

# Title

**Bart-Jan Boverhof\***

\*bjboverhof@gmail.com

## Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum

## 1 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<sup>1-3</sup>. 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<sup>3</sup>. 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<sup>4</sup>. 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<sup>3,5</sup>, human action and shape classification<sup>6,7</sup>, computational molecular biology<sup>8</sup>, brain MR image segmentation<sup>9</sup>, precipitation occurrence<sup>10</sup> and stock market forecasting<sup>11,12</sup>.

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<sup>13,14</sup>. 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<sup>3</sup>. 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<sup>3,13</sup>. 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<sup>15,16</sup>. In short, MCMC sampling algorithms enable researchers to calculate numerical approximations of multi-dimensional integrals<sup>17</sup>. 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<sup>17</sup>.

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<sup>15,16</sup>. Despite the widespread usage of the HMM, still little is known about the character and distribution of the data that determines the effectiveness of this modelling approach. Elemental guidelines concerning model furnishing are until today non-existent within the literature. In this paper we conduct a simulation study, with which we aim to gain insight into the prerequisites for being able to accurately describe intensive longitudinal data by means a HMM. In the conducted simulations we specifically vary (1) the length of the time series, (2) the degree to which a clear distinction between the states exists, (3) the overlap between the hidden states (4) and the amount of observations. Based on these insights we aim to provide a general outline of the required model and data specifications for various research scenarios, and hereby provide practical guidelines that may be utilised by researchers in order to appropriately shape their project.

## 2 Methodology

### 2.1 The Hidden Markov Model

A HMM constitutes three building blocks, being the probability matrices of the conditional probability distribution denoted by  $\theta$ , the transitional probability distribution denoted by  $\Gamma$  and the initial probability distribution denoted by  $\pi$ . The application of

the HMM inhibits a trajectory embodying two different facets: (1) deriving the probabilities of the states at each moment in time  $T$ , summarized in the conditional and transition distribution matrices, and (2) optimize these probability distributions by Gibbs sampling. The most probable state sequence is derived by means of the Forward probability algorithm, estimating the most probable state sequences at times  $T_1, T_2, \dots, T_{-1}$ . After determining the most probable state sequence at moment  $T$ , the Gibbs sampling algorithm samples the states in a backwards manner and updates the probabilities expressed in the conditional and transition distribution matrices accordingly<sup>15,16</sup>.

An intuitive example of this characteristic is provided by the analysis of sleep stages with use of the HMM framework. Although sleep stages such as REM sleep, deep sleep and wakefulness are not directly observable, they each generate distinguishable continuous EEG measurements, from which the particular state of consciousness at each time point  $T$  is inferred<sup>18</sup>.

## **2.2 Simulation setup**

## **3 Results**

## **4 Conclusion & Discussion**

### **4.1 Discussion & Future Research**

## References

1. Eddy, S. R. Hidden markov models. *Curr. opinion structural biology* **6**, 361–365 (1996).
2. Eddy, S. R. Profile hidden markov models. *Bioinforma. (Oxford, England)* **14**, 755–763 (1998).
3. Rabiner, L. & Juang, B. An introduction to hidden markov models. *ieee assp magazine* **3**, 4–16 (1986).
4. Aarts, E. Beyond the average: Choosing and improving statistical methods to optimize inference from complex neuroscience data. (2016).
5. Chen, M.-Y., Kundu, A. & Zhou, J. Off-line handwritten word recognition using a hidden markov model type stochastic network. *IEEE transactions on Pattern analysis Mach. Intell.* **16**, 481–496 (1994).
6. Kundu, A. & He, Y. On optimal order in modeling sequence of letters in words of common language as a markov chain. *Pattern Recognit.* **24**, 603–608 (1991).
7. Yamato, J., Ohya, J. & Ishii, K. Recognizing human action in time-sequential images using hidden markov model. In *CVPR*, vol. 92, 379–385 (1992).
8. Eddy, S. R. Accelerated profile hmm searches. *PLoS computational biology* **7**, e1002195 (2011).
9. Zhang, Y., Brady, J. M. & Smith, S. Hidden markov random field model for segmentation of brain mr image. In *Medical Imaging 2000: Image Processing*, vol. 3979, 1126–1137 (International Society for Optics and Photonics, 2000).
10. Hughes, J. P., Guttorp, P. & Charles, S. P. A non-homogeneous hidden markov model for precipitation occurrence. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **48**, 15–30 (1999).
11. Hassan, M. R. & Nath, B. Stock market forecasting using hidden markov model: a new approach. In *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, 192–196 (IEEE, 2005).
12. Mamon, R. S. & Elliott, R. J. *Hidden Markov models in finance*, vol. 4 (Springer, 2007).
13. Visser, I. Seven things to remember about hidden markov models: A tutorial on markovian models for time series. *J. Math. Psychol.* **55**, 403–415 (2011).
14. Gagniuc, P. A. *Markov chains: from theory to implementation and experimentation* (John Wiley & Sons, 2017).
15. Rydén, T. Em versus markov chain monte carlo for estimation of hidden markov models: A computational perspective. *Bayesian Analysis* **3**, 659–688 (2008).
16. Scott, S. L. Bayesian methods for hidden markov models: Recursive computing in the 21st century. *J. Am. statistical Assoc.* **97**, 337–351 (2002).
17. Lynch, S. M. *Introduction to applied Bayesian statistics and estimation for social scientists* (Springer Science & Business Media, 2007).
18. Flexerand, A., Dorffner, G., Sykacekand, P. & Rezek, I. An automatic, continuous and probabilistic sleep stager based on a hidden markov model. *Appl. Artif. Intell.* **16**, 199–207 (2002).