Bayesian Analysis (2023) **18**, Number 2, pp. 547–577

Bayesian Approximations to Hidden Semi-Markov Models for Telemetric Monitoring of Physical Activity

Beniamino Hadj-Amar[*], Jack Jewson[†], and Mark Fiecas[‡]

Abstract. We propose a Bayesian hidden Markov model for analyzing time series and sequential data where a special structure of the transition probability matrix is embedded to model explicit-duration semi-Markovian dynamics. Our formulation allows for the development of highly flexible and interpretable models that can integrate available prior information on state durations while keeping a moderate computational cost to perform efficient posterior inference. We show the benefits of choosing a Bayesian approach for HSMM estimation over its frequentist counterpart, in terms of model selection and out-of-sample forecasting, also highlighting the computational feasibility of our inference procedure whilst incurring negligible statistical error. The use of our methodology is illustrated in an application relevant to e-Health, where we investigate rest-activity rhythms using telemetric activity data collected via a wearable sensing device. This analysis considers for the first time Bayesian model selection for the form of the explicit state dwell distribution. We further investigate the inclusion of a circadian covariate into the emission density and estimate this in a data-driven manner.

Keywords: Markov switching process, Hamiltonian Monte Carlo, Bayes factor, telemetric activity data, circadian rhythm.

1 Introduction

Recent developments in portable computing technology and the increased popularity of wearable and non-intrusive devices, e.g. smartwatches, bracelets, and smartphones, have provided exciting opportunities to measure and quantify physiological time series that are of interest in many applications, including mobile health monitoring, chronotherapeutic healthcare and cognitive-behavioral treatment of insomnia (Williams et al., 2013; Kaur et al., 2013; Silva et al., 2015; Aung et al., 2017; Huang et al., 2018). The behavioral pattern of alternating sleep and wakefulness in humans can be investigated by measuring gross motor activity. Over the last twenty years, activity-based sleep-wake monitoring has become an important assessment tool for quantifying the quality of sleep (Ancoli-Israel et al., 2003; Sadeh, 2011). Though polysomnography (Douglas et al., 1992), usually carried out within a hospital or at a sleep center, continues to remain the gold standard for diagnosing sleeping disorders, accelerometers have become

[*] Department of Statistics, Rice University, TX 77005-1827, Beniamino.Hadj-Amar@rice.edu

[†] Barcelona Graduate School of Economics, Universitat Pompeu Fabra, Barcelona, Spain, 08005, jack.jewson@upf.edu

[‡] Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, mfiecas@umn.edu

https://doi.org/10.1214/22-BA1318

a practical and inexpensive way to collect non-obtrusive and continuous measurements of rest-activity rhythms over a multitude of days in the individual's home sleep environment (Ancoli-Israel et al., 2015).

Our study investigates the *physical activity* (PA) time-series first considered by Huang et al. (2018) and Hadj-Amar et al. (2019), where a wearable sensing device is fixed to the chest of a user to measure its movement via a triaxial accelerometer (ADXL345, Analog Devices). The tool produces PA counts, defined as the number of times an accelerometer undulation exceeds zero over a specified time interval. Figure 1 displays an example of 4 days of 5-min averaged PA recordings for a healthy subject, providing a total of 1150 data points. Transcribing information from such complex, high-frequency data into interpretable and meaningful statistics is a non-trivial challenge, and there is a need for a data-driven procedure to automate the analysis of these types of measurements. While Huang et al. (2018) addressed this task by proposing a hidden Markov model (HMM) within a frequentist framework, we formulate a more flexible approximate hidden semi-Markov model (HSMM) approach that enables us to explicitly model the dwell time spent in each state. Our proposed modeling approach uses a Bayesian inference paradigm, allowing us to incorporate available prior information for different activity patterns and facilitate consistent and efficient model selection between dwell distribution.

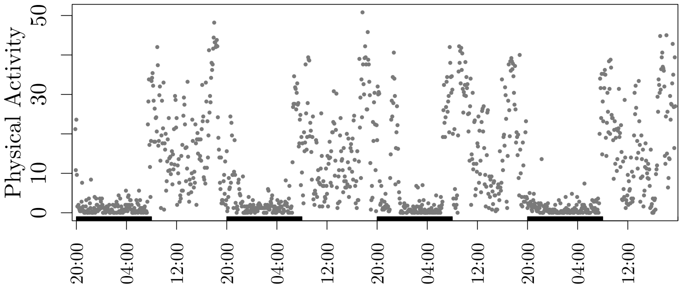


Figure 1: PA time series for a healthy individual. Rectangles on the time axis correspond to periods from 20.00 to 8.00.

We conduct Bayesian inference using a HMM likelihood model that is a reformulation of any given HSMM. We utilize the method of Langrock and Zucchini (2011) to embed the generic state duration distribution within a special transition matrix structure that can approximate the underlying HSMM with arbitrary accuracy. This framework is able to incorporate the extra flexibility of explicitly modeling the state dwell distribution provided by a HSMM, without renouncing the computational tractability, theoretical understanding, and the multitudes of methodological advancements that are available when using an HMM. To the best of our knowledge, such a modeling approach has only previously been treated from a non-Bayesian perspective in the literature, where parameters are estimated either by direct numerical likelihood maximization (MLE) or applying the expectation-maximization (EM) algorithm.

The main practical advantages of a fully Bayesian framework for HSMM inference are that the regularization and uncertainty quantification provided by the prior and posterior distributions can be readily incorporated into improved mechanisms for prediction and model selection. In particular, selecting the HSMM dwell distribution in a data-driven manner and performing predictive inference for future state dwell times.

However, the posterior distribution is rarely available in closed form and the computational burden of approximating the posterior, often by sampling (see e.g. Gelfand and Smith, 1990), is considered a major drawback of the Bayesian approach. In particular, evaluating the likelihood in HSMMs is already computationally burdensome (Guédon, 2003), yielding implementations that are often prohibitively slow. This further motivates the use of the likelihood approximation of Langrock and Zucchini (2011) within a Bayesian framework. Here, we combine their approach with the *stan* probabilistic programming language (Carpenter et al., 2016), further accelerating the likelihood evaluations by proposing a sparse matrix implementation and leveraging *stan*'s compatibility with bridge sampling (Meng and Wong, 1996; Meng and Schilling, 2002; Gronau et al., 2020) to facilitate Bayesian model selection. We provide examples to illustrate the statistical advantages of our Bayesian implementation in terms of prior regularization, forecasting, and model selection and further illustrate that the combination of our approaches can make such inferences computationally feasible (for example, by reducing the time for inference from more than three days to less than two hours), whilst incurring negligible statistical error.

The rest of this article is organized as follows. In Section 2.1, we provide a brief introduction to HMMs and HSMMs. Section 2.2 reviews the HSMM likelihood approximation of Langrock and Zucchini (2011). Section 3 presents our Bayesian framework and inference approach. Using several simulation studies, Section 4 investigates the performance of our proposed procedure when compared with the implementation of Langrock and Zucchini (2011). Section 5 evaluates the trade-off between computational efficiency and statistical accuracy of our method and proposes an approach to investigate the quality of the likelihood approximation for given data. Section 6 illustrates the use of our method to analyze telemetric activity data, and we further investigate the inclusion of spectral information within the emission density in Section 6.1. The *stan* files (and `R` utilities) that were used to implement our experiments are available at `https://github.com/Beniamino92/BayesianApproxHSMM`. The probabilistic programming framework associated with *stan* makes it easy for practitioners to consider further dwell/emission distributions to the ones considered in this paper. Users need only change the corresponding function in our *stan* files.

2 Modeling Approach

2.1 Overview of Hidden Markov and Semi-Markov Models

We now provide a brief introduction to the standard HMM and HSMM approaches before considering the special structure of the transition matrix presented by Zucchini et al. (2017), which allows the state dwell distribution to be generalized with arbitrary accuracy. HMMs, or Markov switching processes, have been shown to be appealing models in addressing learning challenges in time series data and have been successfully applied in fields such as speech recognition (Rabiner, 1989; Jelinek, 1997), digit recognition (Raviv, 1967; Rabiner et al., 1989) as well as biological and physiological data (Langrock et al., 2013; Huang et al., 2018; Hadj-Amar et al., 2021). An HMM is a stochastic process model based on an unobserved (hidden) state sequence $\boldsymbol{s} = (s_1, \ldots, s_T)$ that takes discrete values in the set $\{1, \ldots, K\}$ and whose transition probabilities follow a Markovian structure. Conditioned on this state sequence, the observations $\boldsymbol{y} = (y_1, \ldots, y_T)$ are assumed to be conditionally independent and generated from a parametric family of probability distributions $f(\boldsymbol{\theta}_j)$, which are often called *emission* distributions. This generative process can be outlined as

$$\begin{aligned} s_t \,|\, s_{t-1} &\sim \boldsymbol{\gamma}_{s_{t-1}}, \\ y_t \,|\, s_t &\sim f\,(\boldsymbol{\theta}_{s_t}), \qquad t = 1, \ldots, T, \end{aligned} \tag{2.1}$$

where $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jK})$ denotes the state-specific vector of transition probabilities, $\gamma_{jk} = p\,(\,s_t = k \,|\, s_{t-1} = j)$ with $\sum_k \gamma_{jk} = 1$, and $p\,(\cdot)$ is a generic notation for probability density or mass function, whichever appropriate. The initial state s_0 has distribution $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0K})$ and $\boldsymbol{\theta}_j$ represents the vector of emission parameters modeling state j. HMMs provide a simple and flexible mathematical framework that can be naturally used for many inference tasks, such as signal extraction, smoothing, filtering and forecasting (see e.g. Zucchini et al. 2017). These appealing features are a result of an extensive theoretical and methodological literature that includes several dynamic programming algorithms for computing the likelihood in a straightforward and inexpensive manner (e.g. forward messages scheme, Rabiner 1989). HMMs are also naturally suited for local and global decoding (e.g. Viterbi algorithm, Forney 1973), and the incorporation of trend, seasonality and covariate information in both the observed process and the latent sequence. Although computationally convenient, the Markovian structure of HMMs limits their flexibility. In particular, the *dwell* duration in any state, namely the number of consecutive time points that the Markov chain spends in that state, is implicitly forced to follow a geometric distribution with probability mass function $p_j(d) = (1 - \gamma_{jj})\,\gamma_{jj}^{d-1}$.

A more flexible framework can be formulated using HSMMs, where the generative process of an HMM is augmented by introducing an explicit, state specific, form for the dwell time (Guédon, 2003; Johnson and Willsky, 2013). The state stays unchanged until the duration terminates, at which point there is a Markov transition to a new regime. As depicted in Figure 2, the *super-states* $\boldsymbol{z} = (z_1, \ldots, z_S)$ are generated from a Markov chain prohibiting self-transitions wherein each super-state z_s is associated with a dwell time d_s and a random segment of observations $\boldsymbol{y}_s = (y_{t_s^1}, \ldots, y_{t_s^2})$, where $t_s^1 = 1 + \sum_{r<s} d_r$ and $t_s^2 = t_s^1 + d_s - 1$ represent the first and last index of segment s, and S is the (random) number of segments. Here, d_s represents the length of the dwell duration of z_s. The generative mechanism of an HSMM can be summarized as

$$\begin{aligned} z_s \,|\, z_{s-1} &\sim \boldsymbol{\pi}_{z_{s-1}}, \\ d_s \,|\, z_s &\sim g\,(\boldsymbol{\lambda}_{z_s}), \\ \boldsymbol{y}_s \,|\, z_s &\sim f\,(\boldsymbol{\theta}_{z_s}), \qquad s = 1, \ldots, S, \end{aligned} \tag{2.2}$$

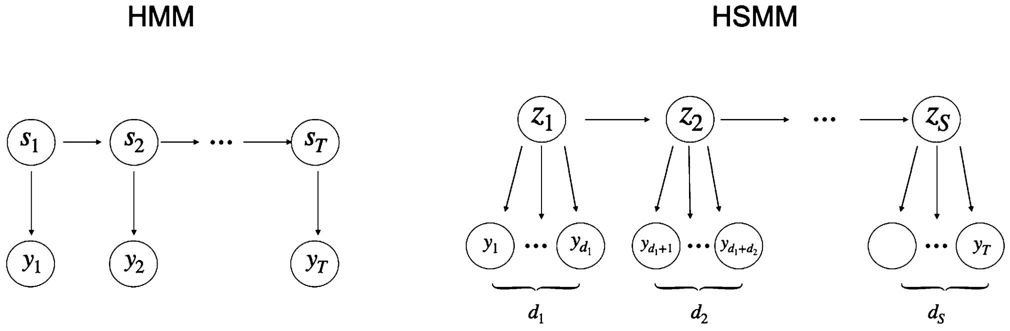


Figure 2: Graphical models: (left) HMM where $y_1, \ldots, y_T$ are the observations and $s_1, \ldots, s_T$ the corresponding hidden state sequence; (right) HSMM where $d_1, \ldots, d_S$ are the random dwell-times associated with each *super* state of the Markov chain $z_1, \ldots, z_S$ where no self-transitions are allowed.

where $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jK})$ are state-specific transition probabilities in which $\pi_{jk} = p\,(\, z_t = k \,|\, z_{t-1} = j,\ z_t \neq j)$ for $j, k = 1, \ldots, K$. Note that $\pi_{jj} = 0$, since self transitions are prohibited. We assume that the initial state has distribution $\boldsymbol{\pi}_0 = (\pi_{01}, \ldots, \pi_{0K})$, namely $\boldsymbol{z}_0 \sim \boldsymbol{\pi}_0$. Here, g denotes a family of dwell distributions parameterized by some state-specific duration parameters $\boldsymbol{\lambda}_j$, which could be either a scalar (e.g. rate of a Poisson distribution), or a vector (e.g. rate and dispersion parameters for negative binomial durations). Unfortunately, this increased flexibility in modeling the state duration has the cost of substantially increasing the computational burden of computing the likelihood: the message-passing procedure for HSMMs requires $\mathcal{O}\,(T^2K + TK^2)$ basic computations for a time series of length T and number of states K, whereas the corresponding forward-backward algorithm for HMMs requires only $\mathcal{O}\,(TK^2)$.

2.2 Approximations to Hidden Semi-Markov Models

In this section we introduce the HSMM likelihood approximation of Langrock and Zucchini (2011). Let us consider an HMM in which $\boldsymbol{y}^\star = (y_1^\star, \ldots, y_T^\star)$ represents the observed process and $\boldsymbol{z}^\star = (z_1^\star, \ldots, z_T^\star)$ denotes the latent discrete-valued sequence of a Markov chain with states $\{1, 2, \ldots, \bar{A}\}$, where $\bar{A} = \sum_{i=1}^{K} a_i$, and $a_1, \ldots, a_K$ are arbitrarily fixed positive integers. Let us define *state aggregates* A_j as

$$A_j = \left\{ a : \sum_{i=0}^{j-1} a_i < a \leq \sum_{i=0}^{j} a_i \right\}, \quad j = 1, \ldots, K, \tag{2.3}$$

where $a_0 = 0$, and each state corresponding to A_j is associated with the same emission distribution $f(\,\boldsymbol{\theta}_j)$ in the HSMM formulation of (2.2), namely $y_t^\star \,\big|\, z_t^\star \in A_j \sim f\,(\boldsymbol{\theta}_j)$. The probabilistic rules governing the transitions between states $\boldsymbol{z}^\star$ are described via the matrix $\boldsymbol{\Phi} = \big\{\phi_{il}\big\}$, where $\phi_{il} = p\,(\, z_t^\star = l \,|\, z_{t-1}^\star = i\,)$, for $i, l = 1, \ldots, \bar{A}$. This matrix has the following structure

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_{11} & \dots & \mathbf{\Phi}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{\Phi}_{K1} & \dots & \mathbf{\Phi}_{KK} \end{bmatrix}, \tag{2.4}$$

where the sub-matrices $\mathbf{\Phi}_{jj}$ along the main diagonal, of dimension $a_j \times a_j$, are defined for $a_j \geq 2$, as

$$\mathbf{\Phi}_{jj} = \begin{bmatrix} 0 & 1 - h_j\,(1) & 0 & \dots & & 0 \\ \vdots & 0 & \ddots & & & \vdots \\ & \vdots & & & & 0 \\ 0 & 0 & \dots & & 0 & 1 - h_j\,(a_j - 1) \\ 0 & 0 & \dots & & 0 & 1 - h_j\,(a_j) \end{bmatrix}, \tag{2.5}$$

and $\mathbf{\Phi}_{jj} = 1 - h_j(1)$, for $a_j = 1$. The $a_j \times a_k$ off-diagonal matrices $\mathbf{\Phi}_{jk}$ are given by

$$\mathbf{\Phi}_{jk} = \begin{bmatrix} \pi_{jk}\, h_j\,(1) & 0 & \dots & 0 \\ \pi_{jk}\, h_j\,(2) & 0 & \dots & 0 \\ \vdots & & & \\ \pi_{jk}\, h_j\,(a_j) & 0 & \dots & 0 \end{bmatrix}, \tag{2.6}$$

where in the case that $a_j = 1$ only the first column is included. Here, π_{jk} are the transition probabilities of an HSMM as in (2.2), and the *hazard rates* $h_j\,(r)$ are specified for $r \in \mathbb{N}_{>0}$ as

$$h_j\,(r) = \frac{p\,(\,d_j = r \,|\, \boldsymbol{\lambda}_j)}{p\,(\,d_j \geq r \,|\, \boldsymbol{\lambda}_j)}, \quad \text{if } p\,(\,d_j \geq r - 1 \,|\, \boldsymbol{\lambda}_j) < 1, \tag{2.7}$$

and 1 otherwise, where $p\,(\,d_j = r \,|\, \boldsymbol{\lambda}_j)$ denotes the probability mass function of the dwell distribution $g\,(\boldsymbol{\lambda}_j)$ for state j. This structure for the matrix $\mathbf{\Phi}$ implies that transitions within state aggregate A_j are determined by diagonal matrices $\mathbf{\Phi}_{jj}$, while transitions between state aggregates A_j and A_k are controlled by off-diagonal matrices $\mathbf{\Phi}_{jk}$. Additionally, a transition from A_j to A_k must enter A_k in $\min(A_k)$. Langrock and Zucchini (2011) showed that this choice of $\mathbf{\Phi}$ allows for the representation of any duration distribution, and yields an HMM that is, at least approximately, a reformulation of the underlying HSMM. In summary, the distribution of $\boldsymbol{y}$ (generated from an underlying HSMM) can be approximated by that of $\boldsymbol{y}^\star$ (modeled using $\mathbf{\Phi}$), and this approximation can be designed to be arbitrarily accurate by choosing a_j adequately large. In fact, the representation of the dwell distribution through $\mathbf{\Phi}$ differs from the true distribution, namely the one in the HSMM formulation of (2.2), only for values larger than a_j, i.e., in the right tail.

3 Bayesian Inference

Bayesian inference for HSMMs has long been plagued by the computational demands of evaluating its likelihood. In this section we use the HSMM likelihood approximation

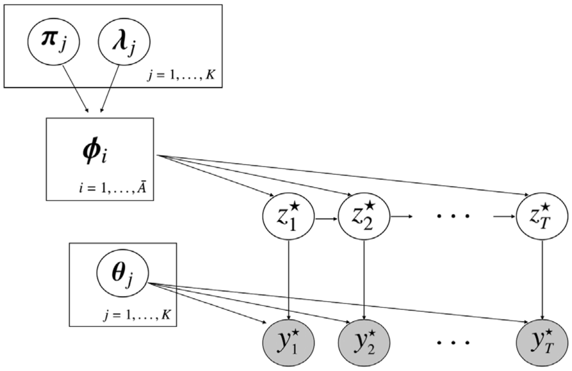


Figure 3: A graphical model for (3.1). Transition probabilities $\boldsymbol{\phi}_j$ are solely determined by $\boldsymbol{\pi}_j$ and $p\,(\,d_j = r \,|\, \boldsymbol{\lambda}_j)$, and thus they are not considered as random variables themselves.

of Langrock and Zucchini (2011) to facilitate efficient Bayesian inference for HSMMs . Extending the model introduced in Section 2.2 to the Bayesian paradigm requires placing priors on the model parameters $\boldsymbol{\eta} = \left\{\, (\boldsymbol{\pi}_j,\, \boldsymbol{\lambda}_j,\, \boldsymbol{\theta}_j)\,\right\}_{j=1}^{K}$. The generative process of our Bayesian model can be summarized by

$$\begin{aligned} \boldsymbol{\pi}_j &\sim \mathrm{Dir}\,(\boldsymbol{\alpha}_0), \qquad (\boldsymbol{\theta}_j, \boldsymbol{\lambda}_j) \sim H \times G, \qquad j = 1, \ldots, K, \\ z_t^\star \,|\, z_{t-1}^\star &\sim \boldsymbol{\phi}_{\,z_{t-1}^\star}, \\ \boldsymbol{y}_t^\star \,|\, z_t^\star \in A_j \;&\sim\; f\,(\,\boldsymbol{\theta}_j), \qquad\qquad\qquad\qquad\qquad t = 1, \ldots, T, \end{aligned} \tag{3.1}$$

where $\mathrm{Dir}(\cdot)$ denotes the Dirichlet distribution over a $(K-2)$ dimensional simplex (since the probability of self transition is forced to be zero) and $\boldsymbol{\alpha}_0$ is a vector of positive reals. Here, H and G represent the priors over emission and duration parameters, respectively, and $\boldsymbol{\phi}_i$ denotes the i^{th} row of the matrix $\boldsymbol{\Phi}$. A graphical model representing the probabilistic structure of our approach is shown in Figure 3, where we remark that the entries of the transition matrix $\boldsymbol{\Phi}$ are entirely determined by the transition probabilities of the Markov chain $\boldsymbol{\pi}_j$ and the values of the durations $p\,(\,d_j = r \,|\, \boldsymbol{\lambda}_j)$.

The posterior distribution for $\boldsymbol{\eta}$ has the following factorization.

$$p\,(\,\boldsymbol{\eta}\,|\,\boldsymbol{y}) \propto \mathscr{L}\,(\boldsymbol{y}\,|\,\boldsymbol{\eta}) \;\times\; \left[\,\prod_{j=1}^{K} p\,(\boldsymbol{\pi}_j) \;\times\; p\,(\boldsymbol{\lambda}_j) \;\times\; p\,(\boldsymbol{\theta}_j)\right], \tag{3.2}$$

where $\mathscr{L}\,(\,\cdot\,)$ denotes the likelihood of the model, $p\,(\boldsymbol{\pi}_j)$ is the density of the Dirichlet prior for transitions probabilities (Eq. 2.2), and $p\,(\boldsymbol{\lambda}_j)$ and $p\,(\boldsymbol{\theta}_j)$ represent the prior densities for dwell and emission parameters, respectively. Since we have formulated an HMM, we can employ well-known techniques that are available to compute the likelihood, and in particular we can express it using the following matrix multiplication (see e.g. Zucchini et al. 2017)

$$\mathscr{L}\,(\boldsymbol{y}\,|\boldsymbol{\eta}) = \boldsymbol{\pi}_0^{\star\,'}\, \boldsymbol{P}\,(y_1)\,\boldsymbol{\Phi}\,\boldsymbol{P}\,(y_2)\,\boldsymbol{\Phi}\,\cdots\,\boldsymbol{\Phi}\,\boldsymbol{P}\,(y_{T-1})\,\boldsymbol{\Phi}\,\boldsymbol{P}\,(y_T)\,\mathbf{1}, \tag{3.3}$$

where the diagonal matrix $\boldsymbol{P}\,(\,y\,)$ of dimension $\bar{A} \times \bar{A}$ is defined as

$$\boldsymbol{P}\,(\,y\,) = \operatorname{diag}\{\, \underbrace{p\,(y\,|\,\boldsymbol{\theta}_1),\, \ldots,\, p\,(y\,|\,\boldsymbol{\theta}_1)}_{a_1 \text{ times}},\, \ldots,\, \underbrace{p\,(y\,|\,\boldsymbol{\theta}_K)\ldots p\,(y\,|\,\boldsymbol{\theta}_K)}_{a_K \text{ times}} \,\}, \tag{3.4}$$

and $p\,(y\,|\,\boldsymbol{\theta}_j)$ is the probability density of the emission distribution $f\,(\,\boldsymbol{\theta}_j\,)$. Here, $\mathbf{1}$ denotes an $\bar{A}$-dimensional column vector with all entries equal to one and $\boldsymbol{\pi}_0^\star$ represents the initial distribution for the state aggregates. Note that if we assume that the underlying Markov chain is stationary, $\boldsymbol{\pi}_0^\star$ is solely determined by the transition probabilities $\boldsymbol{\Phi}$, i.e. $\boldsymbol{\pi}_0^\star = (\boldsymbol{I} - \boldsymbol{\Phi} + \boldsymbol{U}\,)^{-1}\,\mathbf{1}$, where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{U}$ is a square matrix of ones. Alternatively, it is possible to start from a specified state, namely assuming that $\boldsymbol{\pi}_0^\star$ is an appropriate unit vector, e.g. $(1, 0, \ldots, 0)$, as suggested by Leroux and Puterman (1992). We finally note that computation of the likelihood in (3.3) is often subject to numerical underflow and hence its practical implementation usually require appropriate scaling (Zucchini et al., 2017).

While a fully Bayesian framework is desirable for its ability to provide coherent uncertainty quantification for parameter values, a perceived drawback of this approach compared with a frequentist analogue is the increased computation required for estimation. Bayesian posterior distributions are only available in closed form under the very restrictive setting when the likelihood and prior are conjugate. Unfortunately, the model outlined in Section 2.2 does not admit such a conjugate prior form and as a result the corresponding posterior (3.2) is not analytically tractable. However, numerical methods such as Markov Chain Monte Carlo (MCMC) can be employed to sample from this intractable posterior. The last twenty years have seen an explosion of research into MCMC methods and more recently approaches scaling them to high dimensional parameter spaces. The next section outlines one such black box implementation that is used to sample from the posterior in (3.2).

3.1 Hamiltonian Monte Carlo, No-U-Turn Sampler and Stan Modeling Language

One particularly successful posterior sampling algorithm is Hamiltonian Monte Carlo (HMC, Duane et al. 1987), where we refer the reader to Neal et al. (2011) for an excellent introduction. HMC augments the parameter space with a ‘momentum variable’ and uses Hamiltonian dynamics to propose new samples. The gradient information contained within the Hamiltonian dynamics allows HMC to produce proposals that can traverse high dimensional spaces more efficiently than standard random walk MCMC algorithms. However, the performance of HMC samplers is dependent on the tuning of the leapfrog discretization of the Hamiltonian dynamics. The No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014) circumvents this burden. NUTS uses the Hamiltonian dynamics to construct trajectories that move away from the current value of the sampler until they make a ‘U-Turn’ and start coming back, thus maximizing the trajectory distance. An iterative algorithm allows the trajectories to be constructed both forwards and backwards in time, preserving time reversibility. Combined with a stochastic optimization of the step size, NUTS is able to conduct efficient sampling without any hand-tuning.

The *stan* modeling language (Carpenter et al., 2016) provides a probabilistic programming environment facilitating the easy implementation of NUTS. The user needs only define the three components of their model: (i) the inputs to their sampler, e.g. data and prior hyperparameters; (ii) the outputs, e.g. parameters of interest; (iii) the computation required to calculate the unnormalized posterior. Following this, *stan* uses automatic differentiation (Griewank and Walther, 2008) to produce fast and accurate samples from the target posterior. *stan*'s easy-to-use interface and lack of required tuning have seen it implemented in many areas of statistical science. As well as using NUTS to automatically tune the sampler, *stan* is equipped with a variety of warnings and tools to help users diagnose the performance of their sampler. For example, convergence of all quantities of interest is monitored in an automated fashion by comparing variation between and within simulated samples initialized at over-dispersed starting values (Gelman et al., 2017). Additionally, the structure of the transition matrix $\boldsymbol{\Phi}$ allows us to take advantage of *stan*'s sparse matrix implementation to achieve vast computational improvements. Although $\boldsymbol{\Phi}$ has dimension $\bar{A} \times \bar{A}$, each row has at most K non-zero terms (representing within state transitions to the next state aggregate or between state transitions), and as a result only a proportion $(K/\bar{A})$ of the elements of $\boldsymbol{\Phi}$ is non-zero. Hence, for large values of the dwell approximation thresholds $\boldsymbol{a}$, the matrix $\boldsymbol{\Phi}$ exhibits considerable sparsity. The *stan* modeling language implements compressed row storage sparse matrix representation and multiplication, which provides considerable speed up when the sparsity is greater than 90% (Stan Development Team, 2018, Ch. 6). In our applied scenario we consider dwell-approximation thresholds as big as $\boldsymbol{a} = (250, 50, 50)$ with sparsity of greater than 99% allowing us to take considerable advantage of this formulation. Finally, we note that our proposed Bayesian approach may suffer from *label switching* (Stephens, 2000) since the likelihood is invariant under permutations of the labels of the hidden states. However, this issue is easily addressed using order constraints provided by *stan*. This strategy worked well in the simulations and applications presented in the paper, without introducing any noticeable bias in the results.

3.2 Bridge Sampling Estimation of the Marginal Likelihood

The Bayesian paradigm provides a natural framework for selecting between competing models by means of the marginal likelihood, i.e.

$$p(\boldsymbol{y}) = \int \mathscr{L}(\boldsymbol{y} \mid \boldsymbol{\eta})\, p(\boldsymbol{\eta})\, d\boldsymbol{\eta}. \tag{3.5}$$

The ratio of marginal likelihoods from two different models, often called the *Bayes factor* (Kass and Raftery, 1995), can be thought of as the weight of evidence in favor of a model against a competing one. The marginal likelihood in (3.5) corresponds to the normalizer of the posterior $p(\boldsymbol{\eta} \mid \boldsymbol{y})$ (3.2) and is generally the component that makes the posterior analytically intractable. MCMC algorithms, such as the *stan*'s implementation of NUTS introduced above, allow for sampling from the unnormalized posterior, but further work is required to estimate the normalizing constant. Bridge sampling (Meng and Wong, 1996; Meng and Schilling, 2002) provides a general procedure for estimating these marginal likelihoods reliably. While standard Monte Carlo (MC) estimates draw samples from a single distribution, bridge sampling formulates an estimate of the marginal likelihood using the ratio of two MC estimates drawn from different distributions: one being the posterior (which has already been sampled from) and the other being an appropriately chosen proposal distribution $q\left(\boldsymbol{\eta}\right)$. The bridge sampling estimate of the marginal likelihood is then given by

$$p\left(\boldsymbol{y}\right) = \frac{\mathbb{E}_{q(\boldsymbol{\eta})}\left[h(\boldsymbol{\eta})\,\mathscr{L}\left(\boldsymbol{y}\mid\boldsymbol{\eta}\right)p\left(\boldsymbol{\eta}\right)\right]}{\mathbb{E}_{p(\boldsymbol{\eta}\mid\boldsymbol{y})}\left[h(\boldsymbol{\eta})\,q(\boldsymbol{\eta})\right]} \approx \frac{\frac{1}{n_2}\sum_{j=1}^{n_2} h(\tilde{\boldsymbol{\eta}}^{(j)})\,\mathscr{L}\left(\boldsymbol{y}\mid\tilde{\boldsymbol{\eta}}^{(j)}\right)p\left(\tilde{\boldsymbol{\eta}}^{(j)}\right)}{\frac{1}{n_1}\sum_{i=1}^{n_1} h\left(\overline{\boldsymbol{\eta}}^{(i)}\right)q(\overline{\boldsymbol{\eta}}^{(i)})},$$

where $h(\boldsymbol{\eta})$ is an appropriately selected *bridge function* and $p(\boldsymbol{\eta})$ denotes the joint prior distribution. Here, $\{\overline{\boldsymbol{\eta}}^{(1)},\ldots,\overline{\boldsymbol{\eta}}^{(n_1)}\}$ and $\{\tilde{\boldsymbol{\eta}}^{(1)},\ldots,\tilde{\boldsymbol{\eta}}^{(n_2)}\}$ represent n_1 and n_2 samples drawn from the posterior $p\left(\boldsymbol{\eta}\mid\boldsymbol{y}\right)$ and the proposal distribution $q(\boldsymbol{\eta})$, respectively. This estimator can be implemented in `R` using the package `bridgesampling` (Gronau et al., 2020), whose compatibility with *stan* makes it particularly straightforward to estimate the marginal likelihood directly from a *stan* output. This package implements the method of Meng and Wong (1996) to choose the optimal bridge function minimizing the estimator mean-squared error and constructs a multivariate normal proposal distribution whose mean and variance match those of the sample from the posterior.

3.3 Comparable Dwell Priors

Model selection based on marginal likelihoods can be very sensitive to prior specifications. In fact, Bayes factors are only defined when the marginal likelihood under each competing model is proper (Robert, 2007; Gelman et al., 2013). As a result, it is important to include any available prior information into the Bayesian modeling in order to use these quantities in a credible manner. Reliably characterizing the prior for the dwell distributions is particularly important for the experiments considered in Section 6, since we use Bayesian marginal likelihoods to select between the dwell distributions associated with HSMMs and HMMs. For instance, if we believe that the length of sleep for an average person is between 7 and 8 hours we would choose a prior that reflects those beliefs in all competing models. However, we need to ensure that we encode this information in *comparable priors* in order to perform 'fair' Bayes factor selection amongst a set of dwell-distributions. Our aim is to infer which dwell distribution, and not which prior specification, is most appropriate for the data at hand.

For example, suppose we consider selecting between geometric (i.e. an HMM), negative binomial or Poisson distributions (i.e. an HSMM), to model the dwell durations of our data. While a Poisson random variable, shifted away from zero to consider strictly positive dwells, has its mean $\lambda_j + 1$ and variance λ_j described by the same parameter λ_j, the negative binomial allows for further modeling of the precision through an additional factor ρ_j. In both negative binomial and Poisson HSMMs, the parameters λ_j are usually assigned a prior $\lambda_j \sim \text{Gamma}\left(a_{0j}, b_{0j}\right)$ with mean $\mathbb{E}\left[\lambda_j\right] = a_{0j}/b_{0j}$ and variance $\text{Var}\left[\lambda_j\right] = a_{0j}/b_{0j}^2$. In order to develop an interpretable comparison of all competing models, we parameterize the geometric dwell distribution associated with state j in the standard HMM (2.1) as also being characterized by the mean dwell length

$\tau_j = 1/(1-\gamma_{jj})$, where the geometric is also shifted to only consider strictly positive support and γ_{jj} represents the probability of self-transition. Under a Dirichlet prior for the state-specific vector of transition probabilities $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{jK}) \sim \text{Dirichlet}(\boldsymbol{v}_j)$, with $\boldsymbol{v}_j = (v_{j1}, \dots, v_{jK})$ and $\beta_j = \sum_{i \neq j} v_{ji}$, the mean and variance of the prior mean dwell under an HMM are given by

$$\mathbb{E}\left[\tau_j\right] = \frac{v_{jj} + \beta_j - 1}{\beta_j - 1} \text{ and } \operatorname{Var}\left[\tau_j\right] = \frac{(v_{jj} + \beta_j - 1)(v_{jj} + \beta_j - 2)}{(\beta_j - 1)(\beta_j - 2)} - \left(\frac{v_{jj} + \beta_j - 1}{\beta_j - 1}\right)^2$$

for $\beta_j > 2$ (the derivation of this result is provided in the Supplementary Material (Hadj-Amar et al., 2022)).

We therefore argue that a comparable prior specification requires hyper-parameters $\{a_{0j}, b_{0j}\}_{j=1}^K$ and $\{\boldsymbol{v}_j\}_{j=1}^K$ be chosen in a way that satisfy $\mathbb{E}\left[\tau_j\right] = \mathbb{E}[\lambda_j + 1]$ and $\operatorname{Var}\left[\tau_j\right] = \operatorname{Var}\left[\lambda_j + 1\right]$, ensuring the dwell distribution in each state has the same prior mean and variance across models. The prior mean can be interpreted as a best a priori guess for the average dwell time in each state, and the variance reflects the confidence in this prior belief. In addition, since the negative binomial distribution is further parameterized by a dispersion parameter ρ_j, we center our prior belief at $\rho_j = 1$, which is the value that recovers geometric dwell durations (namely an HMM) when $\lambda_j = \gamma_{jj}/(1-\gamma_{jj})$. Between state transition probabilities, i.e. the non-diagonal entries of the transition matrix, as well as the emission parameters, are shared between the HMM and HSMM, and thus we may place a prior specification on these parameters that is common across all models.

4 A Comparison with Langrock and Zucchini (2011)

This section presents several simulation studies. Firstly, we show that our Bayesian implementation provides similar point estimates as the methodology of Langrock and Zucchini (2011), serving as a "sanity check". We then proceed to illustrate the benefits adopting a Bayesian paradigm can bring to HSMM modeling.

4.1 Parameter Estimation

For our first example, we simulated $T = 200$ data points from a three-state HSMM (2.2). Conditional on each state j, the observations are generated from a Normal $\left(\mu_j, \sigma_j^2\right)$, and the dwell durations are Poisson(λ_j) distributed. We consider relatively large values for λ_j in order to evaluate the quality of the HSMM approximation provided by (3.1). The full specification is provided in Table 1 and a realization of this model is shown in Figure 4 (a, top). The dwell approximation thresholds $\boldsymbol{a}$ are set equal to $(30, 30, 30)$ and we placed a Gamma$(0.01, 0.01)$ prior on the Poisson rates λ_j. The transition probabilities $\boldsymbol{\pi}_j$ are distributed as Dirichlet$(1, 1)$ and the priors for the Gaussian emissions are given as Normal$(0, 10^2)$ and Inverse-Gamma$(2, 0.5)$ for locations μ_j and scale σ_j^2 , respectively. Overall, this prior specification is considered weakly informative (Gelman et al., 2013, 2017).

Table 1 shows estimation results for our proposed Bayesian methodology as well as the analogous frequentist approach (EM) of Langrock and Zucchini (2011), which will be referred to as LZ-2011. Figure 4 (a) displays: (top) a graphical posterior predictive check consisting of the observations alongside 100 draws from the estimated posterior predictive (Gelman et al., 2013); (bottom) the most likely hidden state sequence, i.e. $\arg\max_{\boldsymbol{z}} p\,(\,\boldsymbol{z} \,|\, \boldsymbol{y},\, \boldsymbol{\eta}\,)$, which is estimated via the Viterbi algorithm (see e.g. Zucchini et al. 2017) using plug-in Bayes estimates of the model parameters; In order to assess the goodness of fit of the model, we also verified normality of the pseudo-residual (see Supplementary Material).

	True	LZ-2011	Proposed		True	LZ-2011	Proposed		True	LZ-2011	Proposed
μ_1	5	4.96	4.95 (4.66–5.24)	σ_3	1	1.01	1.08 (0.90–1.20)	π_{13}	0.70	0.50	0.5 (0.13–0.87)
μ_2	14	14.02	14.02 (13.67–14.37)	λ_1	20	23.47	23.36 (17.03–30.57)	π_{21}	0.20	0.00	0.20 (0.01–0.53)
μ_3	30	30.19	30.18 (29.98–30.38)	λ_2	30	27.22	27.05 (22.43–32.19)	π_{23}	0.80	1.00	0.80 (0.47–0.99)
σ_1	1	1.09	1.15 (0.95–1.40)	λ_3	20	19.98	20.00 (15.93–24.46)	π_{31}	0.10	0.33	0.40 (0.10–0.76)
σ_2	2	1.90	1.95 (1.73–2.22)	π_{12}	0.30	0.50	0.50 (0.13–0.87)	π_{32}	0.90	0.67	0.60 (0.24–0.90)

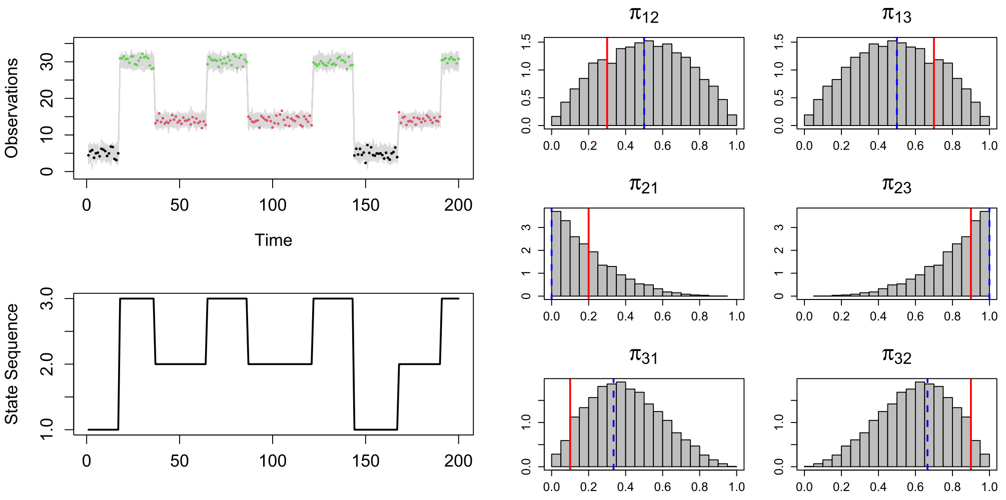
Table 1: Illustrative Example. True model parameterization and corresponding estimates obtained via the EM algorithm and our proposed Bayesian approach. For the latter, we also report 95% credible intervals estimated from the posterior sample.

In general, both methods satisfactorily retrieve the correct pre-fixed duration and emission parameters and the posterior predictive checks indicate that our posterior sampler is performing adequately. The implementation of Langrock and Zucchini (2011) suffers from a lack of regularization, for example in the estimation of π_{21} as 0, and is not currently available with an automatic method to quantify parameter uncertainty. While augmenting the approach of Langrock and Zucchini (2011) by adding regularization penalties to parameters and producing confidence measures such as standard errors and bootstrap estimates is possible, such features are automatic to our Bayesian adaptation. Further, such an approach allows this uncertainty to be incorporated into methods for prediction and model selection making the Bayesian paradigm appealing for HSMM modeling.

4.2 Forecasting

A key feature of HSMMs is their ability to be able to capture and forecast when and for how long the model will be in a given state. We compare the forecasting properties of the method presented by Langrock and Zucchini (2011) and our proposed Bayesian approach. We simulated 20 'un-seen' time series, $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \ldots, \tilde{y}_H)$, where $\tilde{y}_h = y_{T+h}$, $h = 1, \ldots, H$ and $H = 100, 300, 500$ denotes the forecast horizon, from the model as in Table 1. We used the logarithmic score (log-score) to measure predictive performances. Let $\hat{\boldsymbol{\eta}}$ be the frequentist (MLE/EM) parameter estimate and define the log-score

$$L_{\,\mathrm{freq}}(\tilde{\boldsymbol{y}}) = \sum_{h=1}^{H} -\log p\,(\tilde{y}_h \,|\, \hat{\boldsymbol{\eta}}),$$



(a) (b)

Figure 4: (a, top) a realization (dots) of a three-state HSMM with Gaussian emissions and Poisson durations, where different colors correspond to (true) different latent states. Grey lines represent 100 samples drawn from the estimated posterior predictive distribution. (a, bottom) Most likely hidden state sequence estimated via the Viterbi algorithm; (b) estimated posterior distribution of the transition probabilities π_{jk}, where vertical solid red and blue dotted lines represent true values and EM estimates, respectively.

where $p\,(\tilde{y}_h \,|\, \hat{\boldsymbol{\eta}})$ denotes the forecast density function (see Supplementary Material for an explicit expression). Our Bayesian framework does not assume a point estimate $\hat{\boldsymbol{\eta}}$ but considers instead a posterior distribution $p\,(\boldsymbol{\eta} \,|\, \boldsymbol{y})$, which is integrated over to produce a predictive density. Given M MCMC samples drawn from the posterior, $\{\boldsymbol{\eta}^{(i)}\}_{i=1}^{M} \sim \pi\,(\boldsymbol{\eta} \,|\, \boldsymbol{y})$, the log-score of the predictive density can be approximated as

$$L_{\text{Bayes}}(\tilde{\boldsymbol{y}}) = \sum_{h=1}^{H} - \log p\,(\tilde{y}_h \,|\, \boldsymbol{y}) = \sum_{h=1}^{H} - \log \int p\,(\tilde{y}_h \,|\, \boldsymbol{\eta})\, p\,(\boldsymbol{\eta} \,|\, \boldsymbol{y})\, d\boldsymbol{\eta}$$
$$\approx \sum_{h=1}^{H} - \log \left(\frac{1}{M} \sum_{i=1}^{M} p\,(\tilde{y}_h \,|\, \boldsymbol{\eta}^{(i)}) \right).$$

Figure 5 presents box-plots of log-scores for LZ-2011 and our proposed Bayesian approach. It is clear that our Bayesian methodology typically produces a much lower predictive log-score than the frequentist procedure. The approach by Langrock and Zucchini (2011) which uses plug-in estimates for parameters, is known to 'under-estimate' the true predictive variance thus yielding large values of the log-score (Jewson et al., 2018). On the other hand, our Bayesian paradigm integrates over the parameters and hence is more accurately able to capture the true forecast distribution. As a result, it produces significantly smaller log-score estimates.

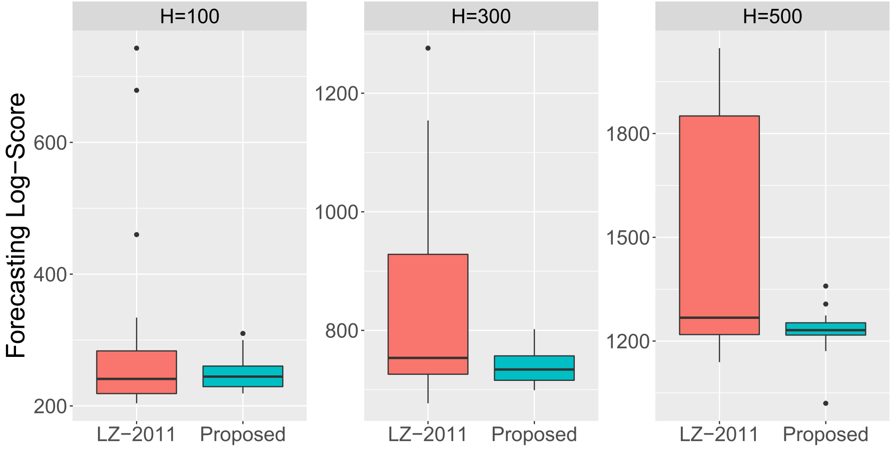


Figure 5: Boxplots of log-scores for LZ-2011 (via EM) and our Bayesian methodology, with three different forecast horizons $H = 100, 300, 500$.

4.3 Dwell Distribution Selection

An important consideration is whether to formulate an HMM or to extend the dwell distribution beyond a geometric one (i.e., an HSMM). Ideally, the data should be used to drive such a decision. In this section, we compare the frequentist methods for doing so, namely Akaike's information criterion (AIC, Akaike 1973) and Bayesian information criterion (BIC, Schwarz et al. 1978), with their Bayesian counterpart, namely the marginal likelihood. We choose not to consider other Bayesian inspired information criteria (e.g. Spiegelhalter et al., 2002; Watanabe, 2010; Gelman et al., 2014) as our goal here is to compare standard frequentist methods used previously in the literature to conduct model selection for HMMs and HSMMs (e.g. Langrock and Zucchini, 2011; Huang et al., 2018) with the canonical Bayesian analog. Although the performance of Bayesian model selection can be sensitive to the specification of the prior, we gave specific consideration to specifying this with model selection in mind in Section 3.3.

Consistency for Nested Models

A special feature of the negative binomial dwell distribution is that the geometric dwell distribution associated with HMMs is nested within it. Taking $\rho = 1$ for the negative binomial exactly corresponds to the geometric distribution. An important consideration when selecting between nested models is complexity penalization. For the same data set, the more complicated of two nested models will always achieve a higher in-sample likelihood score than the simpler model. Therefore, in order to achieve consistent model selection among nested models, the extra parameters of the more complex models must be penalized. In this scenario, the AIC $:= -2\,\mathscr{L}\,(\boldsymbol{y}\,|\,\boldsymbol{\eta}) + 2p$ where p denotes the number of parameters included in the model, is known not to provide consistent model selection when the data is generated from the simpler model (see e.g. Fujikoshi, 1985). On the other hand, performing model selection using the marginal likelihood can be shown to be consistent (see e.g. O'Hagan and Forster 2004), provided some weak conditions on the prior are satisfied. Therefore, when following a Bayesian paradigm, the correct data generating model is selected with probability one as T tends to infinity. Here we show that under the approximate HSMM likelihood model, Bayesian model selection appears to maintain its desirable properties.

We simulated 20 time series from a two-state HMM with Gaussian emission parameters $\boldsymbol{\mu} = (1, 4)$ and $\boldsymbol{\sigma}^2 = (1, 1.5)$, and diagonal entries of the transition matrix set to $(\gamma_{11}, \gamma_{22}) = (0.7, 0.6)$. To model this data we considered the HMM and a HSMM with negative binomial durations. For the HSMM approximation, we considered $\boldsymbol{a} = (3, 3)$, $(5, 5)$ and $(10, 10)$ in order to investigate how the dwell approximation affects the model selection performance. We use prior distributions that are comparable as explained in Section 3.3, the exact prior specifications are presented in the Supplementary Material. Figure 6 (top) displays box-plots of the difference between the model selection criteria (namely marginal likelihood and AIC) achieved by the HMM and the HSMM, for increasing sample size $T = 500, 5000, 10000$ and values for $\boldsymbol{a}$. We negate the AIC such that maximizing both criteria is desirable. Thus, positive values for the difference correspond to correctly selecting the simpler data generating model, i.e. the HMM. As the sample size T increases, the marginal likelihood appears to converge to a positive value, and the variance across repeats decreases, indicating consistent selection of the correct model. On the other hand, even for large T there are still occasions when the AIC strongly favors the incorrect, more complicated model. Further, such performance appears consistent across values of $\boldsymbol{a}$.

Complexity Penalization

Unlike the AIC, the BIC $:= -2\,\mathscr{L}\,(\boldsymbol{y}\,|\,\boldsymbol{\eta}) + p \log T$ penalizes complexity in a manner that depends on the sample size T. This is termed 'Bayesian' because it corresponds to the Laplace approximation of the marginal likelihood of the data (Konishi and Kitagawa, 2008), often interpreted as considering a uniform prior for the model parameters (Bhat and Kumar, 2010; Sodhi and Ehrlich, 2010). Though the uniform distribution may be viewed as naturally uninformative, it is well known that using the marginal likelihood assuming an uninformative prior specification can lead to the selection of the simplest model independently of the data (see e.g. Lindley, 1957; Jeffreys, 1998; Jennison, 1997). As a result, while BIC can provide consistent selection of nested models, it can punish extra complexity in an excessive manner.

To investigate how the approximate HSMM likelihood model affects this model selection behavior, we consider data generated from an HSMM with the same formulation as above except that in this scenario the dwell distribution is a negative binomial parameterized by state-specific parameters $\boldsymbol{\lambda} = (3.33, 2.50)$ and $\boldsymbol{\rho} = (2, 0.5)$. Note that the data generating HSMM has two more parameters than the HMM. For the HSMM

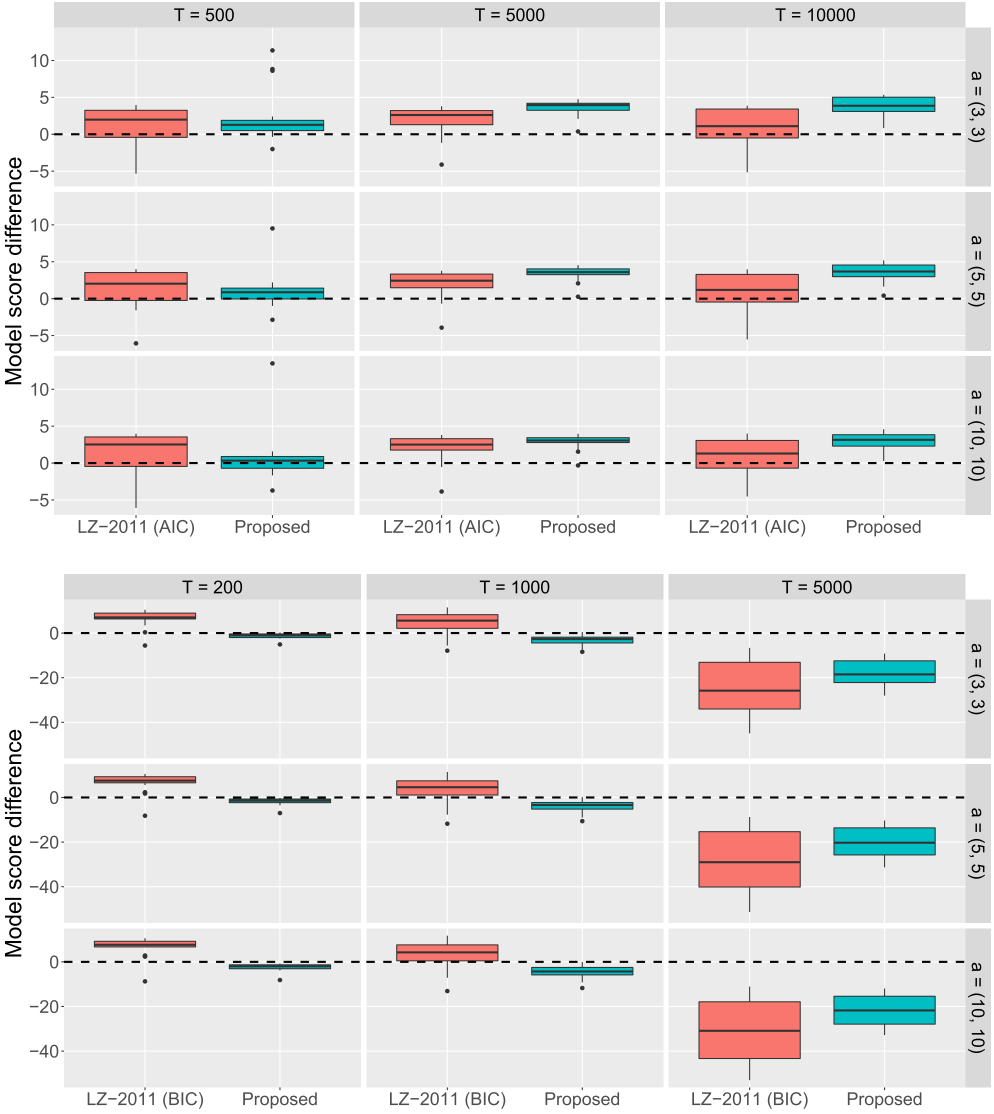


Figure 6: AIC, BIC and the marginal likelihood for nested models. (Top) model score differences between a negative binomial duration HSMM with $\boldsymbol{a} = (3, 3), (5, 5)$ and $(10, 10)$ and an HMM when the data is generated from the HMM. Positive values of the model score difference correspond to correctly selecting the simpler model. (Bottom) model score differences between a negative binomial duration HSMM approximated by a threshold $\boldsymbol{a} = (3, 3), (5, 5)$ and $(10, 10)$ and an HMM when the data is generated from the HSMM. Negative values of the model score difference correspond to correctly selecting the more complicated model. Note that here we interpret the difference between the AIC/BIC and the log-marginal likelihood values of two models as quantitatively comparable for model selection decisions, with being greater than or less than 0 corresponding to the selection decision.

approximation, we consider $\boldsymbol{a} = (3, 3)$, $(5, 5)$ and $(10, 10)$, where the largest of these provides negligible truncation of the right tail of the dwell distribution given the data generating parameters. Figure 6 (bottom) shows box-plots of the difference between the model scores (marginal-likelihood and BIC) across 20 simulated time series when fitting the HMM and HSMM, for increasing sample size $T = 200, 1000, 5000$ and values for $\boldsymbol{a}$. We negate the BIC so that the preferred model maximizes both criteria. Unlike the experiments described above, the data is now from the less parsimonious HSMM approach and therefore negative values for the difference in score correspond to correctly selecting the more complicated model. For small sample sizes, e.g. $T = 200, 1000$, the complexity penalty of the BIC appears to be too large, so that in almost all of the 20 repeat experiments the simple model is incorrectly favored over the correct data generating model, i.e. the HSMM. On the other hand, the marginal likelihood is able to correctly select the more complicated model across almost all simulations and sample sizes. Although for smaller $\boldsymbol{a}$ the HSMM approximation is 'closer' to a HMM, we still see that the model selection performance is consistent across the different values of $\boldsymbol{a}$.

5 Approximation Accuracy and Computational Time

The previous section motivated why the Bayesian paradigm can improve statistical inferences for HSMMs. Next, we investigate the computational feasibility of such an approach and the trade-off between computational efficiency and statistical accuracy achieved by our Bayesian approximate HSMM implementation. In particular, we compare our Bayesian approximate HSMM method for different values of the threshold $\boldsymbol{a}$ with a Bayesian implementation of the exact HSMM, while also illustrating the computational savings made by our sparse matrix implementation. For the exact HSMM, the full-forward recursion is used to evaluate the likelihood (see e.g. Guédon 2003 or Economou et al. 2014). In order to provide a fair comparison, we coded the forward recursion outlined in Economou et al. (2014) in *stan* also. We then compare the computational resources required to sample from the approximate and exact HSMM posteriors with the accuracy of the posterior mean parameter estimates with respect to their data generating values.

We generate $T = 5000$ observations from two different HSMMs with Poisson durations (both with $K = 5$ states and the same Gaussian emission distributions). For the two different datasets, we consider the following dwell parameters: (i) *short dwells*, i.e. $\boldsymbol{\lambda} = (2, 5, 8, 1, 4)$, where the average time spent in each state is fairly small and (ii) *one long dwell*, i.e. $\boldsymbol{\lambda} = (2, 5, 25, 1, 4)$, where four states have short average dwell time and one where the average dwell time is much longer. We also consider two approximation thresholds: $\boldsymbol{a}_1 = (10, 10, 10, 10, 10)$, namely a fixed approximation threshold for all five states, and $\boldsymbol{a}_2 = (10, 10, 30, 10, 10)$, a 'hybrid' model where four of the states have short dwell thresholds and one has a longer threshold. The emission parameters were set to $\boldsymbol{\mu} = (1, 2, 3.5, 6, 10)$ and $\boldsymbol{\sigma}^2 = (1^2, 0.5^2, 0.75^2, 1.5^2, 2.5^2)$, and we specify priors $\mu_j \sim \mathcal{N}(0, 10^2)$, $\sigma_j^2 \sim \mathcal{IG}(2, 0.5)$, $\lambda_j \sim \mathcal{G}(0.01, 0.01)$ and $\gamma_j \sim \mathcal{D}(1, \dots, 1)$ for $j = 1, \dots, 5$.

The results are presented in Table 2. Across both datasets and approximation thresholds, the sparse implementation takes less than half the time of the non-sparse implementation, with the saving greater when the dwell thresholds are larger (and the matrix

Short dwells	Time (hours)	ESS		MSE	
			μ	σ^2	λ
Approx: $\boldsymbol{a_1}$	2.62	986.50	3.72×10^{-2}	2.88×10^{-3}	0.25
Approx (SPARSE): $\boldsymbol{a_1}$	1.30	975.20	3.84×10^{-2}	3.06×10^{-3}	0.26
Approx: $\boldsymbol{a_2}$	3.94	961.76	3.84×10^{-2}	3.05×10^{-3}	0.17
Approx: (SPARSE): $\boldsymbol{a_2}$	1.78	978.82	3.96×10^{-2}	3.01×10^{-3}	0.18
Exact	81.15	933.28	4.02×10^{-2}	3.24×10^{-3}	0.19
One long dwell					
Approx: $\boldsymbol{a_1}$	3.33	984.51	1.76×10^{-2}	4.68×10^{-2}	128.50
Approx: (SPARSE): $\boldsymbol{a_1}$	1.78	981.90	1.73×10^{-2}	4.84×10^{-2}	128.51
Approx: $\boldsymbol{a_2}$	5.08	993.89	1.50×10^{-2}	4.84×10^{-2}	1.25
Approx: (SPARSE): $\boldsymbol{a_2}$	2.21	983.47	1.51×10^{-2}	4.82×10^{-2}	1.25
Exact	101.35	980.59	1.65×10^{-2}	4.66×10^{-2}	1.12

Table 2: Computational time (hours), effective sample size (ESS) and mean squared error (MSE) of posterior mean parameters. The results are reported using the approximate HSMM for different dwell approximations $\boldsymbol{a}$ (with the corresponding sparse implementation), and the exact HSMM implementation.

$\boldsymbol{\Phi}$, (3.4), is sparser). Furthermore, the HSMM approximations are considerably faster than the full HSMM implementation. For the *short dwell* dataset the full HSMM takes close to 3.5 days while the sparse implementations of the HSMM approximation both require less than 2 hours. Similarly, for the *one long dwell* dataset, the full HSMM takes over 4 days to run while again the sparse HSMM approximations require around 2 hours. The quoted Effective Sample Size (ESS, e.g. Gelman et al. 2013) values are calculated using the *LaplaceDemons* package in R and are averaged across parameters. These show that the ESS of all the generated samples is close to 1000 and thus the time comparisons are indeed fair. Further, we expect the difference to become starker as the number of observations T increases. While, the approximate HSMM scales linearly in T and quadratically in $\sum_{j=1}^{K} a_j$, the full HSMM in the worst case is quadratic in T (Langrock and Zucchini, 2011).

Lastly, we see that the savings in computation time come at very little cost in statistical accuracy. We measure the statistical accuracy of the vector-valued parameter $\hat{\boldsymbol{\theta}}$ to estimate $\boldsymbol{\theta}^*$ using its mean squared error (MSE $= \sum_{j=1}^{K}(\hat{\theta}_j - \theta_j^*)^2$). All methods achieve almost identically MSE values for the emission parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. For the *short dwell data*, the $\boldsymbol{a}_1$ approximation has slightly higher MSE for $\boldsymbol{\lambda}$ while the $\boldsymbol{a}_2$ approximation performed comparably to the HSMM. Clearly, increasing the approximation threshold improves statistical accuracy. On the other hand, the *one long dwell* shows that if the dwell threshold is set too low, as is the case with $\boldsymbol{a}_1$, large errors in the dwell estimation can be made. However, in this example the higher dwell approximation $\boldsymbol{a}_2$ once again performs comparably with the full HSMM, whilst requiring only 2% of the computational time.

5.1 Setting the Dwell Threshold

The results of Section 5 indicate that while vast computational savings are possible using the approximate HSMM likelihood, care must be taken not to set the dwell approximation threshold $\boldsymbol{a}$ too low. We propose initializing $\boldsymbol{a}$ based on the prior distribution for the dwell times d_j, $\pi(d_j) = \int \pi(d_j; \lambda)\pi(\lambda)d\lambda$. Noting that any dwell time $d_j < \boldsymbol{a}_j$ is not approximated, we recommend initializing $\tilde{\boldsymbol{a}}$ such that $d_j \leq \tilde{\boldsymbol{a}}_j$ with high probability for all $j = 1, \ldots, K$.

Such an initialization however does not guarantee the accuracy of the HSMM modeling, particularly in the absence of informative prior beliefs. We therefore, propose a diagnostic method to check that $\tilde{\boldsymbol{a}}$ is not too small.

1. Initialize $\tilde{\boldsymbol{a}}$ and conduct inference on the observed data. Record posterior mean parameter estimates $\hat{\boldsymbol{\eta}}_{obs}(\tilde{\boldsymbol{a}})$
2. Generate data $\tilde{y}_{gen}$ from an exact HSMM with generating parameters $\hat{\boldsymbol{\eta}}_{obs}(\tilde{\boldsymbol{a}})$. Note that generation from an exact HSMM is easier than inference on its parameters
3. Continuing with $\tilde{\boldsymbol{a}}$, conduct inference on the generated data and record posterior mean parameter estimates $\hat{\boldsymbol{\eta}}_{gen}(\tilde{\boldsymbol{a}})$
4. Compare dwell distribution parameters $\hat{\lambda}_{obs}(\tilde{\boldsymbol{a}})$ and $\hat{\lambda}_{gen}(\tilde{\boldsymbol{a}})$

The estimates $\hat{\lambda}_{obs}(\tilde{\boldsymbol{a}})$ provide the best guess estimate of the parameters of the HSMM underlying the data for fixed $\tilde{\boldsymbol{a}}$. Generating from this exact HSMM given by these estimates allows us to verify the accuracy of the proposed model. If the estimates are not accurate then little confidence can be had that $\hat{\lambda}_{obs}(\tilde{\boldsymbol{a}})$ accurately represents the dwell distribution of the underlying HSMM. If $\hat{\lambda}_{gen}(\tilde{\boldsymbol{a}})_j$ is not considered a satisfactory estimate of $\hat{\lambda}_{obs}(\tilde{\boldsymbol{a}})_j$, then $\tilde{\boldsymbol{a}}_j$ must be increased. Conveniently, this can be done for each state j independently. Further, if $\hat{\lambda}_{gen}(\tilde{\boldsymbol{a}})_j$ is considered accurate enough, then there is also the possibility to decrease $\tilde{\boldsymbol{a}}_j$ based on the inferred dwell distribution. Although the above procedure requires the fitting of the model several times, we believe the computational savings of our model when compared with the exact HSMM inference demonstrated in Table 2 render this worthwhile. This procedure is implemented to set the dwell-approximation threshold for the physical activity time series analyzed in the next section.

6 Telemetric Activity Data

In this section, we return to the physical activity (PA) time series that Huang et al. (2018) analyzed using a frequentist HMM. We seek to conduct a similar study but within a Bayesian framework and consider the extra flexibility afforded by our proposed methodology to investigate departures from the HMM. Further, in Section 6.1 we consider the inclusion of spectral information within the HMM and HSMM emission densities.

We consider three-state HSMMs with Poisson (λ_j) and Neg-Binomial (λ_j, ρ_j) dwell durations, shifted to have strictly non-negative support and approximated via thresholds $\boldsymbol{a}_P = (160, 40, 25)$ and $\boldsymbol{a}_{NB} = (250, 50, 50)$ respectively. These are fitted to the square root of the PA time series shown in Figure 1, wherein we assume that transformed observations are generated from Normal(μ_j, σ_j^2) distributions, as in Huang et al. (2018). We specified $K = 3$ states, in agreement with findings of Migueles et al. (2017) and Huang et al. (2018), where they collected results from more than forty experiments on PA time series. In their studies, for each individual the lowest level of activity corresponds to the sleeping period, which usually happens during the night, while the other two phases are mostly associated with movements happening in the daytime. Henceforth, these different telemetric activities are represented as inactive (IA), moderately active (MA) and highly active (HA) states. The setting of $\boldsymbol{a}$ followed the iterative process outlined in Section 5.1, initializing $\tilde{a}_j$ giving prior probability of 0.9 that $d_j < a_j$. This choice also reflects a trade-off between accurately capturing the states with which we have considerable prior information, i.e. IA, whilst improving the computational efficiency of the other states over a standard HSMM formulation.

We assume that the night rest period of a healthy individual is generally between 7 and 8 hours. The parameter of the dwell duration of the IA state, λ_{IA}, is hence assigned a Gamma prior with hyperparameters that reflect mean 90 (i.e. 7.5×12) and variance 36 (i.e. $[0.5 \times 12]^2$), the latter was chosen to account for some variability amongst people. Since we do not have significant prior knowledge on how long people spend in the MA and HA states, we assigned λ_{MA} and λ_{HA} Gamma priors with mean 24 (i.e. 2 hours) and variance 324 (i.e. $[1.5 \times 12]^2$) to reflect a higher degree of uncertainty. Transition probabilities from state IA, $\boldsymbol{\pi}_{\text{IA}}$, are specified as Dirichlet with equal prior probability of switching to any of the active states MA or HA. On the other hand, active states usually alternate between each other more frequently than with IA (Huang et al., 2018), and therefore we set the prior for $\boldsymbol{\pi}_{\text{MA}}$ so that transitions from MA to HA are four times more likely than switching from MA to IA (a similar argument can be made for $\boldsymbol{\pi}_{\text{HA}}$). Finally, the inverse of dispersion parameters ρ_j^{-1} were given Gamma $(2, 2)$ priors, and the parameters of the Gaussian emissions were assigned $\mu_j \sim$ Normal $(\bar{y}, 4)$ and $\sigma_j^2 \sim$ Inverse-Gamma $(2, 0.5)$, where $\bar{y}$ denotes the sample mean.

For each proposed model our Bayesian procedure is run for 6,000 iterations, 1,000 of which are discarded as burn-in. Firstly, we consider selecting which of the competing dwell distributions, i.e. the geometric dwell characterizing the HMM and the Poisson and negative binomial HSMM extensions, is most supported by the observed data. As explained in Section 3.3, we specified hyperparameters for these competing models so that the corresponding priors match the means and variances of the informative prior specification given above. In order to measure the gain of including available prior knowledge into the model, we also investigated a weakly informative prior setting (as in Section 4.1). Table 3 displays the bridge sampling estimates of the marginal likelihood for the different models and posterior means of the corresponding dwell parameters. It is clear that integrating into the model available prior information improves performance greatly. In addition, modeling dwell durations as either negative binomial or geometric provides a better approximation to the data compared to a Poisson model. Furthermore, the Bayes factor 18.36 (i.e. $\exp\{-1632.42 + 1635.33\}$) suggests that there is

	log-marg lik	λ_{IA}	λ_{MA}	λ_{HA}	ρ_{IA}	ρ_{MA}	ρ_{HA}
Poisson†	−1751.02	88.32 (86.28–89.28)	34.79 (29.08–43.02)	18.55 (14.45–22.47)	–	–	–
Geometric†	−1653.67	45.57 (26.97–74.42)	10.53 (7.49–14.53)	8.60 (6.13–11.94)	–	–	–
Neg-Binom†	−1649.00	46.25 (21.12–88.14	10.46 (6.22–16.94)	8.37 (5.44–12.31)	0.61 (0.29–1.08)	0.61 (0.33–0.98)	1.22 (0.60–2.26)
Poisson	−1732.16	88.39 (86.65–89.27)	33.61 (28.76–40.55)	17.98 (14.35–22.04)	–	–	–
Geometric	−1635.33	88.72 (79.63–98.70)	13.42 (9.49–18.68)	10.97 (7.91–15.05)	–	–	–
Neg-Binom	**−1632.42**	87.97 (78.40–97.75)	12.07 (7.33–18.84)	9.12 (5.99–13.19)	0.67 (0.33–1.19)	0.71 (0.36–1.15)	1.25 (0.60–2.22)

Table 3: Telemetric activity data. Log-marginal likelihood for different dwell distributions (Poisson, geometric and negative binomial), where the superscript † denotes a weakly informative prior specification. Geometric durations are characterized by their mean dwell length $\lambda_j = 1/(1-\gamma_{jj})$ where γ_{jj} represents the probability of self-transition. Estimated posterior means of the dwell parameters are reported with a 90% credible intervals estimated from the posterior sample.

strong evidence (Kass and Raftery, 1995) in favor of the HSMM with negative binomial durations in comparison to a standard HMM. This is also reflected by the estimated posterior means of the parameters ρ_j which differ from one, hence showing some departure from geometric dwell durations. These 'dispersion' parameters are smaller than one for the IA and MA states indicating a larger fitted variance of the dwell times under the negative binomial HSMM than the geometric HMM. Combined with their estimated means, this may explain the improved performance of the negative binomial dwell model over the HMM. The increased variance allows the time series to better capture the short transitions to IA states seen in the fitted model (Figure 7). This also explains why the Poisson HSMM performs poorly for this dataset; the fitted Poisson dwell distribution for the IA state can be seen to have a much smaller variance than the geometric and negative binomial alternatives. Plots comparing the posterior predictive dwell time for the IA, MA, and HA states estimated under the three proposed dwell distributions are provided in the Supplementary Material. Future work could consider more complex dwell distributions to reflect the different patterns of human sleep. For example, a natural extension to the results presented here could be to look at whether a two-component mixture distribution (e.g. Poisson) can aid in better capturing the short excursions to the IA seen in Figure 7. In the Supplementary Material, we have further investigated the different state classifications provided by the optimal proposed model (using negative binomial durations) with respect to Poisson and geometric dwells.

Posterior means of the emission parameters were $y_t|_{\text{IA}} \sim \text{Normal}(0.93, 0.47)$, $y_t|_{\text{MA}} \sim \text{Normal}(3.17, 1.28)$ and $y_t|_{\text{HA}} \sim \text{Normal}(5.38, 0.54)$. The IA state naturally corresponds to the state with the lowest mean activity and the MA state appears to have largest variance in activity levels. Posterior means of the dwell parameters in Table 3 show that this individual sleeps an average of 7 and a half hours per night. In Figure 8, we display posterior histograms of the transition probabilities between different states. There ap-

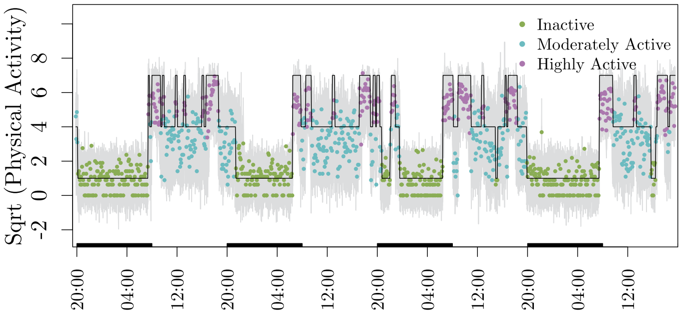


Figure 7: Square root of the PA time series along with simulated observations from the fitted model with negative binomial dwell-time. The piecewise horizontal line represents the estimated state sequence. Rectangles on the time axis correspond to periods from 20.00 to 8.00. IA state happens during night, whereas days are characterized by many switches between MA and HA states. This picture is best viewed in color.

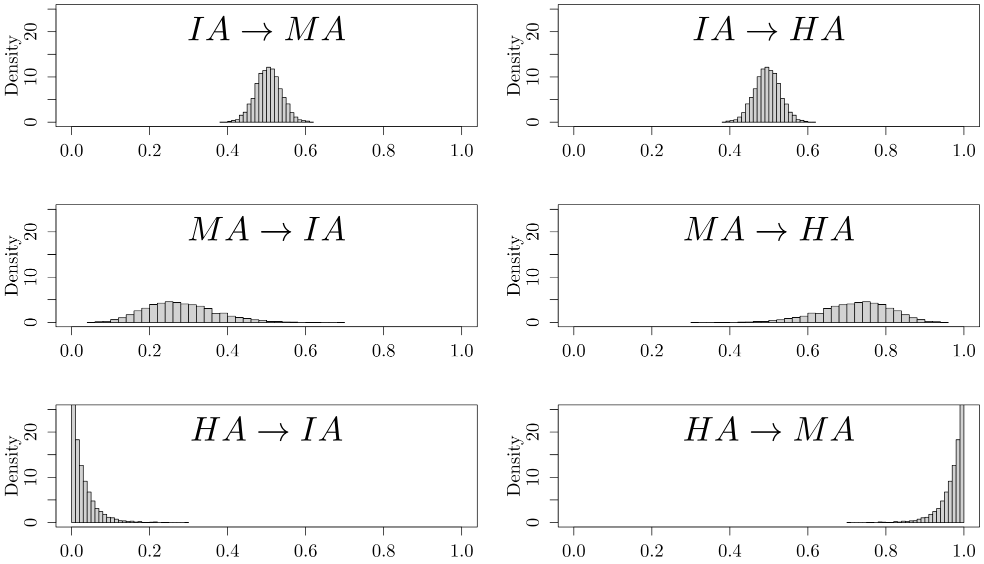


Figure 8: Estimated posterior density histograms of the transition probabilities between IA, MA, and HA states from the proposed HSMM with negative binomial dwell-times.

pears to be high chances of switching between active states, since the posterior means for $\pi_{\text{HA}\to\text{MA}}$ and $\pi_{\text{MA}\to\text{HA}}$ are close to one, though the latter exhibits larger variance. Additionally, the posterior probability of transitioning from HA to IA is very close to zero, which is reasonable considering that it is very unlikely that an individual would go to sleep straight after having performed intense physical activity. Figure 7 shows the transformed time series as well as simulated data from the predictive distribution, and the estimated hidden state sequence using the Viterbi algorithm. It can be seen that the IA state occurs during the night whereas days are characterized by many switches between the MA and HA states. Our results are in agreement with Huang et al. (2018).

6.1 Harmonic Emissions

Huang et al. (2018) further extended the standard Gaussian HMM for the PA recordings by allowing the state transition dynamics to depend on body's circadian periodicity (24 hours). In a similar vein, we investigate the inclusion of spectral information within the emission density, and study how this affects the HMM and HSMM models considered in the previous section. Specifically, we consider that the observations are generated from state-specific *harmonic emissions* of the form $y_t \,|\, z_t = j \sim \mathcal{N}\,(\,\mu_j(t), \sigma_j^2)$, with oscillatory mean defined as

$$\mu_j(t) = \beta_j^{(0)} + \beta_j^{(1)} \cos(2\pi\hat{\omega}t) + \beta_j^{(2)} \sin(2\pi\hat{\omega}t). \tag{6.1}$$

This emission density is hence expressed as a sum of a sine and a cosine (weighted by the linear coefficients $\beta_j^{(1)}$ and $\beta_j^{(2)}$) oscillating at frequency $\hat{\omega}$, plus a state-specific intercept $\beta_j^{(0)}$. While Huang et al. (2018) choose a priori the 24-hour periodicity included in the basis function, in our study we estimate this directly from the data. The next section describes our approach for identifying the frequency $\hat{\omega}$ driving the overall variation in the PA time series.

Identifying the Periodicity

We define $\hat{\omega}$ as the posterior mean of the frequency ω under the periodic model in (6.2) defined below, i.e. $\hat{\omega} := \mathbb{E}\,(\omega \,|\, \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2)$, with $\boldsymbol{\beta} = (\beta^{(1)}, \beta^{(2)})$. In this preliminary step to the proposed model with harmonic emissions (6.1), we first assume the data to be generated by the following stationary periodic process

$$y_t = \beta^{(1)} \cos(2\pi\omega t) + \beta^{(2)} \sin(2\pi\omega) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\omega^2), \qquad t = 1, \ldots, T, \tag{6.2}$$

where we have developed a Metropolis-within-Gibbs sampler to obtain samples from the posterior distribution of the frequency

$$p\,(\omega \,|\, \boldsymbol{\beta},\, \sigma^2,\, \boldsymbol{y}) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_t \Big\{ y_t - \beta^{(1)} \cos(2\pi\omega t) - \beta^{(2)} \sin(2\pi\omega) \Big\}^2 \right] \mathbb{1}_{\left[\,\omega \,\in\, (0,\, \phi_\omega)\right]}, \tag{6.3}$$

where ϕ_ω is a pre-specified upper bound for the frequency and may be chosen to reflect prior information about the value of ω, for example focusing only on low frequencies (e.g. $0 < \phi_\omega < 0.1$). Full details of the sampling scheme and our prior choice are provided in the Supplementary Material. This algorithm is similar to the within-model move of the "segment model" presented in Hadj-Amar et al. (2019, 2021), but with the number of frequencies fixed at one.

We ran the sampler for 5000 iterations using software written in Julia 1.6 which took around 3 seconds on an Intel$^{®}$ Core$^{\text{TM}}$ i5 2 GHz Processor with 16 GB RAM. Figure 9

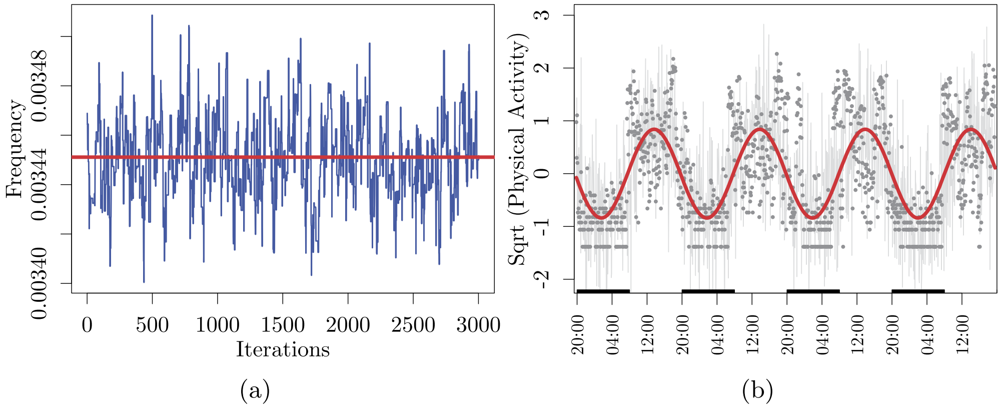


Figure 9: Identifying the periodicity via the periodic model in (6.2). Panel (a) shows the trace plot (after burn-in) of the posterior distribution of the frequency. Panel (b) displays draws from the posterior predictive as well as the estimated periodic signal. In both plots, the red represents posterior mean.

(a) shows the trace plot (after burn-in) of the posterior sample of the frequency where the acceptance rate (28%) was roughly tuned to be optimal (Roberts et al., 2001). We also highlight in red the posterior mean $\hat{\omega} = 0.003453$. In Figure 9 (b) we display 20 draws from the posterior predictive distribution of the stationary periodic model and the posterior mean of the oscillatory signal. This shows that the model predictions appear to capture some of the structure of the PA time series. However, there also appears to be temporal structure not captured by the global circadian harmonic. As a result, in the next section we will use the global $\hat{\omega} = 0.003453$ as the circadian covariate for the emissions of the harmonic HMM and HSMM (6.1), allowing the harmonic parameters $(\beta_j^{(0)}, \beta_j^{(1)}, \beta_j^{(2)})$ to vary by state in order to better capture the temporal structure.

Results

Given the point estimate for $\hat{\omega} = 0.003453$, we then applied the HMM and HSMM approximations with Poisson and negative binomial dwells to the PA time series (using $K = 3$ states). Our prior specification follows the discussion in Section 6 for σ_j^2, λ_j, γ_j and ρ_j, where appropriate, while the intercept of the harmonic mean model $\beta_j^{(0)}$ is given the same prior as μ_j from the standard Gaussian emission model. The additional parameters of the harmonic model $\beta_j^{(1)}$ and $\beta_j^{(2)}$ are both assumed a priori $\mathcal{N}(0, 2^2)$.

Table 4 (top) provides the log-marginal likelihoods of the different models and posterior mean estimates of the parameter of their dwell-distributions, along with the 90% credibility intervals provided by their posteriors. It is clear that the marginal likelihood favors the negative binomial dwell distribution, with the standard HMM (geometric dwell) being the next most favorable. Further, when comparing Table 4 with Table 3, we see that the inclusion of harmonic emissions results in an increase of the marginal

	log-marg lik	λ_{IA}	λ_{MA}	λ_{HA}	ρ_{IA}	ρ_{MA}	ρ_{HA}
Poisson	−1727.24	88.29 (86.42–89.25)	44.68 (41.80–47.57)	21.62 (18.52–25.05)	–	–	–
Geometric	−1629.40	88.24 (79.08–98.05)	15.53 (10.19–23.08)	12.15 (8.36–17.61)	–	–	–
Neg-Binom	**−1625.61**	87.54 (77.66–97.34)	14.32 (7.82–24.02)	10.67 (6.44–16.20)	0.65 (0.31–1.17)	0.64 (0.33–1.13)	1.7 (0.63–3.88)

	IA			MA			HA		
	$\beta^{(0)}$	$\beta^{(1)}$	$\beta^{(2)}$	$\beta^{(0)}$	$\beta^{(1)}$	$\beta^{(2)}$	$\beta^{(0)}$	$\beta^{(1)}$	$\beta^{(2)}$
Gaussian	0.93 (0.88–0.98)	–	–	3.18 (3.03–3.33)	–	–	5.38 (5.27–5.51)	–	–
Harmonic	1.36 (1.26–1.46)	0.04 (−0.05–0.13)	−0.60 (−0.72–−0.47)	3.32 (2.94–3.65)	−0.11 (−0.34–0.13)	−0.24 (−0.69–0.17)	5.46 (5.32–5.60)	0.20 (0.07–0.33)	−0.23 (−0.61–0.16)

Table 4: Telemetric activity data with harmonic emissions. (Top) Log-marginal likelihood for different dwell durations (i.e. Poisson, geometric and negative binomial). Geometric durations are characterized by their mean dwell length $\lambda_j = 1/(1-\gamma_{jj})$ where γ_{jj} represents the probability of self-transition. (Bottom) Parameters of the mean of the Gaussian and harmonic emission distributions under the selected negative binomial dwell distribution. Estimated posterior means of the parameters are reported with a 90% credible intervals estimated from the posterior sample.

likelihood by a factor ranging between 6 and 7 on the log-scale for all dwell distributions, thus supporting its integration in our model.

Following the selection of the negative binomial dwell distribution for both the standard Gaussian and harmonic emission models, Table 4 (bottom) provides the posterior mean values for the parameters of these emission distributions, along with the 90% credibility intervals provided by the posterior. These results show that even with a global estimate for the periodicity, there are differences between the estimated parameters in each state, supporting the combination of the periodic time-series model with a hidden state model. Furthermore, there are clear differences between the estimated emissions of the harmonic model compared with the estimated Gaussian emissions in the standard model (where $\beta_j^{(0)} = \mu_j$ and $\beta_j^{(1)}$ and $\beta_j^{(2)}$ were both 0). In particular, the intercept $\beta_{\mathrm{IA}}^{(0)}$ in the IA state differs non-negligibly when using the harmonic model instead of the standard Gaussian, as do $\beta_{\mathrm{IA}}^{(2)}$ and $\beta_{\mathrm{HA}}^{(1)}$, whose 90% credibility intervals do not cover 0. This all supports the selection of the harmonic model over the standard Gaussian emissions.

7 Concluding Summaries

We presented a Bayesian model for analyzing time series data based on an HSMM formulation with the goal of analyzing physical activity data collected from wearable sensing devices. We facilitate the computational feasibility of Bayesian inference for HSMMs via the likelihood approximation introduced by Langrock and Zucchini (2011), in which a special structure of the transition matrix is embedded to model the state duration distributions. We utilize the *stan* modeling language and deploy a sparse matrix formulation to further leverage the efficiency of the approximate likelihood. We showed the advantages of choosing a Bayesian paradigm over its frequentist counterpart in terms of incorporation of prior information, quantification of uncertainty, model selection, and forecasting. We additionally demonstrated the ability of the HSMM approximation to drastically reduce the computational burden of the Bayesian inference (for example reducing the time for inference on $T = 5000$ observations from > 3 days to < 2 hours), whilst incurring negligible statistical error. The proposed approach allows for the efficient implementation of highly flexible and interpretable models that incorporate available prior information on state durations. An avenue not explored in the current paper is how our model compares to particle filtering methods. For example, a referee suggested that an algorithm sampling the filtering distribution using an adaptation of the sequential Monte Carlo (SMC) sampler of Yildirim et al. (2013) inside one of the two particle MCMC algorithms of Whiteley et al. (2009) could prove competitive for HSMM inference. Further work could define, implement and compare such an approach to ours.

The analysis of physical activity data demonstrated that our model was able to learn the probabilistic dynamics governing the transitions between different activity patterns during the day as well as characterizing the sleep duration overnight. We were also able to illustrate the flexibility of the proposed model by adding harmonic covariates to the emission distribution, extending further the analysis of Huang et al. (2018). Future work will investigate the further inclusion of covariates into these time series models as well as computationally and statistically efficient approaches for conducting variable selection among these (George and McCulloch, 1993; Rossell and Telesca, 2017). We will also consider extending our methodology to account for higher-dimensional multivariate time series, where computational tractability is further challenging.

Supplementary Material

Supplementary Material to "Bayesian Approximations to Hidden Semi-Markov Models" (DOI: 10.1214/22-BA1318SUPP; .pdf). Supplementary materials are available and include further details about dwell durations, forecasting functions, graphs of normal pseudo-residuals, and further analysis of the PA time series results. Code that implements the methodology is available as online supplemental material (see also `https://github.com/Beniamino92/BayesianApproxHSMM`).

References

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle." In *Second International Symposium on Information Theory*, 267–281. MR0483125. 560

Ancoli-Israel, S., Cole, R., Alessi, C., Chambers, M., Moorcroft, W., and Pollak, C. P. (2003). "The role of actigraphy in the study of sleep and circadian rhythms." *Sleep*, 26(3): 342–392. 547

Ancoli-Israel, S., Martin, J. L., Blackwell, T., Buenaver, L., Liu, L., Meltzer, L. J., Sadeh, A., Spira, A. P., and Taylor, D. J. (2015). "The SBSM guide to actigraphy monitoring: clinical and research applications." *Behavioral Sleep Medicine*, 13(sup1): S4–S38. 548

Aung, M. H., Matthews, M., and Choudhury, T. (2017). "Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies." *Depression and Anxiety*, 34(7): 603–609. 547

Bhat, H. S. and Kumar, N. (2010). "On the derivation of the Bayesian Information Criterion." *School of Natural Sciences, University of California*. 561

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2016). "Stan: A probabilistic programming language." *Journal of Statistical Software*, 20. 549, 555

Douglas, N. J., Thomas, S., and Jan, M. A. (1992). "Clinical value of polysomnography." *The Lancet*, 339(8789): 347–350. 547

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). "Hybrid Monte Carlo." *Physics Letters B*, 195(2): 216–222. MR3960671. doi: `https://doi.org/10.1016/0370-2693(87)91197-x`. 554

Economou, T., Bailey, T. C., and Kapelan, Z. (2014). "MCMC implementation for Bayesian hidden semi-Markov models with illustrative applications." *Statistics and Computing*, 24(5): 739–752. MR3229694. doi: https://doi.org/10.1007/s11222-013-9399-z. 563

Forney, G. D. (1973). "The Viterbi algorithm." *Proceedings of the IEEE*, 61(3): 268–278. MR0439384. 550

Fujikoshi, Y. (1985). "Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria." *Journal of Multivariate Analysis*, 17(1): 27–37. MR0797518. doi: https://doi.org/10.1016/0047-259X(85)90092-2. 561

Gelfand, A. E. and Smith, A. F. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85(410): 398–409. MR1141740. 549

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press. MR3235677. 556, 557, 558, 564

Gelman, A., Hwang, J., and Vehtari, A. (2014). "Understanding predictive information criteria for Bayesian models." *Statistics and Computing*, 24(6): 997–1016. MR3253850. doi: https://doi.org/10.1007/s11222-013-9416-2. 560

Gelman, A., Simpson, D., and Betancourt, M. (2017). "The prior can often only be understood in the context of the likelihood." *Entropy*, 19(10): 555. 555, 557

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 573

Griewank, A. and Walther, A. (2008). *Evaluating derivatives: principles and techniques of algorithmic differentiation*, volume 105. SIAM. MR2454953. doi: https://doi.org/10.1137/1.9780898717761. 555

Gronau, Q., Singmann, H., and Wagenmakers, E.-J. (2020). "Bridgesampling: An R package for estimating normalizing constants." *Journal of Statistical Software*, 92(10). 549, 556

Guédon, Y. (2003). "Estimating hidden semi-Markov chains from discrete sequences." *Journal of Computational and Graphical Statistics*, 12(3): 604–639. MR2002638. doi: https://doi.org/10.1198/1061860032030. 549, 550, 563

Hadj-Amar, B., Finkenstädt, B., Fiecas, M., and Huckstepp, R. (2021). "Identifying the recurrence of sleep apnea using a harmonic hidden Markov model." *The Annals of Applied Statistics*, 15(3): 1171. MR4316645. doi: https://doi.org/10.1214/21-aoas1455. 550, 569

Hadj-Amar, B., Finkenstädt, B., Fiecas, M., Lévi, F., and Huckstepp, R. (2019). "Bayesian Model Search for Nonstationary Periodic Time Series." *Journal of the American Statistical Association*, 1–29. MR4143468. doi: https://doi.org/10.1080/01621459.2019.1623043. 548, 569

Hadj-Amar, B., Jewson, J., and Fiecas, M. (2022). "Supplementary Material to "Bayesian Approximations to Hidden Semi-Markov Models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1318SUPP. 557

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research*, 15(1): 1593–1623. MR3214779. 554

Huang, Q., Cohen, D., Komarzynski, S., Li, X.-M., Innominato, P., Lévi, F., and Finkenstädt, B. (2018). "Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data." *Journal of The Royal Society Interface*, 15(139): 20170885. 547, 548, 550, 560, 565, 566, 569, 572

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford. MR1647885. 561

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. MIT press. 550

Jennison, C. (1997). "Discussion of "On Bayesian analysis of mixtures with an unknown number of components" by S. Richardson and P. J. Green." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4): 778–779. MR1483213. doi: https://doi.org/10.1111/1467-9868.00095. 561

Jewson, J., Smith, J., and Holmes, C. (2018). "Principles of Bayesian inference using general divergence criteria." *Entropy*, 20(6): 442. MR3879894. doi: https://doi.org/10.3390/e20060442. 559

Johnson, M. J. and Willsky, A. S. (2013). "Bayesian nonparametric hidden semi-Markov models." *Journal of Machine Learning Research*, 14(Feb): 673–701. MR3033344. 550

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90(430): 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 555, 567

Kaur, G., Phillips, C., Wong, K., and Saini, B. (2013). "Timing is important in medication administration: a timely review of chronotherapy research." *International Journal of Clinical Pharmacy*, 35(3): 344–358. 547

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media. MR2367855. doi: https://doi.org/10.1007/978-0-387-71887-3. 561

Langrock, R., Swihart, B. J., Caffo, B. S., Punjabi, N. M., and Crainiceanu, C. M. (2013). "Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms." *Statistics in Medicine*, 32(19): 3342–3356. MR3074361. doi: https://doi.org/10.1002/sim.5747. 550

Langrock, R. and Zucchini, W. (2011). "Hidden Markov models with arbitrary state dwell-time distributions." *Computational Statistics & Data Analysis*, 55(1): 715–724. MR2736590. doi: https://doi.org/10.1016/j.csda.2010.06.015. 548, 549, 551, 552, 553, 557, 558, 559, 560, 564, 572

Leroux, B. G. and Puterman, M. L. (1992). "Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models." *Biometrics*, 545–558. 554

Lindley, D. V. (1957). "A statistical paradox." *Biometrika*, 44(1/2): 187–192. MR0087273. doi: https://doi.org/10.1093/biomet/44.1-2.179. 561

Meng, X.-L. and Schilling, S. (2002). "Warp bridge sampling." *Journal of Computational and Graphical Statistics*, 11(3): 552–586. MR1938446. doi: https://doi.org/10.1198/106186002457. 549, 555

Meng, X.-L. and Wong, W. H. (1996). "Simulating ratios of normalizing constants via a simple identity: a theoretical exploration." *Statistica Sinica*, 831–860. MR1422406. 549, 555, 556

Migueles, J. H., Cadenas-Sanchez, C., Ekelund, U., Nyström, C. D., Mora-Gonzalez, J., Löf, M., Labayen, I., Ruiz, J. R., and Ortega, F. B. (2017). "Accelerometer data collection and processing criteria to assess physical activity and other outcomes: a systematic review and practical considerations." *Sports Medicine*, 47(9): 1821–1845. 566

Neal, R. M. et al. (2011). "MCMC using Hamiltonian dynamics." *Handbook of Markov chain Monte Carlo*, 2(11): 2. MR2858447. 554

O'Hagan, A. and Forster, J. J. (2004). *Kendall's advanced theory of statistics, volume 2B: Bayesian inference*, volume 2. Arnold. MR3237119. 561

Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." *Proceedings of the IEEE*, 77(2): 257–286. 550

Rabiner, L. R., Wilpon, J. G., and Soong, F. K. (1989). "High performance connected digit recognition using hidden Markov models." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(8): 1214–1225. 550

Raviv, J. (1967). "Decision making in Markov chains applied to the problem of pattern recognition." *IEEE Transactions on Information Theory*, 13(4): 536–551. 550

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media. MR2723361. 556

Roberts, G. O., Rosenthal, J. S., et al. (2001). "Optimal scaling for various Metropolis-Hastings algorithms." *Statistical science*, 16(4): 351–367. MR1888450. doi: https://doi.org/10.1214/ss/1015346320. 570

Rossell, D. and Telesca, D. (2017). "Nonlocal priors for high-dimensional estimation." *Journal of the American Statistical Association*, 112(517): 254–265. MR3646569. doi: https://doi.org/10.1080/01621459.2015.1130634. 573

Sadeh, A. (2011). "The role and validity of actigraphy in sleep medicine: an update." *Sleep Medicine Reviews*, 15(4): 259–267. 547

Schwarz, G. et al. (1978). "Estimating the dimension of a model." *The Annals of Statistics*, 6(2): 461–464. MR0468014. 560

Silva, B. M., Rodrigues, J. J., de la Torre Díez, I., López-Coronado, M., and Saleem, K. (2015). "Mobile-health: A review of current state in 2015." *Journal of Biomedical Informatics*, 56: 265–272. 547

Sodhi, N. S. and Ehrlich, P. R. (2010). *Conservation biology for all*. Oxford University Press. 561

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583–639. MR1979380. doi: https://doi.org/10.1111/1467-9868.00353. 560

Stan Development Team (2018). "Stan Functions Reference." URL https://mc-stan.org/docs/2_23/functions-reference/index.html 555

Stephens, M. (2000). "Dealing with label switching in mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 795–809. MR1796293. doi: https://doi.org/10.1111/1467-9868.00265. 555

Watanabe, S. (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of Machine Learning Research*, 11(Dec): 3571–3594. MR2756194. 560

Whiteley, N., Andrieu, C., and Doucet, A. (2009). "Particle Markov chain Monte Carlo for multiple change-point problems." *Department of Mathematics, Bristol University, Bristol, UK, Technical Report*, 911. 572

Williams, J., Roth, A., Vatthauer, K., and McCrae, C. S. (2013). "Cognitive behavioral treatment of insomnia." *Chest*, 143(2): 554–565. 547

Yildirim, S., Singh, S. S., and Doucet, A. (2013). "An online expectation–maximization algorithm for changepoint models." *Journal of Computational and Graphical Statistics*, 22(4): 906–926. MR3173749. doi: https://doi.org/10.1080/10618600.2012.674653. 572

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov models for time series: an introduction using R*. CRC press. MR3618333. 549, 550, 553, 554, 558

Acknowledgments

The authors would like to thank the Editor, the Referee, the Associate Editor, David Rossell, and Marina Vannucci for their insightful and valuable comments.