of N were consequently characterized as:

$$\theta_3 = \begin{bmatrix} 0.87 & 0.06 & 0.07 \\ 0.03 & 0.92 & 0.05 \\ 0.07 & 0.01 & 0.92 \end{bmatrix}, \theta_4 = \begin{bmatrix} 0.68 & 0.24 & 0.06 & 0.02 \\ 0.02 & 0.22 & 0.75 & 0.01 \\ 0.03 & 0.03 & 0.01 & 0.93 \end{bmatrix},$$

$$\theta_6 = \begin{bmatrix} 0.62 & 0.31 & 0.03 & 0.02 & 0.01 & 0.01 \\ 0.02 & 0.01 & 0.55 & 0.38 & 0.01 & 0.03 \\ 0.03 & 0.03 & 0.01 & 0.02 & 0.58 & 0.33 \end{bmatrix},$$

$$\theta_8 = \begin{bmatrix} 0.34 & 0.25 & 0.25 & 0.05 & 0.02 & 0.03 & 0.04 & 0.02 \\ 0.02 & 0.02 & 0.01 & 0.42 & 0.29 & 0.21 & 0.01 & 0.02 \\ 0.01 & 0.03 & 0.01 & 0.02 & 0.01 & 0.03 & 0.40 & 0.49 \end{bmatrix},$$

$$\theta_{10} = \begin{bmatrix} 0.34 & 0.23 & 0.21 & 0.05 & 0.02 & 0.03 & 0.04 & 0.02 & 0.04 & 0.02 \\ 0.02 & 0.02 & 0.01 & 0.32 & 0.22 & 0.19 & 0.16 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.01 & 0.01 & 0.02 & 0.01 & 0.02 & 0.01 & 0.36 & 0.29 & 0.25 \end{bmatrix}.$$

Note that although the probability mass that is assigned to each event type shifts between the varying conditional matrices, it remains relatively equal for the sum of event types that are key indicators for a particular state. This simulation design represents how increases in N function to introduce pattern variability into the observation sequence of the HMM. Given the value ranges of the input variables, the HMM simulation function consequently proceeded to generate an HMM data sequence for each value combination of N and L. For example, table 1 shows that in scenario one, an HMM with three event types generated an HMM observation sequence with a length 1000. The sum of all simulation combinations between N and L resulted in a total of $5 \cdot 5 = 25$ simulated HMM data sequences. The reader is once more referred to table 1 below for an overview of the complete set of simulation scenarios.

Table 1: Set of HMM data simulation scenarios.

		N				
		3	4	6	8	10
L	1000	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
	2000	Scenario 6	Scenario 7	Scenario 8	Scenario 9	Scenario 10
	4000	Scenario 11	Scenario 12	Scenario 13	Scenario 14	Scenario 15
	6000	Scenario 16	Scenario 17	Scenario 18	Scenario 19	Scenario 20
	8000	Scenario 21	Scenario 22	Scenario 23	Scenario 24	Scenario 25

The value specifications of the Gibbs sampler model were consequently able to vary with respect to the number of iterations, the number of samples, the burn-in period and its starting values. For all scenarios, the Gibbs sampler procedure was repeated for a total of 4000 iterations, with a burn-in period of 1000. For scenarios 1 - 5, five-hundred samples were drawn for each scenario, for scenarios 6 - 15, two-hundred-and-fifty samples were drawn each scenario, while for scenarios 16 - 25, one-hundred samples were drawn for each scenario. As was alluded to earlier, due to constraints in time and computing power, this study was unfortunately unable to draw equally large sample sizes for all scenarios. As a consequence it had to impose limitations on the sample sizes for HMM data sequences with higher values for N. The starting values for the transition probability matrix were specified as

$$\Gamma_s = \begin{bmatrix} 0.40 & 0.30 & 0.30 \\ 0.30 & 0.40 & 0.30 \\ 0.30 & 0.30 & 0.40 \end{bmatrix},$$

whereas the starting values for the respective conditional probability matrices were specified as:

$$\theta_{3s} = \begin{bmatrix} 0.40 & 0.30 & 0.20 \\ 0.30 & 0.40 & 0.30 \\ 0.30 & 0.30 & 0.40 \end{bmatrix}, \theta_{4s} = \begin{bmatrix} 0.30 & 0.30 & 0.20 & 0.20 \\ 0.20 & 0.30 & 0.30 & 0.20 \\ 0.20 & 0.20 & 0.20 & 0.40 \end{bmatrix},$$

$$\theta_{6s} = \begin{bmatrix} 0.24 & 0.24 & 0.13 & 0.13 & 0.13 & 0.13 \\ 0.13 & 0.13 & 0.24 & 0.24 & 0.13 & 0.13 \\ 0.13 & 0.13 & 0.13 & 0.13 & 0.24 & 0.24 \end{bmatrix},$$

$$\theta_{8s} = \begin{bmatrix} 0.20 & 0.15 & 0.15 & 0.10 & 0.10 & 0.10 & 0.10 & 0.10 \\ 0.10 & 0.10 & 0.10 & 0.20 & 0.15 & 0.15 & 0.10 & 0.10 \\ 0.10 & 0.10 & 0.10 & 0.10 & 0.10 & 0.10 & 0.20 & 0.20 \end{bmatrix},$$

$$\theta_{10s} = \begin{bmatrix} 0.20 & 0.12 & 0.12 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 \\ 0.07 & 0.07 & 0.07 & 0.18 & 0.14 & 0.14 & 0.12 & 0.07 & 0.07 & 0.07 \\ 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.08 & 0.20 & 0.12 & 0.12 \end{bmatrix}.$$

Note that the starting values have been specified so as to slightly steer the estimates of the Gibbs sampler in the direction of the true underlying parameter values of the model. This was done to prevent label switching. The reader is referred to Scott (2002) for a discussion on this issue. Also note that the starting values for the first row of θ_{3s} were misspecified by the researcher. Due to this specification error, the likelihood of the occurence of label switching was increased for all the three event type scenarios. Initial priors for the Gibbs sampler were given uniform and symmetric Dirichlet distributions, for reasons explicated in appendix B and in the discussion on Gibbs sampling on pages 10-14. The Gibbs sampler model was subsequently ran on each scenario, resulting in sample ensemble estimates of the latent structure underlying each of the twenty-five scenarios.

Results

The accuracy of the Gibbs samplers ensemble parameter estimates for the varying scenarios will be assessed by comparing them with with their true parameter values, which were specified on pages 20 and 21 of the previous section. Unfortunately, a number of scenarios suffered from severe label switching. Specifically, all scenarios with three event types, and all scenarios with an observation sequence of length 1000 displayed switched components. As such, these scenarios had to be excluded from the analysis. Some incidental label switching also ocurred for the scenarios with four event types and associated sequence lengths 2000 and 4000, and the scenario with eight event types and associated sequence length 2000. However, in these particular instances only a single sample value of the total ensemble sample displayed switched components. The influence of label switching on these scenarios parameter estimates was therefore negligible. After excluding the switched component scenarios, a total of sixteen scenarios were incorporated into the analysis. Three statistical measures were subsequently utilized to facilitate the comparison between the estimated and true parameter values, the average bias, the Root Mean Square Error of Approximation or RMSEA, and the coverage. The definitions of these statistical measures are first summarily discussed. The average bias of an ensemble sample of estimators is the difference between each estimator's posterior mean of the estimated value, and the true value of the parameter being estimated, averaged over all samples (Lehmann, & Casella, 2006). It is obtained with the argument

$$\text{Bias}_{\xi}(\hat{\xi}) = \frac{1}{Z} \cdot \sum_{z=1}^{Z} (\hat{\xi} - \xi)$$

with sample size $z \in \{1, ..., Z\}$, $\hat{\xi}$ as the posterior mean of the estimator's estimated value, and ξ as a real number representing the true value of the parameter being estimated (Lehmann, & Casella, 2006). The RMSEA represents the sample variation of the difference between the estimated value and the value of what is estimated over the sample size $z \in \{1, ..., Z\}$, and is given by the argument (Lehmann, & Casella, 2006):

$$\text{RMSEA}_{\xi}(\hat{\xi}) = \sqrt{\frac{1}{Z} \cdot \sum_{z=1}^{Z} ((\hat{\xi} - \xi)^2)}.$$

Note that the interpretation of the bias and the RMSEA are relative to the scale of the estimated parameter, i.e., a bias of 0.05 is relatively large if the parameter has a true value of 0.10, but relatively small if it has a true value of 1. To aid interpretations of the results, relative definitions for the bias and RMSEA have additionally been specified, which express the bias and RMSEA in percentages relative to the true parameter value.

Moreover, due to the fact that the distribution of the probability mass varies for the respective conditional matrices, specifying relative measures for the bias and RMSEA will enable their inter-scenario comparison. For the sake of being thorough, these relative measures have been obtained with the arguments:

$$\text{Bias}_{\xi\text{rel}}[\hat{\xi}] = \frac{1}{Z} \cdot \sum_{z=1}^{Z}((\hat{\xi} - \xi)/\xi), \text{ and } \text{RMSEA}_{\xi\text{rel}}(\hat{\xi}) = \sqrt{\frac{1}{Z} \cdot \sum_{z=1}^{Z}(((\hat{\xi} - \xi)/\xi)^2)}.$$

The coverage of the 95% credibility interval finally represents the percentage of the number of times over the samples that the interval contains the true parameter value. The size of the credibility interval is also assessed, so as to diagnose the precision of the Gibbs samplers' estimates.

The results of the simulation study are consequently evaluated on the basis of the set of diagnostic statistical measures. The reader is referred to figure 8 for an overview of the respective biases for the parameters of the transition probabilities (Γ) for all scenarios. From left to right, the bias has been plotted for each of the number of event type categories N, with the length of the sequence L as the grouping variable. The respective X-axes depict the value of the bias, whereas the respective Y-axes depict each of the parameters in Γ.

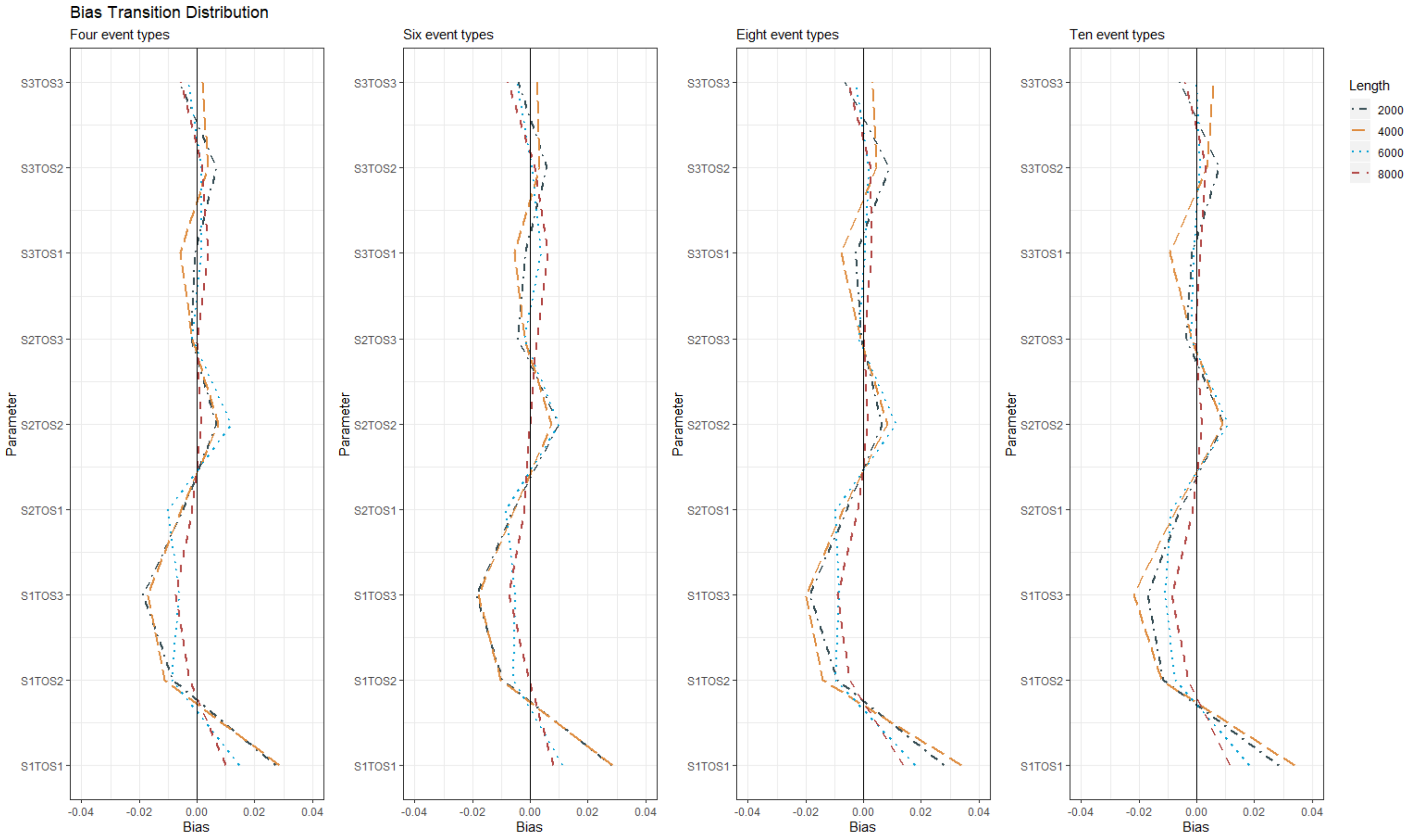


Figure 8: Bias of the parameters of the transition probability distribution for the set of event type categories, with grouping on length. Bias is shown on the X-axis, the parameters in Γ on the Y-axis.

First note that the scale of the X-axes range from -0.04 to 0.04 for all event type categories, indicating that the bias is generally low for all respective parameters of Γ. Upon further inspection it becomes strikingly apparent that the bias of the transition probability parameters display highly similar patterns over each of the event type categories. This is an indication that, for the here specified HMM scenarios at least, the Gibbs samplers' estimates of these transition probability parameters will approach certain estimator values regardless of variations in N. Transition probability parameter estimates furthermore improve for all event type categories when L increases, especially for sequence lengths 6000 and 8000, albeit it only slightly. These preliminary results are consequently reassessed by inspecting the relative bias of the transition probabilities, for which the reader is referred to figure 9. As becomes readily apparent from the graph, the relative bias is lowest for the self-transition probabilities, whereas it is relatively high for the between-transition probabilities. This is an intuitive finding, because slight under- or overestimations will be more impactful for parameters with low true values as opposed to those with high true values. Moreover, as in the absolute bias plot, the transition probability parameters display highly similar patterns over each of the event type categories, indicating that this variable is redundant to bias reduction in Γ. The bias furthermore decreases notably for increases in the length of the event observation sequence, again specifically for lengths 6000 and 8000. The relative bias thus reaffirms the notion that there is no effect on the accuracy of the Gibbs samplers' estimates of the transition probabilities for increases in N, while there is for L.

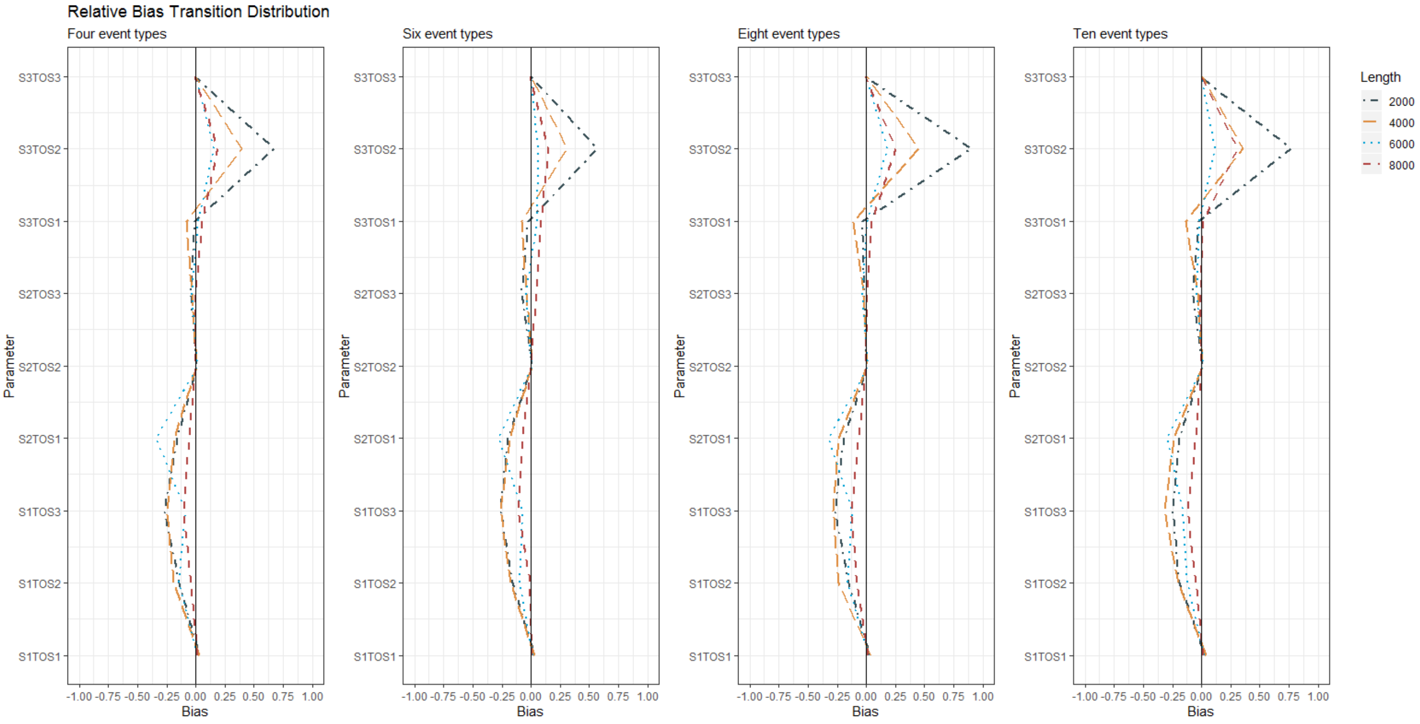


Figure 9: Relative bias of the parameters of the transition probability distribution for the set of event type categories, with grouping on length. Bias is shown on the X-axis, the parameters in Γ on the Y-axis.

The reader is consequently referred to figure 10 for an overview of the biases for the parameters of the conditional probabilities (θ) for all scenarios. As in the transition probability plots, the bias has been plotted from left to right for the number of event type categories N, with the length of the sequence L as its grouping variable. The respective X-axes depict the value of the bias, whereas the respective Y-axes depict each of the parameters in θ.

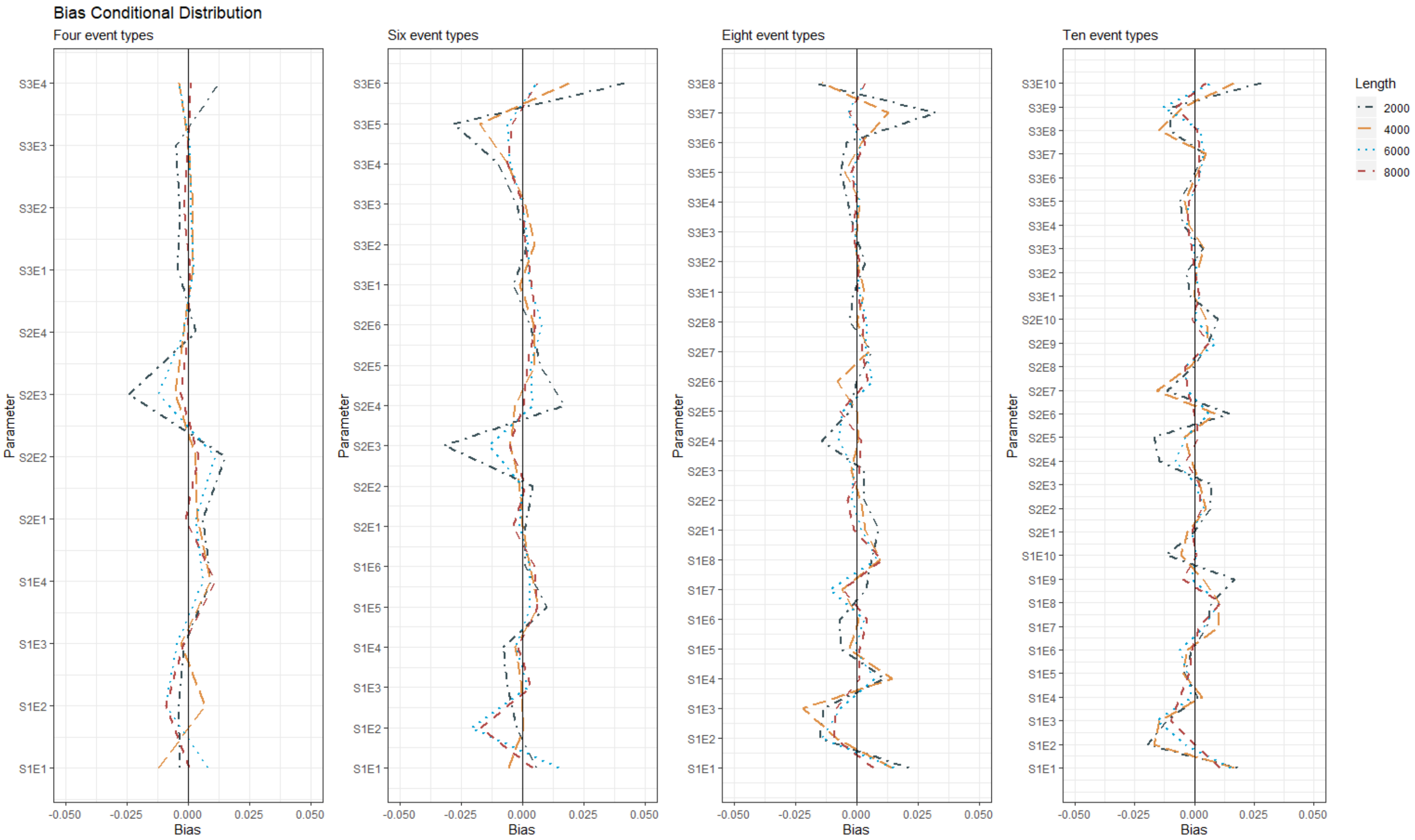


Figure 10: Bias of the parameters of the conditional probability distribution for the set of event type categories, with grouping on length. Bias is shown on the X-axis, the parameters in θ on the Y-axis.

Note that the scale of the X-axes range from -0.05 to 0.05 over all event type categories, indicating that the bias is generally low for all respective parameters of θ. Further inspection of the plot shows that the bias is generally most under- and overestimated for key state components, such as S1E2, S2E3, and S3E5, in the six event type category for example. This is an intuitive finding however, because the true probabilities for these components are much higher than those of non-key event components, sensitising them to adopting larger absolute bias values relative to non-key state components. As such, these values should additionally be evaluated from a relative perspective to assess the true severity of these biases. The plot furthermore seems to suggest the presence of an effect of N on the accuracy of the Gibbs samplers estimates of θ. The parameter progression of the bias is slightly more narrow for the four, eight and ten event type plots, as opposed to the six event type plot.

This distinction is however not very explicit, and should therefore also be re-evaluated in the context of the relative bias. Parameter estimates furthermore improve for increases in L over all event type categories, most notably so for the step from 2000 to any of the higher sequence lengths. The reader is consequently directed to figure 11 for an overview of the relative bias of the conditional probabilities. As was expected, reinspection of the S1E2, S2E3, and S3E5 parameters shows that the relative bias for the key state components is much lower than for the non-key state components. Upon further inspection of the plot, it becomes strikingly clear that the conditional probability biases are generally lowest in the four event type category. This is especially true when coupled with a sequence length of 8000, for which the conditional probability estimates closely approximate the true values for almost all parameters of θ. The relative bias patterns for the remaining event types are subsequently somewhat more fluctuative as opposed to that of event type four, but generally approach close paramater approximations of the true conditional probability values for higher sequence lengths of L. For all event types scenarios therefore, an increase in length of the sequence generally results in a decrease in the bias, especially for the step from length 2000 to any of the higher lengths.

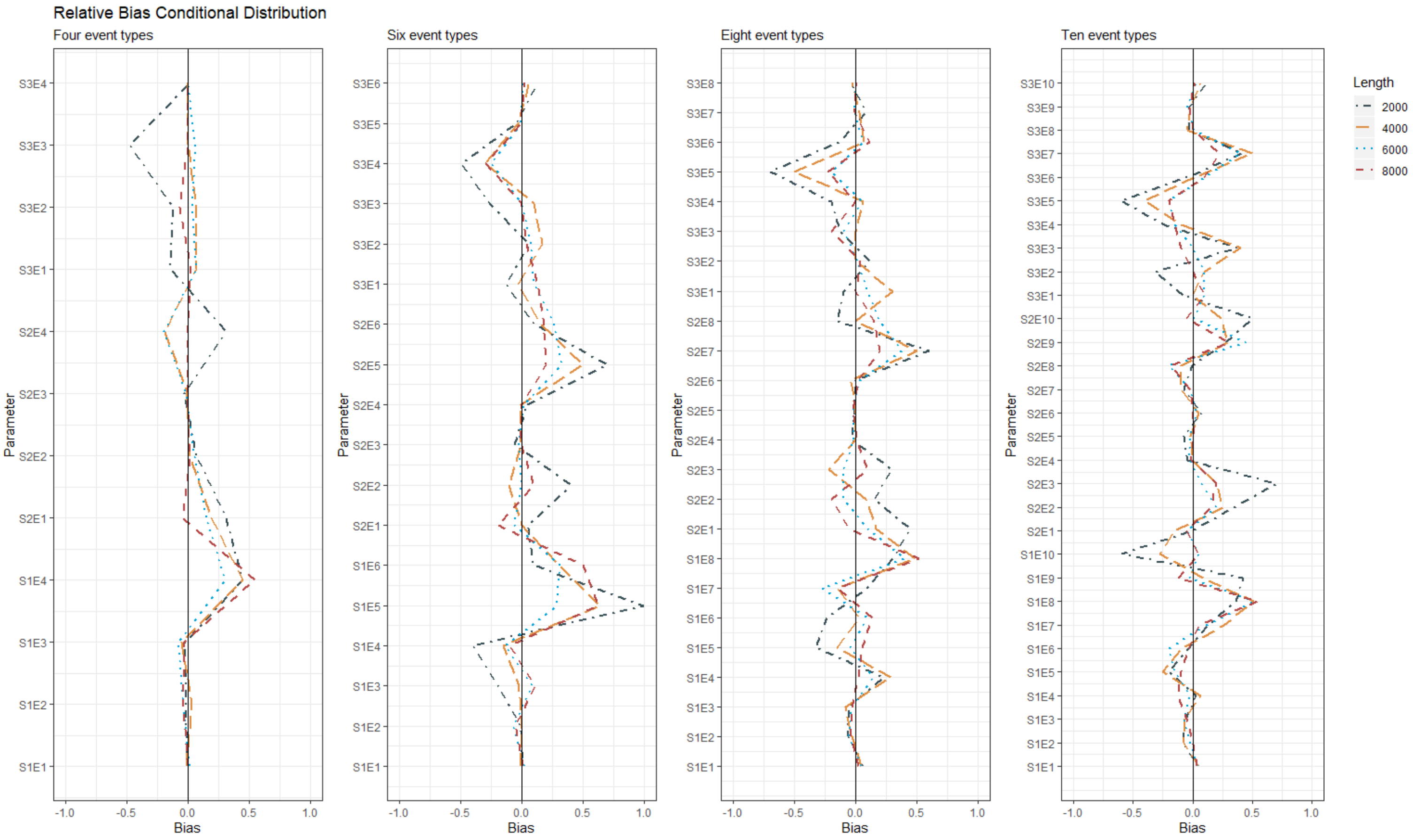


Figure 11: Relative bias of the parameters of the conditional probability distribution for the set of event type categories, with grouping on length. Bias is shown on the X-axis, the parameters in θ on the Y-axis.

The reader is consequently referred to figure 12 for an overview of the absolute and relative RMSEA of the transition probabilities for all scenarios. The figure on the left indicates that the RMSEA for the transition probabilities remains relatively even between sequence lengths 2000 and 4000 - even increasing slightly for the eigth and ten event types - after which it decreases starkly for all event types in the subsequent 6000 and 8000 length ranges. The RMSEA values for the four and six event type categories are lowest throughout the graph, but the values for the entire event type set are generally in close proximity to one another, ultimately converging in the 6000 and 8000 range. These findings are in line with the reported findings on the bias, which indicated that increases in L - specifically for lengths 6000 and 8000 - and not so much N were important for ensuring low bias values for the parameters of the transition probability distribution. The right figure similarly shows that the relative RMSEA decreases sharply as the length of the sequence increases, whereas relatively little differences in the RMSEA are observed between the different event types. The four and six event types can once again be said to be somewhat more optimal than the others, albeit it only very slightly so.

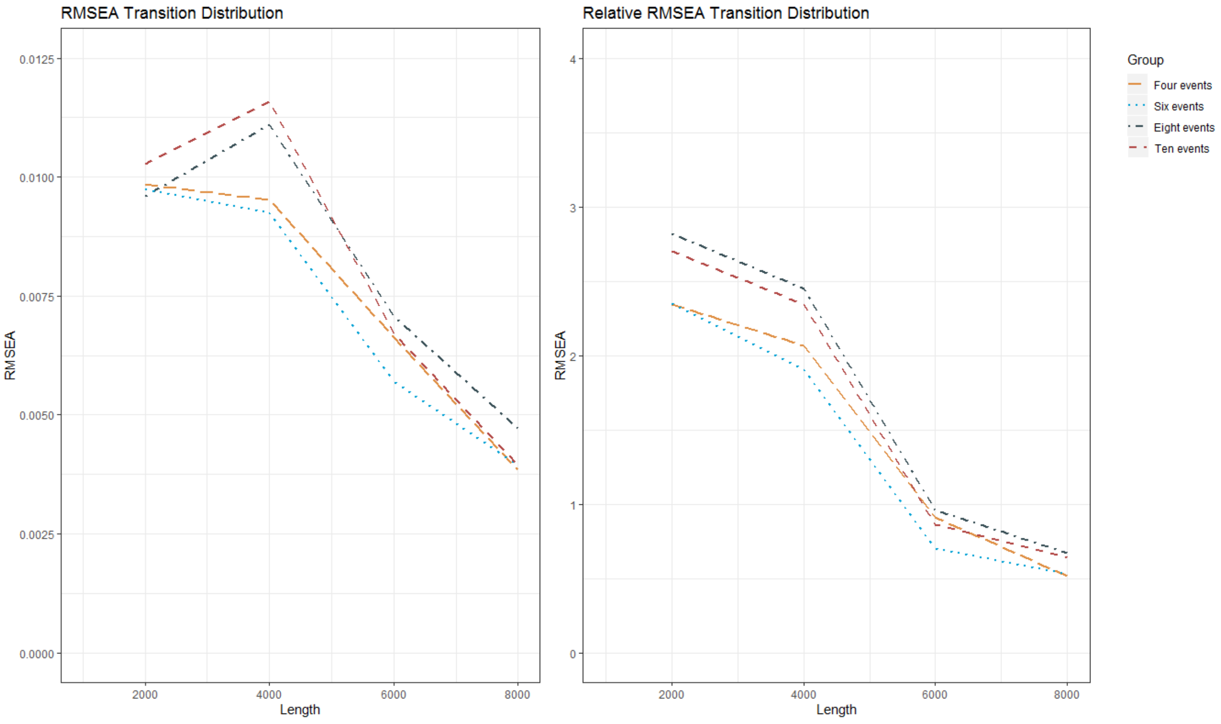


Figure 12: Absolute and relative RMSEA for the transition probability distributions for all scenarios. The length of the associated sequence is shown on the X-axis, the RMSEA - absolute on the left, relative on the right - is shown on the Y-axis. Grouping is by event type.

The reader is subsequently directed to figure 13 for an overview of the absolute and relative RMSEA of the conditional probabilities for all scenarios. The figure on the left indicates that at length 2000, the RMSEA for the four event type category is high relative to all other event types. This is most probably a consequence of the absolute squared bias being inflated for this particular event type, as a result of it inherently having higher conditional probabilities due to having less components relative to the other event types. As such, its more sensitive to adopting higher absolute bias values, and should therefore be contextualized relative to its true scale. Initial inspection of the relative bias plot on the right shows that in the context of the true scale, the RMSEA is actually lower than for the other event types. Once again redirecting the reader's attention the the left figure, it becomes strikingly apparant that for all event types, the RMSEA decreases when the length of the sequence increases. This decrease is furthermore strongest for the four event type category, and less so but nonetheless still quite substantial for the remaining event types. At the 6000 and 8000 sequence lengths, the RMSEA values for the respective event types converge at very low values, indicating that there is no substantial difference in the values of the RMSEA between event types at that point. The figure on the right consequently shows that on a relative level, the RMSEA decreases sharply when the length does as well, regardless of the number of event types. However, similarly to what was reported in the findings of the relative bias of the conditional probabilities, the figure shows slightly more optimal values for four as opposed to other event types.

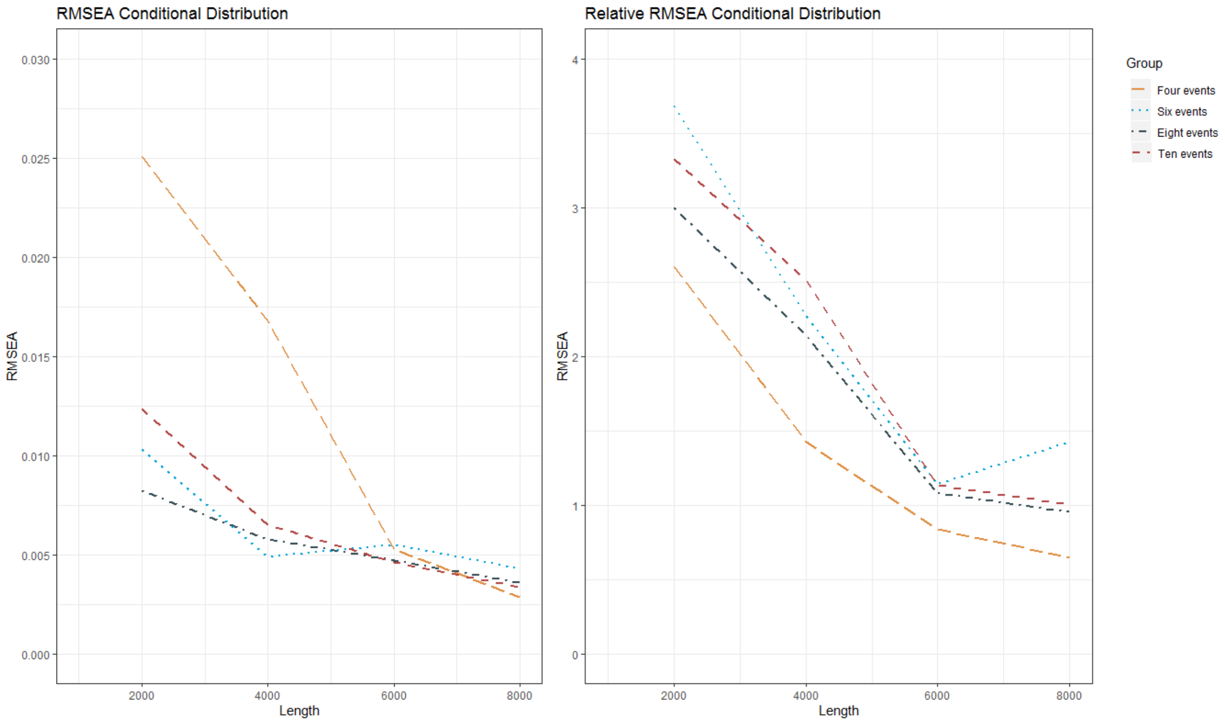


Figure 13: Absolute and relative RMSEA for the conditional probability distribution for all scenarios. The length of the associated sequence is shown on the X-axis, the RMSEA - absolute on the left, relative on the right - is shown on the Y-axis. Grouping is by event type.

The reader is referred to figure 14 for an overview of the coverage for both the transition and conditional probabilities. As becomes apparent from the left graph in the figure, the coverage for the transition probabilities is relatively low for all event types at length 4000 - about 75% to 80% - whereas it is higher for length 6000 - between 90% and 99%, and generally approaches 1 for lengths 2000 and 8000. As such, these findings indicate that parameter estimates of the transition probabilities fall outside the 95% coverage interval to the highest degree for the 4000 sequence length, to a lesser degree for the 6000 sequence length, while almost always falling inside for the 2000 and 8000 lengths. The figure on the right furthermore indicates that the coverage for the conditional probabilities is generally high for all scenarios, regardless of L or N. To properly diagnose these findings, especially with regards to the transition probability coverage of lengths 4000 and 6000, the average size of the credibility interval is additionally assessed.

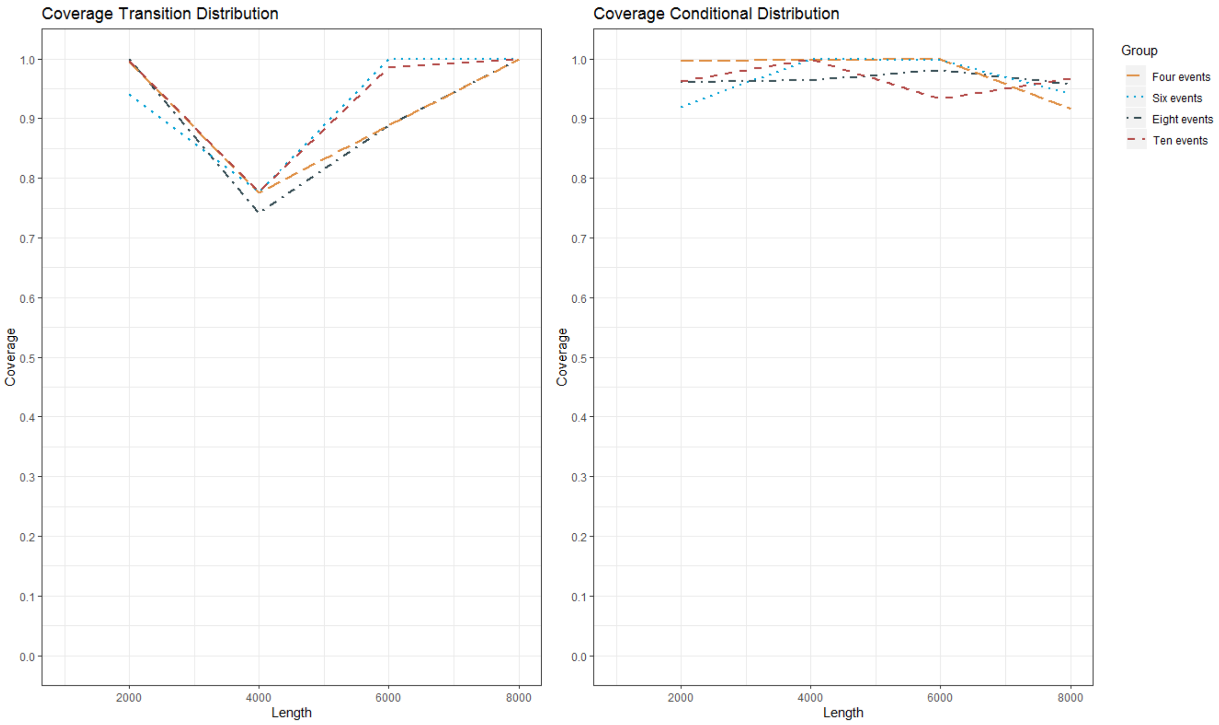


Figure 14: Coverage conditional and transition probability distributions for all scenarios. The length of the associated sequence is shown on the X-axis, the coverage - transition probabilities on the left, conditional probabilities on the right - is shown on the Y-axis. Grouping is by event type.

The reader is therefore referred to figure 15 on the next page, which shows the average size of the credibility interval for both the transition and conditional probabilities. Upon inspection of the left figure, it becomes clear that the Gibbs samplers' estimates of the transition probabilities become more precise as L increases. However, as indicated by the coverage, the average size of its credibility interval is too narrow to permit 95% of the Gibbs samplers' estimates to fall inside it for sequence lengths 4000 and 6000. As such, the Gibbs sampler model is too certain in its estimates of the transition probabilities for these two length types.

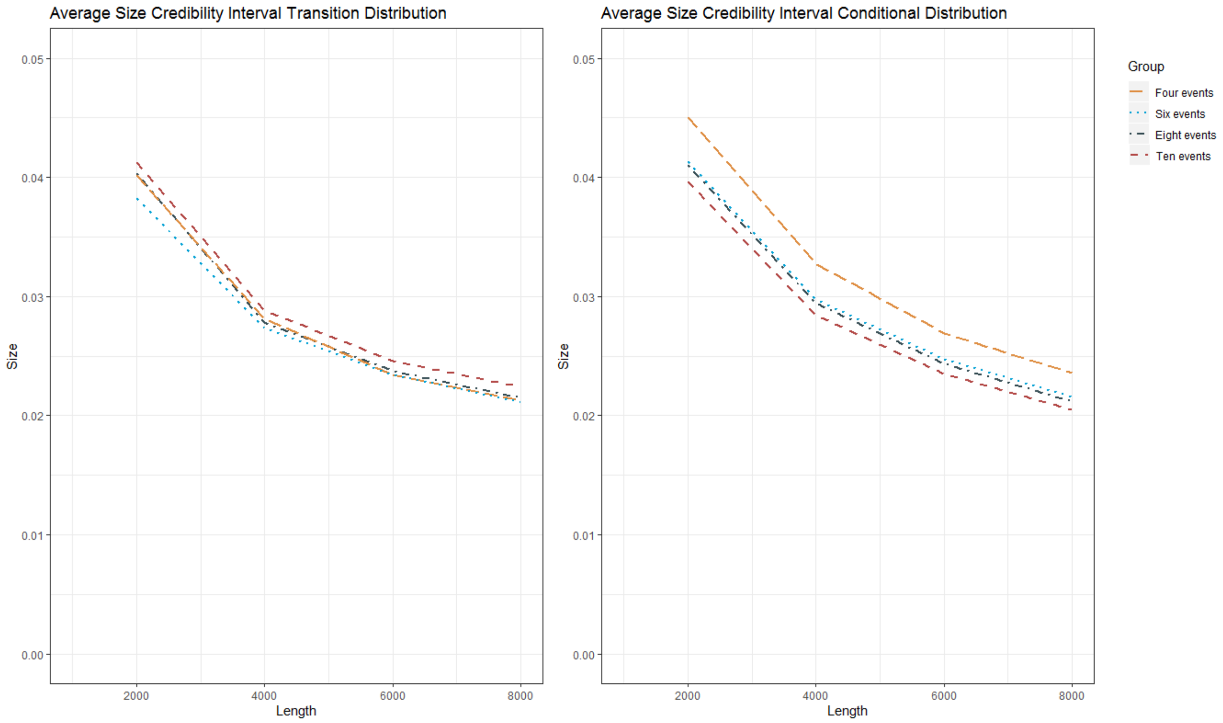


Figure 15: Average size of credibility interval for the conditional and transition probability distributions for all scenarios. The length of the associated sequence is shown on the X-axis, the coverage - transition probabilities on the left, conditional probabilities on the right - is shown on the Y-axis. Grouping is by event type.

The average sizes of the credibility intervals for the transition probabilities are consequently appropriate for lengths 2000 and 8000. Although the accuracy of the estimator values for length 2000 are generally inferior to those of the other sequence lengths, the Gibbs sampler appropriately adjusts the width of the credibility interval for over 95% of its estimates to fall within it. The estimates for length 8000 are finally both highly accurate and precise. It has the lowest average credibility interval and the highest estimator accuracy, while simultaneously permitting over 95% of its estimator values to fall within the credibility interval. There are furthermore no differences in the average size of the credibility interval for the transition probabilities as a result of the number of event types. Inspection of the graph on the right finally shows that as the the length of the sequence increases, the average size of the credibility interval for the conditional probabilities decrease quite sharply. Similarly to the transition probabilities, there exist no substantial differences for the average size of the credibility interval between event types. Due to the fact that the coverage of the conditional probabilities were high for all lengths and event types scenarios, the precision of these estimates can be said to generally improve for increases in L.

Conclusion & discussion

This study has examined how the length of, and the number of event types inherent to the event observation sequence affect the Gibbs samplers' capacity to produce accurate HMM estimates. It moreover has sought to explicate for which single and combined value input ranges these estimates are the most optimal. With regards to the first query, the following research question was formulated: How do increases in the number of event types, and the length of the event observation sequence affect the accuracy of the Gibbs samplers' estimates? The associated hypothesis stated that increases in both variables would result in improved HMM Gibbs sampler parameter estimates. The results consequently indicated that for the transition parameter probabilities, the effect of the number of event types was redundant, whereas for the length of the input sequence it had a notable, albeit relatively nominal effect. For the conditional parameter probabilities, increases in the number of event types did not result in improved estimator values. To the contrary, the lowest of all evaluated event types, namely four, produced the best Gibbs sampler estimates with regards to these parameters. An increase in the length of the observation sequence furthermore had a continually beneficial effect on the estimates of the conditional probabilities. The first hypothesis is therefore rejected for the number of event types variable, whereas it is confirmed for the length of the input sequence. As such, this study concludes that increases in the sequence length have a continouous beneficial effect on the estimates of the Gibbs sampler for both the transition and conditional probabilities. Increases in the number of event types are however redundant to improving Gibbs sampler estimates for both the transition and conditional HMM parameters.

The second research question consequently sought to explicate for which single and combined value input ranges the Gibbs samplers' HMM approximations were the most optimal. With regards to the single variable optimal input value, this value was hypothesized to be situated in the lower but not the lowest bounds of the event type variable, and the upper quartile for the length of the input sequence. The results indicated that the four event type category best enabled the Gibbs sampler to produce optimal estimates for both the transition and conditional parameter probabilities. However, due to the fact that the three event type category had to be removed from the analysis, it was unfortunately not possible to investigate the implications of a transition between the three and four event type categories. As such it cannot definitively be assessed whether the lower bound of the variable is more beneficial to estimation as opposed to its lowest bound. Furthermore, recall that the effect of all event type variables was redundant for improving estimation accuracy for the transition probabilities. The reason why the results nonetheless indicate that the four event type is the most optimal value, is a consequence of the fact that it imposes the least computational strain on the Gibbs sampler procedure from among the set of event types.

The results furthermore indicate an input sequence length of 8000 to be the optimal value for producing optimized HMM Gibbs sampler estimates. Although input lengths of 4000 and 6000 were generally also able to produce acceptable parameter estimates, both the accuracy and precision of the Gibbs samplers' estimates for length 8000 were sufficiently superior for it to be the natural length of choice. The hypothesis is therefore preliminarily accepted for the event type variable, and confirmed for the length of the event sequence variable. As such, this study concludes that the individual optimal input values for the Gibbs sampler - given a three state system with distinct underlying parameter distributions - are equal to four for the observation event type variable, and 8000 for the sequence length variable.

With regards to the unconstrained combined optimal input value for these two variables, the optimal value ranges were hypothesized to be situated in the intersection between the lower bound values of the number of event types, and the upper quartile values for the length of the observed event sequence. As such, one would expect optimal HMM estimates to result from a Gibbs sampler input configuration consisting of four observation event types, and an event observation length of 8000. Recall that the results indicated the conditional probability parameters to be best estimated by precisely these input specifications. It has furthermore already been established that an input sequence length of 8000 produces optimal estimates for the transition probability parameters, and that the four event type variable similarly represents the optimal input specification for estimating these parameters on account of computational efficiency. The unconstrained optimization hypothesis is therefore confirmed. This study consequently concludes the optimal combined Gibbs sampler input specification - for a three event system with distinct underlying parameter distributions - to consist of a value of four for the event observation type variable, and a value of 8000 for the sequence input length variable. The constrained combined optimal input value hypothesis finally states that, given the constraint, estimates will be maximized in accordance with the optimal individual variable values. The results indicate that, given any particular constrained event type variable, a value of 8000 for the length of the event sequence will represent the superior input specification. The bias, RMSEA, coverage and average credibility interval all support the notion that for any given event type, a sequence input length of 8000 will produce the most accurate and precise estimates of the parameters of the HMM. Given any particular constrained input sequence length, the results show that, in general, the four and six event type categories are superior to the eigth and ten event type categories for estimating transition parameter probabilities. For the conditional parameter probabilities, the four event type variable is however convincingly superior to all other event types over the different sequence length categories.

As such, this study concludes, in the confines of the here defined ranges of the variables of interest, a sequence length of 8000 to represent the optimal input specification for any particular event type, and an event type of four to represent the optimal input specification for any particular sequence length.

In conclusion, this study has established initial tentative references on both the general and optimal input relations between the here presented variables of interest, and the functioning of the Gibbs sampler. Its main findings indicate - for a three state system with relatively distinct HMM parameter probabilities, and in the here defined value ranges of the variables of interest - the optimal Gibbs sampler input configuration to consist of four event types, and an associated sequence length of 8000. This finding generalizes to single variable optimal inputs, combined optimal inputs, and contextual optimal inputs. As such, this study provides support to the extant literature that single long runs and increases in pattern variability are beneficial factors to the Gibbs samplers' optimal functioning. It moreover provides some nuance to the literature, in that although introducing pattern variation is beneficial to the Gibbs samplers' estimates, it does not seem to extend beyond the lower bounds of the event type variable.

The principal drawback of this study pertains to the fact that nine of the twenty-five HMM scenarios had to be excluded from the analysis due to severe label switching. In large part this was a consequence of misspecification by the researcher, and in some part was a result of the specified starting values for the 1000 length types requiring more highly skewed steering distributions than originally anticipated. A second, methodological drawback of this study related to the fact that unequal samples were drawn for constructing the ensemble sample parameter estimates of the different scenarios. This could potentially have skewed results somewhat for the more computationally intensive scenario contexts. Replication of the here presented work with properly specified starting value specifications, and Gibbs samples of about 500 for each scenario, could therefore first provide feedback on whether its results are replicable and generally valid. It could additionally enable the elucidation of one of the original objectives of this paper, namely whether the increase from the three to four event types, and from lengths 1000 to 2000, are as beneficial to the Gibbs samplers' functioning as indicated by the literature.

The here presented results consequently suggest a number of future research avenues. One of this study's major findings indicated one additional event type relative to the number of states to be the most beneficial to the Gibbs samplers' functioning. However, does this finding also translate to HMM structures which consist of more than three states? Put differently, does there exist a general mechanism where one additional event type relative to the number of states in the system will introduce an optimized amount of pattern variation into the event observation sequence?

Or does this relation behave differently, as per an exponential growth relation for example, where increases in the number of states necessitate continually increasingly increases in the number of event types? The second major finding of this study indicated a length of 8000 to be the universal optimal Gibbs sampler input value for the sequence length variable. Is this 8000 length however also the optimal value in the context of an extended scale, i.e., does it constitute the actual optimization threshold, or is this threshold situated at an even higher value? Recall that the effect of this variable was furthemore hypothesized to have a logarithmic growth curve, i.e., should at one point start to become redundant to the improvement of estimator values. At what point do additional increases in the sequence length therefore actually start to display this redundancy? It could moreover also be of interest to asses the role of additional event observation sequence variables. One of the most intuitive and relevant would be the shape of the probability distributions of the transition and conditional parameters of the HMM. The central query of interest would than relate to the question of whether uniform parameter distributions have a more or less beneficial effect on the Gibbs samplers' estimates as opposed to more distinctly specified distributions. The role of this variable could consequently be contextualized with that of the here considered variables. For instance, recall that the effect of the event type variable on improving estimator accuracy was redundant for the transition parameter probabilities. It could hower be the case that increases in the pattern variation were made redundant exactly due to the high distinction of the probability distribution of this parameter. It would consequently be of interest to assess whether increases in the pattern variation have a more notable impact on the Gibbs samplers' functioning, as opposed to more uniform probability distributions. This notion could likewise be extended to the conditional parameters of the HMM. The behaviour of the sequence length variable could consequently also be contextualized in regards to this query. Are initial increases in the length more notable in contexts with highly uniform transition and conditional parameter probability distributions for example? And are its optimal and redundant values situated on different points of the variable's scale in such contexts? Finally, the constrained and unconstrained combined variable hypotheses could also be assessed with inclusion of the parameter distribution shape variable, so as to explicate a wider array of both absolute and contextual optimized Gibbs sampler inputs. This would consequently provide researchers with a broad set of references and guidelines on general and optimal Gibbs sampler input specifications for single sequence HMM learning.

References

Aarts, E. (2016). *Beyond the average. Choosing and improving statistical methods to optimize inference from complex neuroscience data* (Doctoral dissertation). Utrecht University, Utrecht, The Netherlands.

Banks, J., Carson, J. S., Nelson, B. L., & Nicol, D. (2010). *Discrete-Event System Simulation.* (5 ed.) Upper Saddle River, NJ: Prentice Hall.

Brooks, S., Gelman, A., Jones, G., & Meng, X. L. (Eds.). (2011). *Handbook of markov chain monte carlo.* CRC press.

Cappé, O., Moulines, E., & Rydén, T. (2009, June). Inference in hidden markov models. *In Proceedings of EUSFLAT conference* (pp. 14-16).

Chen, M. Y., Kundu, A., & Zhou, J. (1994). Off-line handwritten word recognition using a hidden Markov model type stochastic network. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5), 481-496.

Chen, M. H., & Schmeiser, B. (1993). Performance of the Gibbs, hit-and-run, and Metropolis samplers. *Journal of computational and graphical statistics, 2*(3), 251-272.

Chudova, D., & Smyth, P. (2002, July). Pattern discovery in sequences under a markov assumption. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 153-162). ACM.

Eddy, S. R. (1996). Hidden markov models. *Current opinion in structural biology*, 6(3), 361-365.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England), 14*(9), 755-763.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS computational biology, 7*(10), e1002195.

Fine, S., Singer, Y., & Tishby, N. (1998). The hierarchical hidden Markov model: Analysis and applications. *Machine learning, 32*(1), 41-62.

Flexerand, A., Dorffner, G., Sykacekand, P., & Rezek, I. (2002). An automatic, continuous and probabilistic sleep stager based on a hidden Markov model.
Applied Artificial Intelligence, 16(3), 199-207.

Gagniuc, P. A. (2017). *Markov chains: from theory to implementation and experimentation.* John Wiley & Sons.

Gao, J., & Johnson, M. (2008, October). A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 344-352). Association for Computational Linguistics.

Hassan, M. R., & Nath, B. (2005, September). Stock market forecasting using hidden Markov model: a new approach. *In 5th International Conference on Intelligent Systems Design and Applications (ISDA'05)* (pp. 192-196). IEEE.

He, Y., & Kundu, A. (1991, April). Shape classification using hidden markov model. In [Proceedings] ICASSP 91: 1991 *International Conference on Acoustics, Speech, and Signal Processing* (pp. 2373-2376). IEEE.

Hsu, P. L., & Robbins, H. (1947). Complete convergence and the law of large numbers. *Proceedings of the National Academy of Sciences of the United States of America, 33*(2), 25.

Hughes, J. P., Guttorp, P., & Charles, S. P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 48*(1), 15-30.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing (Vol. 3).* London: Pearson.

Lehmann, E. L., & Casella, G. (2006). *Theory of point estimation.* Springer Science & Business Media.

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

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine.*

Rabiner, L. R., & Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE assp magazine, 3*(1), 4-16.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286.

Raftery, A. E., & Lewis, S. M. (1992). [Practical Markov Chain Monte Carlo]: comment: one long run with diagnostics: implementation strategies for Markov Chain Monte Carlo. *Statistical science, 7*(4), 493-497.

Rydén, T. (2008). EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Analysis, 3*(4), 659-688.

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.

Van Helden, J., André, B., & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of molecular biology, 281*(5), 827-842.

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.

Yamato, J., Ohya, J., & Ishii, K. (1992, June). Recognizing human action in time-sequential images using hidden markov model. *In Proceedings 1992 IEEE Computer Society conference on computer vision and pattern recognition* (pp. 379-385). IEEE.

Yildirim, I. (2012). Bayesian inference: Gibbs sampling. *Technical Note, University of Rochester.*

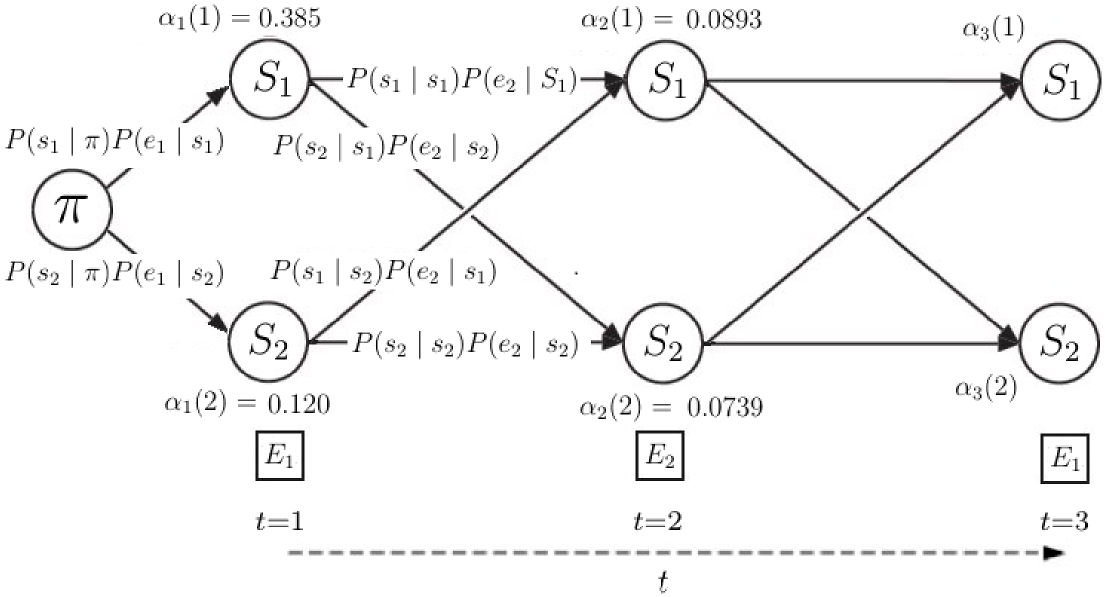
Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging, 20*(1), 45-57.

Appendix A - Likelihood and decoding

Recall that Rabiner (1989) identified three problems that need solving in order for the HMM to be applicable in real-world applications:

Problem 1 (Likelihood / Filtering): Given an observed event sequence $E = \{E_1, E_2, ..., E_T\}$, and an HMM model $\lambda = (\Gamma, \theta, \pi)$, how does one efficiently compute $P(E \mid \lambda)$?
Problem 2 (Decoding / Smoothing): Given the observed event sequence $E = \{E_1, E_2, ..., E_t\}$, and an HMM model $\lambda = (\Gamma, \theta, \pi)$, how does one choose a corresponding state sequence $S = \{S_1, S_2, ..., S_T\}$ which is optimal in some meaningful sense?
Problem 3 (Learning / Training): How does one adjust the model parameters $\lambda = (\Gamma, \theta, \pi)$, so as to maximize $P(E \mid \lambda)$?

The first problem can be adressed with the forward algorithm (Jurafsky & Martin, 2014; Rabiner, 1989). The forward algorithm computes the observation probability of an event sequence by summing over the probabilities of all possible hidden state paths that could function to generate it (Jurafsky & Martin, 2014). It does so by implicitly folding each path into a single forward trellis, as pictured below in figure 16.



$\alpha_1(1) = 0.385$
$\alpha_2(1) = 0.0893$
$\alpha_3(1)$
S_1
$P(s_1 \mid s_1)P(e_2 \mid S_1)$
S_1
S_1
$P(s_1 \mid \pi)P(e_1 \mid s_1)$
$P(s_2 \mid s_1)P(e_2 \mid s_2)$
π
$P(s_2 \mid \pi)P(e_1 \mid s_2)$
$P(s_1 \mid s_2)P(e_2 \mid s_1)$
S_2
$P(s_2 \mid s_2)P(e_2 \mid s_2)$
S_2
S_2
$\alpha_1(2) = 0.120$
$\alpha_2(2) = 0.0739$
$\alpha_3(2)$
E_1
E_2
E_1
t=1
t=2
t=3
t

Figure 16: Forward trellis depicting the observation likelihood computational procedure for two out of three time points for the arbitrary event sequence $E = \{E_1, E_2, E_1\}$ with associated state set $S = \{S_1, S_2\}$, given λ.

As per the illustration, each cell of the forward algorithm trellis $\alpha_t(j)$ represents the probability of being in state j after the first t observations, given λ (Jurafsky & Martin, 2014):

$$\alpha_t(j) = P(E_1, E_2, ..., E_t, S_t = j \mid \lambda). \qquad (11)$$

For a given state S_j at time t, the value $\alpha_t(j)$ is determined by the argument

$$\alpha_t(j) = \sum_{i=1}^{m} \alpha_{t-1}(i)\gamma_{ij}\theta_j(E_t) \qquad (12)$$

where $\alpha_{t-1}(i)$ represents the forward path probability from the previous time step, γ_{ij} represents the transition probability from the previous state S_i to the current state S_j, and $\theta_j(E_t)$ represents the state observation likelihood of observing E_t given the current state S_j (Jurafsky & Martin, 2014). A formal definition statement of the forward algorithm is provided by the step-wise argument (Jurafsky & Martin, 2014):

(1) Initialization:

$$\alpha_1(j) = \pi_j\theta_j(E_1); \quad 1 \le j \le m;$$

(2) Recursion:

$$\alpha_t(j) = \sum_{i=1}^{m} \alpha_{t-1}(i)\gamma_{ij}\theta_j(E_t); \quad 1 \le j \le m, 1 \le t \le T;$$

(3) Termination:

$$P(E \mid \lambda) = \sum_{i=1}^{m} \alpha_T(i).$$

To set ideas, suppose that the system in figure 5 is adequately characterized by the following parameter sets

$$\pi = \begin{bmatrix} 0.70 & 0.30 \end{bmatrix}, \theta = \begin{bmatrix} 0.55 & 0.25 & 0.20 \\ 0.40 & 0.50 & 0.10 \end{bmatrix}, \Gamma = \begin{bmatrix} 0.85 & 0.15 \\ 0.25 & 0.75 \end{bmatrix}$$

and provides a description of the observed event sequence $E = \{E_1, E_2, E_3\}$. The first step initializes the system, i.e., given the initial probability distribution π and the initial observation E_1, determine the probabilities for $\alpha_1(1)$ and $\alpha_1(2)$:

$$\begin{aligned} \alpha_1(1) &= P(S_1 \mid \pi)P(E_1 \mid S_1) = 0.70 \cdot 0.55 = 0.385; \\ \alpha_1(2) &= P(S_2 \mid \pi)P(E_1 \mid S_2) = 0.30 \cdot 0.40 = 0.120. \end{aligned}$$

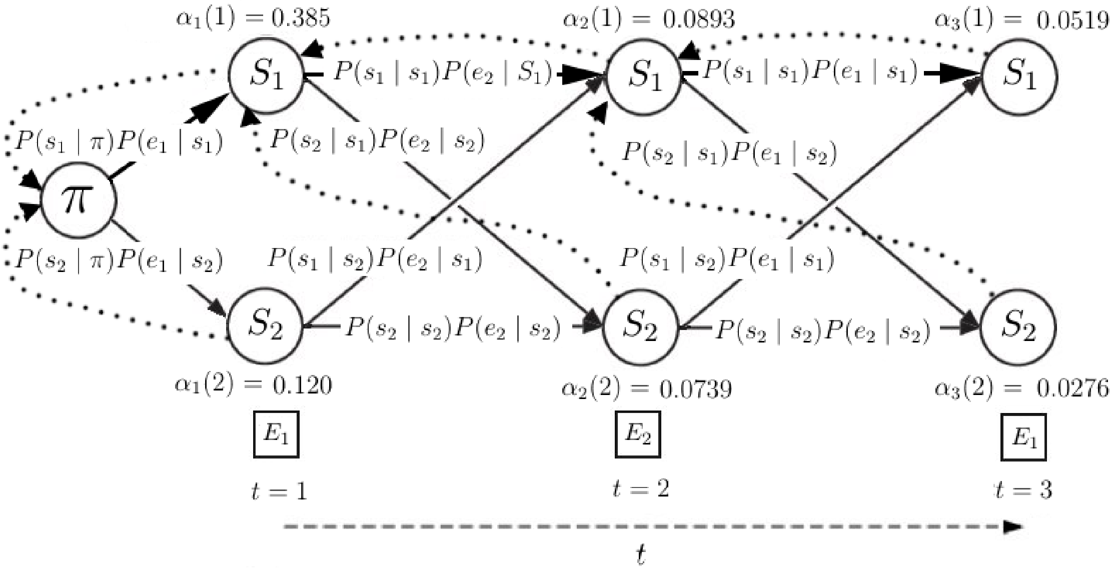
The second step computes the forward probabilities over the time span T by determining $\gamma_{ij}\theta_j(E_t)$ for each of the paths in the system, multiplying the acquired value with the previous path probability α_{t-1}, to consequently sum over the total of respective pathways for each state in current time t:

$$\begin{aligned} \alpha_2(1) &= 0.385 \cdot P(S_1 \mid S_1)P(E_2 \mid S_1) + 0.12 \cdot P(S_1 \mid S_2)P(E_2 \mid S_1) \\ &= 0.385 \cdot (0.85 \cdot 0.25) + 0.12 \cdot (0.25 \cdot 0.25) = 0.0818 + 0.0075 = 0.0893; \\ \alpha_2(2) &= 0.385 \cdot P(S_2 \mid S_1)P(E_2 \mid S_2) + 0.12 \cdot P(S_2 \mid S_2)P(E_2 \mid S_2) \\ &= 0.385 \cdot (0.15 \cdot 0.50) + 0.12 \cdot (0.75 \cdot 0.50) = 0.0289 + 0.045 = 0.0739; \end{aligned}$$

$$\begin{aligned}\alpha_3(1) &= 0.0893 \cdot P(S_1 \mid S_1)P(E_1 \mid S_1) + 0.0739 \cdot P(S_1 \mid S_2)P(E_1 \mid S_1)\\ &= 0.0893 \cdot (0.85 \cdot 0.55) + 0.0739 \cdot (0.25 \cdot 0.55) = 0.0417 + 0.0102 = 0.0519;\\ \alpha_3(2) &= 0.0893 \cdot P(S_2 \mid S_1)P(E_1 \mid S_2) + 0.0739 \cdot P(S_2 \mid S_2)P(E_2 \mid S_2)\\ &= 0.0893 \cdot (0.15 \cdot 0.40) + 0.0739 \cdot (0.75 \cdot 0.40) = 0.0054 + 0.0222 = 0.0276.\end{aligned}$$

The third step relates to the termination of recursion at time T, and gives $P(E \mid \lambda)$ with $\sum_{i=1}^{m} \alpha_T(i) = \alpha_3(1) + \alpha_3(2) = 0.0519 + 0.0276 = 0.0795$. The forward algorithm addionally enables assessment of the probability of being in a certain state at time t given the input sequence E (Jurafsky & Martin, 2014). For instance, what is the probability of S_1 being the active underlying condition at time two, given E? This conditional probability is represented by $P(S_{t=2} = S_1 \mid E_1, E_2)$, where the numerator is given by the forward probability $\alpha_2(1)$, and the denominator is the probability of seeing E given the HMM, i.e., the sum of the forward probabilities at time two $\sum_i \alpha_2(i)$. In the example presented here, the probability of observing state S_1 over state S_2 at $t = 2$ equals $0.0893/(0.0893 + 0.0739) = 0.5472$ or 54.72%.

The second query concerns identifying the state sequence S which provides an optimal description of the observed event sequence E in some predefined sense. The problem is typically adressed with use of the Viterbi algorithm, which is procedurally identical to the forward algorithm except for the fact that it takes the maximum instead of the sum over the previous path probabilities (Jurafsky & Martin, 2014). See figure 17 below for a graphical illustration of the Viterbi algorithm procedure.



$\alpha_1(1) = 0.385$
$\alpha_2(1) = 0.0893$
$\alpha_3(1) = 0.0519$
S_1
$P(s_1 \mid s_1)P(e_2 \mid S_1)$
S_1
$P(s_1 \mid s_1)P(e_1 \mid s_1)$
S_1
$P(s_1 \mid \pi)P(e_1 \mid s_1)$
$P(s_2 \mid s_1)P(e_2 \mid s_2)$
$P(s_2 \mid s_1)P(e_1 \mid s_2)$
π
$P(s_2 \mid \pi)P(e_1 \mid s_2)$
$P(s_1 \mid s_2)P(e_2 \mid s_1)$
$P(s_1 \mid s_2)P(e_1 \mid s_1)$
S_2
$P(s_2 \mid s_2)P(e_2 \mid s_2)$
S_2
$P(s_2 \mid s_2)P(e_2 \mid s_2)$
S_2
$\alpha_1(2) = 0.120$
$\alpha_2(2) = 0.0739$
$\alpha_3(2) = 0.0276$
E_1
E_2
E_1
$t = 1$
$t = 2$
$t = 3$
t

Figure 17: Viterbi trellis depicting the computation of the most probable state sequence (arrows in bold) for the event sequence $E = \{E_1, E_2, E_1\}$, given the state set $S = \{S_1, S_2\}$ and λ. Note that as each path transitions to subsequent states in time, backpointers (dotted arrows) store in memory the most probable path that led to the particular state.

In short, each cell of the Viterbi trellis, $v_t(j)$, represents the probability that the HMM is in state j after t observations and has transitioned according to the most probable state sequence ordening $S_1, ..., S_{t-1}$ (Jurafsky & Martin, 2014). The value for each cell $v_t(j)$ is determined recursively by identifying the most probable path that led to it (Jurafsky & Martin, 2014)

$$v_t(j) = \max_{S_1,...,S_{t-1}} P(S_1, ..., S_{t-1}, E_1, E_2, ..., E_t, S_t = j \mid \lambda) \tag{13}$$

For a given state S_j at time t, the value $v_t(j)$ is consequently computed as

$$v_t(j) = \max_{i=1}^{m} v_{t-1}(i)\gamma_{ij}\theta_j(E_t) \tag{14}$$

where $v_{t-1}(j)$ represents the Viterbi path probability from the previous time step, γ_{ij} represents the transition probability from the previous state S_i to the current state S_j and $\theta_j(E_t)$ represents the state observation likelihood of the observation symbol E_t given the current state S_j (Jurafsky & Martin, 2014). Since the Viterbi algorithm has to not only identify the probability of the observation sequence, but also the likelihood of the underlying state sequence, it employs a backtrace which tracks the most probable state sequence path per each subsequent cell in the recursion computation. The backtrace is retraced after termination of the recursion in order to identify the optimal state sequence (Jurafsky & Martin, 2014). A formal definition statement of the Viterbi algorithm is given by the following step-wise argument as presented by Jurafsky & Martin (2014):

(1) Initialization:

$$\begin{aligned} v_1(j) &= \pi_j\theta_j(E_1); \quad 1 \leq j \leq m \\ bt_1(j) &= 0; \quad 1 \leq j \leq m \end{aligned}$$

(2) Recursion:

$$\begin{aligned} v_t(j) &= \max_{i=1}^{m} v_{t-1}(i)\gamma_{ij}\theta_j(E_t); \quad 1 \leq j \leq m, 1 \leq t \leq T \\ bt_t(j) &= \arg\max_{i=1}^{m} v_{t-1}(i)\gamma_{ij}\theta_j(E_t); \quad 1 \leq j \leq m, 1 \leq t \leq T \end{aligned}$$

(3) Termination:

$$\begin{aligned} \text{Optimal pathway} &\rightarrow P^* = \max_{i=1}^{m} v_T(i) \\ \text{Backtrace} &\rightarrow S_T^* = \arg\max_{i=1}^{m} v_T(i) \end{aligned}$$

Appendix B - The uniform Dirichlet prior

To set ideas, consider the fact that the Dirichlet is a generalization of the Beta distribution into multiple dimensions. The Beta probability density function is of the form (Lynch, 2007)

$$\{x_1, x_2(= 1 - x_1)\} \sim \frac{1}{B(\alpha, \beta)} x_1^{\alpha-1} x_2^{\beta-1} \qquad (16)$$

and is identical to the probability density function of the Dirichlet if $K = 2$. By basing initial interpretations of the Dirichlet on the Beta distribution, one can consequently generalize to the general Dirichlet case where, unlike in the Beta distribution, $K > 2$ is a valid argument. The Beta distribution is a univariate distribution of a random variable $X \in (0, 1)$ parameterized by α and β, and a conjugate prior for binomial parameters, meaning that combining a binomial likelihood function with a prior Beta distribution will result in a posterior that has a Beta distribution (Lynch, 2007). In essence, the function of choosing a conjugate over a non-conjugate prior is one of algebraic convenience (Lynch, 2007). Conjugate priors give closed form posterior expressions, making them less burdensome to compute, and give intuition to the Bayesian process by more transparently showing how a likelihood function updates a prior distribution (Lynch, 2007). Since the Dirichlet is multi-dimensional extenstion of the Beta distribution, it is a conjugate for distributions types that permit variables consisting of multiple discrete states, i.e., a multinomial distribution, meaning that combining a Dirichlet prior with a multinomial likelihood function will result in a posterior with a Dirichlet distribution (Lynch, 2007). For a more in-depth discussion of conjugate priors, the reader is referred to Lynch (2007).

In order to concretize diffuse priors in the context of the Beta distribution, suppose that α and β constitute respective black (A) and white (B) marble draws with replacement from an urn. First consider a scenario in which some prior knowledge exists about the probability distribution of α and β. Assume that it has previously been established that the value that best describes the likelihood for drawing a black marble from the urn x_1 approximates 0.27, but that it could also reasonably be situated somewhere else in the likelihood interval $[0.20, 0.35]$. The inverse naturally holds for x_2 with $(1 - x_1) = (1 - 0.27) = 0.73$, since $\sum_{i=1}^{K} x_i = 1$, with confidence interval [0.64, 0.81]. Given these specifications, the prior Beta distribution can be represented as Beta(α = 81, β = 219), since the mean for $x_1 = (\alpha)/(\alpha + \beta) = (81)/(81 + 219) = 0.27$, the inverse value of which can be determined for x_2 when β replaces α in the numerator. and which has a distribution that is almost entirely dispersed over [0.20, 0.35] (see figure 18 on the next page). Note that the values for α and β represent pseudocounts for A and B, i.e., these draws have not been formally observed, but implicitly assume that prior the consideration of any empirical evidence, eighty-one marbles A and two-hundred and nine-teen marbles B have already been drawn from the urn.

Their function is to assign probability mass to each of the respective binomial events, to express a degree of belief about their expected frequencies (Lynch, 2007). Now suppose that a single black marble (A) is drawn with replacement from the urn, which is subsequently incorporated in the binomial belief assesment. The resulting updated posterior Beta distribution, given the specified prior distribution and the binomial likelihood of the data given the prior, will be Beta($\alpha = 81 + A$, $\beta = 219 + B$) = Beta($\alpha = 81 + 1 = 82$, $\beta = 219 + 0 = 219$). One can consequently redefine the binomial probabilities $x_1 = (82)/(82 + 219) = 0.2724$ and $x_2 = (1 - 0.2724) = 0.7276$, where the probability for x_1 has shifted slightly to the right since the evidence has assigned it additional probability mass. The more evidence one has, the more the curve will shift to accommodate it, and the more it will narrow due to the added certainty of additional proof (Lynch, 2007). If one would draw an additional three-hundred marbles from the urn, resulting in one-hundred and two white (A) and one-hundred and ninety-eight black (B) marble draws, this would translate into the posterior Beta distribution Beta($\alpha = 82 + 102$, $\beta = 219 + 198$) = Beta($\alpha = 184$, $\beta = 417$) with binomial probabilities $x_1 = (184)/(184 + 417) = 0.3062$, $x_2 = 1 - 0.3062 = 0.6938$. Given enough evidence, one will ultimately approach the true value for the binomial probability distribution, which are taken to equal $x_1 = 1/3$ and $x_2 = 2/3$ in this particular example. The argument holds since the posterior probabilities of the evidence based Beta($\alpha = 184$, $\beta = 417$) with $x_1 = 0.3062$ and $x_2 = 0.6938$ approximate these true values more closely than the binomial probabilities of the original prior Beta($\alpha = 81$, $\beta = 219$) with $x_1 = 0.27$ and $x_2 = 0.73$. Updating the original prior with evidence regarding the relative frequency of marble draws has therefore resulted in an improved estimate of its bimomial probability. See figure 7 below for a curve comparison of Beta($\alpha = 81$, $\beta = 219$) and Beta($\alpha = 184$, $\beta = 417$).

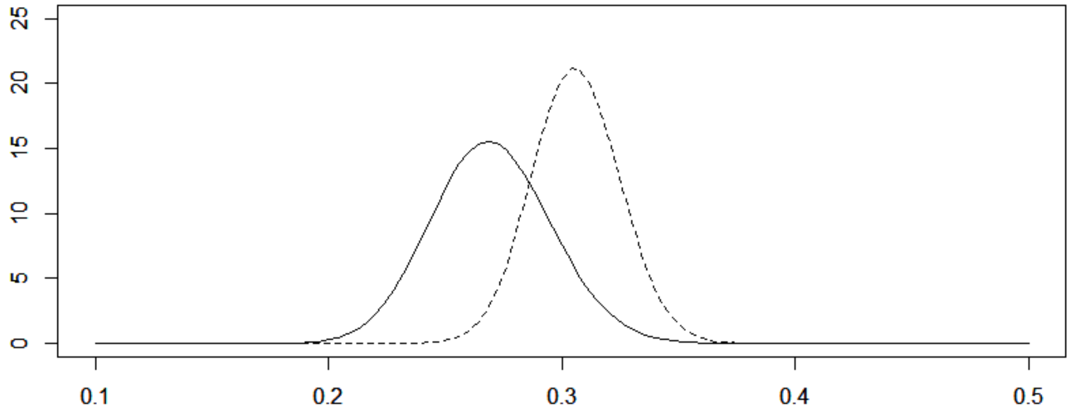


Figure 18: Curve comparison of Beta($\alpha = 81$, $\beta = 219$), graphed in solid line, and Beta($\alpha = 184$, $\beta = 417$), graphed in dashed line.

Now consider a situation wherein there exists no prior information regarding the probability distribution of α and β. In that case the default argument is to assume that the probabilities x_1 and x_2 are equivalent to one another (Lynch, 2007). A straightforward way to represent this probability set is to let $\alpha = \beta = 1$, since $x_1 = (1)/(1+1) = 0.50$, where the inverse holds for x_2. Such a discrete, uniform probability distribution has a symmetric shape, i.e., the probability mass function is reflected around a vertical line, and is uniform in the sense that every one of n values has an equal probability of occurence $1/n$ (Lynch, 2007). Note that the values for α and β again constitute pseudocounts, it is assumed that prior any evidence input, a single draw has been observed for each marble type. The uniform Beta prior can consequently be characterized as Beta($\alpha + A = 1 + A, \beta + B = 1 + B$), so that the only element influencing the belief assesment regarding the distribution of α and β is empirical evidence in the form of data input. Note that it will generally require more empirical evidence to approach the true binomial probability values in the case of a diffuse Beta prior as opposed to a more strictly specified Beta prior.

Now extend the Beta-binomial case to the Dirichlet-multinomial case. Instead of parameterizing $K = 2$ variables, define a parameterization for $K \geq 2$ variables, denoted by $\alpha_1, \alpha_2, ..., \alpha_k$. Analogous to the previous example, instead of drawing black and white marbles, one can now draw N marbles appearing in K colours from the Dirichlet-multinomial, with associated probabilities $x_1, x_2, ..., x_K$. The parameters $\alpha_1, ..., \alpha_K$ can again be thought as apriori pseudocounts of marbles of each color, which are updated by summing observed counts for each category $\alpha_1 + n_1, ...\alpha_k + n_k$. One can consequently assign probability mass to each α_k, so as to define one's prior degree of belief about their respective expected frequencies. The diffuse Dirichlet is than defined analogously to the diffuse Beta distribution, so that it is symmetric and uniform with $\alpha_1 = ... = \alpha_k = 1$. See figure 19 below for a set of trivariate Dirichlet distributions with differing parameter specifications.

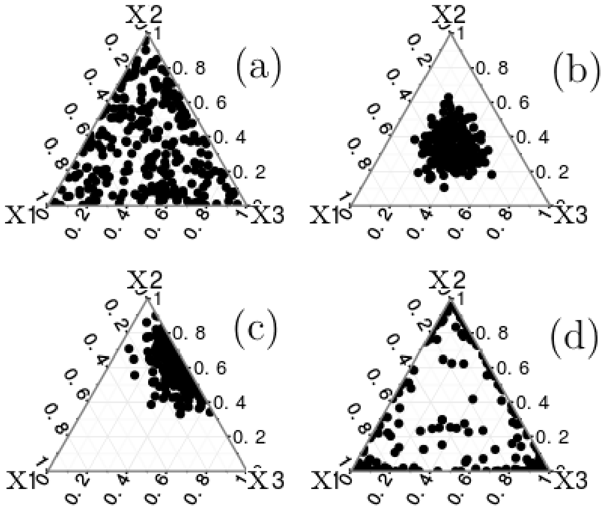


Figure 19: Trivariate Dirichlet distributions, parameterized by (a) $\alpha_1 = \alpha_2 = \alpha_3 = 1$, (b) $\alpha_1 = \alpha_2 = \alpha_3 = 10$, (c) $\alpha_1 = 1, \alpha_2 = 10, \alpha_3 = 5$, (d) $\alpha_1 = \alpha_2 = \alpha_3 = 0.2$. Note the diffuse range of values in (a).